ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΤΜΗΜΑΤΟΣ ΕΦΑΡΜΟΣΜΕΝΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΕΞΑΓΩΓΗ ΟΝΤΟΛΟΓΙΩΝ ΑΠΟ ΑΔΟΜΗΤΟ ΚΕΙΜΕΝΟ ΜΕ ΤΗ ΧΡΗΣΗ ΚΑΤΑΛΛΗΛΩΝ ΕΡΓΑΛΕΙΩΝ

Διπλωματική Εργασία

του

Παπατζίκου Νικόλαου

Θεσσαλονίκη, Ιούνιος 2020

ΕΞΑΓΩΓΗ ΟΝΤΟΛΟΓΙΩΝ ΑΠΟ ΑΔΟΜΗΤΟ ΚΕΙΜΕΝΟ ΜΕ ΤΗ ΧΡΗΣΗ ΚΑΤΑΛΛΗΛΩΝ ΕΡΓΑΛΕΙΩΝ

Παπατζίκος Νικόλαος

Δίπλωμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών, ΑΠΘ, 1999
Μεταπτυχιακό Δίπλωμα ειδίκευσης στα Πληροφοριακά Συστήματα, ΠΑ.ΜΑΚ., 2002

Διπλωματική Εργασία

υποβαλλόμενη για τη μερική εκπλήρωση των απαιτήσεων του

ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΤΙΤΛΟΥ ΣΠΟΥΔΩΝ ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΠΛΗΡΟΦΟΡΙΚΗ

Επιβλέπουσα Καθηγήτρια
Κολωνιάρη Γεωργία

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 26/06/2020

Κολωνιάρη Γεωργία	Ευαγγελίδης Γεώργιος	Κεραμόπουλος Ευκλείδης
.....	.....	.....

Παπατζίκος Νικόλαος

.....

Περίληψη

Στον Σημασιολογικό Ιστό, οι οντολογίες ορίζουν τις έννοιες και τις σχέσεις που χρησιμοποιούνται για να περιγράψουν και να αντιπροσωπεύσουν έναν τομέα ενδιαφέροντος. Οι οντολογίες χρησιμοποιούνται για να ταξινομήσουν τους όρους που μπορούν να χρησιμοποιηθούν σε μια συγκεκριμένη εφαρμογή, να χαρακτηρίσουν πιθανές σχέσεις και να καθορίσουν πιθανούς περιορισμούς στη χρήση αυτών των όρων. Η *OWL* (*Web Ontology Language*) είναι μια γλώσσα Σημασιολογικού Ιστού που έχει σχεδιαστεί για χρήση από εφαρμογές οι οποίες χρειάζεται να επεξεργάζονται το περιεχόμενο των πληροφοριών αντί να παρουσιάζουν μόνο πληροφορίες στον άνθρωπο. Η οντολογία *OWL* περιγράφει την ιεραρχική οργάνωση των ιδεών σε έναν τομέα, με τρόπο που μπορεί να αναλυθεί και να κατανοηθεί από το λογισμικό.

Συνήθως, ο ορισμός οντολογιών για έναν τομέα ενδιαφέροντος γίνεται από ειδικούς του τομέα και είναι μια δύσκολη και χρονοβόρα εργασία. Η αυτόματη εξαγωγή οντολογιών από δομημένη πληροφορία έχει επιλύσει επιμέρους αυτό το πρόβλημα. Ωστόσο, αφού περισσότερη πληροφορία βρίσκεται σε αδόμητη μορφή διαθέσιμη στον παγκόσμιο ιστό, ένα ενδιαφέρον και πιο δύσκολο πρόβλημα είναι η αυτόματη εξαγωγή οντολογιών από έγγραφα σε φυσική γλώσσα.

Στόχος της εργασίας είναι η πειραματική μελέτη της εφαρμογής εκμάθησης οντολογιών *Text2Onto* για την αυτόματη εξαγωγή οντολογικής γνώσης, σε γλώσσα *OWL*, από κείμενο σε φυσική γλώσσα. Η εφαρμογή *Text2Onto* είναι ένα από τα ελάχιστα εργαλεία εκμάθησης οντολογιών το οποίο έχει αναπτυχθεί για να υποστηρίξει τη δημιουργία οντολογιών από αδόμητο κείμενο.

Η μεθοδολογία που ακολουθήθηκε συνοψίζεται στα παρακάτω βήματα:

- Μελέτη της εφαρμογής *Text2Onto* και της σχετικής τεκμηρίωσης. Πειραματισμός με τη χρήση της εφαρμογής.
- Πειραματισμός με τη χρήση διαφόρων ειδών δεδομένων ως είσοδο. Θεωρήθηκε σκόπιμο να επιλεχθεί ως τομέας πειραματισμού, ένας τομέας ενδιαφέροντος διαδεδομένος σε ένα ευρύ φάσμα ανθρώπων, γι' αυτό και επιλέχθηκε ο τομέας των κινηματογραφικών ταινιών και τηλεοπτικών σειρών. Επιλέχθηκαν ως δεδομένα, κριτικές και στοιχεία ταινιών και σειρών από διάφορες πηγές.

- Πειραματισμός με ποσοτικές μεταβολές των δεδομένων και τη χρήση διαφορετικού συνδυασμού αλγορίθμων της εφαρμογής *Text2Onto*. Μελέτη των αντίστοιχων αποτελεσμάτων οντολογιών που προκύπτουν.
- Ποσοτική και ποιοτική ανάλυση των αποτελεσμάτων με έλεγχο και επεξεργασία του εξαγόμενου αρχείου (*.owl*). Είσοδος των εξαγόμενων αποτελεσμάτων στο περιβάλλον οντολογιών *Protégé*. Σύγκριση της εξαγόμενης οντολογίας με ένα πρότυπο οντολογίας σχετικά με κινηματογραφικές ταινίες. Για όλα τα παραπάνω αναπτύχθηκαν διαδικασίες κώδικα σε γλώσσα προγραμματισμού *Javascript*.
- Δημιουργία γραφημάτων σχετικών με την ερευνητική δραστηριότητα με σκοπό την αποτίμησή της.
- Εξαγωγή συμπερασμάτων για το αν μπορούν και σε τι βαθμό να αξιοποιηθούν οι κριτικές ταινιών για τη δημιουργία οντολογίας.
- Εξαγωγή συμπερασμάτων σχετικά με την έκταση των δυνατοτήτων της εφαρμογής και τους τρόπους χρησιμοποίησής της ως εργαλείο στον τομέα των οντολογιών.

Λέξεις Κλειδιά: Σημασιολογικός Ιστός, Οντολογίες, *OWL*, *Text2Onto*, Κινηματογραφικές ταινίες, *Protégé*.

Abstract

In Semantic Web, ontologies define the concepts and relationships used to describe and represent an area of interest. Ontologies are used to classify the terms that can be used in a particular application, to characterize possible relationships, and to determine possible limitations in the use of those terms. *OWL* (***W**eb **O**ntology **L**anguage*) is a Semantic Web language designed for use by applications that need to process the content of information instead of presenting only information to humans. OWL ontology describes the hierarchical organization of ideas in a field, in a way that can be analyzed and understood by software.

Usually, the definition of ontologies for an area of interest is done by experts in the field and is a difficult and time consuming task. Automatic extraction of structured ontologies has solved this problem individually. However, since more information is available in an unstructured form available on the World Wide Web, a more interesting and difficult problem is the automatic extraction of ontologies from documents in natural language.

The aim of this work is the experimental study of Text2Onto, an ontology learning application for the automatic extraction of ontological knowledge, in OWL language, from text in natural language. Text2Onto is one of the few ontology learning tools that has been developed to support the creation of ontologies from unstructured text.

The methodology followed is summarized in the following steps:

- Study of Text2Onto application and the relevant documentation. Experimenting with the application.
- Experimentation using different types of data as input. It was considered appropriate to be chosen as a field of experimentation, an area of interest widespread in a wide range of people, which is why the field of film and television series was chosen. Data and reviews of films and series from various sources were selected as data.
- Experimentation with quantitative data also changes the use of a different combination of algorithms by Text2Onto application. Study of the corresponding ontological results that come up.
- Quantitative and qualitative analysis of the results with control and process the exported file (.owl). Entry of the output results into the Protégé ontology

environment. Comparison of extracted ontology with an ontology standard for movies. For all of the above, some scripts have been developed in Javascript programming language.

- Creation of graphs related to the research activity in order to evaluate it.
- Drawing conclusions about whether and to what extent film reviews can be used to create an ontology.
- Drawing conclusions about the extent of the application's capabilities and how to use it as a tool in the field of ontologies.

Keywords: Semantic Web, Ontologies, OWL, Text2Onto, Cinema movies, Protégé.

Πρόλογος – Ευχαριστίες

Η επιλογή του θέματος της συγκεκριμένης εργασίας ξεκίνησε κατά τη διάρκεια προσωπικής έρευνας περί ευρέσεως θέματος διπλωματικής εργασίας για την εκπλήρωση των απαιτήσεων του μεταπτυχιακού τίτλου σπουδών στην Εφαρμοσμένη Πληροφορική του Πανεπιστημίου Μακεδονίας. Το θέμα επιλέχθηκε από την προτεινόμενη λίστα θεμάτων διπλωματικών εργασιών του ακαδημαϊκού έτους 2018-19 καθώς υπήρχε προτίμηση για ένα θέμα σχετικό με δεδομένα και πληροφορία. Κατά τη διάρκεια εκπόνησης της εργασίας υπήρξαν προβληματισμοί για την πορεία και τα αποτελέσματα των πειραμάτων της ερευνητικής προσπάθειας αλλά ταυτόχρονα δημιουργούνταν και νέες προκλήσεις επάνω στο πεδίο των πειραματισμών και των αναλύσεων των αποτελεσμάτων. Έγινε προσπάθεια για την ύπαρξη μιας σαφούς μεθοδολογίας με την εξαγωγή των βέλτιστων δυνατών συμπερασμάτων που να εκπληρώνουν τους στόχους της εργασίας.

Θερμές ευχαριστίες στην επιβλέπουσα καθηγήτρια, κα Κολωνιάρη Γεωργία για την καθοδήγησή της και τις καίριες και χρήσιμες παρατηρήσεις της σε όλη τη διάρκεια εκπόνησης της εργασίας.

Περιεχόμενα

Κατάλογος Εικόνων

Κατάλογος Πινάκων

1 Εισαγωγή

1.1 Πρόβλημα – Σημαντικότητα του θέματος

Για την εξέταση του θέματος της αναπαράστασης πληροφορίας και των δομών γνώσης είναι απαραίτητο να κατανοηθεί η διαδικασία της κατάταξης, εξόρυξης και της λογικής δομής της πληροφορίας και της γνώσης. Οι δομές γνώσης και αναπαράστασης πληροφορίας βρίσκουν εφαρμογή σε πολλούς και ζωτικούς τομείς της σύγχρονης κοινωνίας όπως η υγεία, η εθνική ασφάλεια, η εκπαίδευση καθώς και στον τομέα των επιχειρήσεων. Μέσω του Διαδικτύου και των πληροφοριακών μέσων η εφαρμογή τους επεκτείνεται και σε επιπλέον τομείς (Τσουκαλά (2018, σ.2)).

Η αύξηση του πλήθους των πληροφοριών είχε ως αποτέλεσμα την ανάγκη μοντελοποίησης ή κατάταξής της, ώστε να είναι δυνατή και η ανάκτηση της επιθυμητής πληροφορίας. Η μοντελοποίηση της κατάταξης πληροφοριών αποτελεί τη λεγόμενη "Οντολογία", την προέλευση δηλαδή και τη λογική σημασιολογική κατάταξη μιας έννοιας ή πληροφορίας, η οποία χρησιμοποιείται από διάφορα πληροφοριακά συστήματα για διάφορους σκοπούς (Τσουκαλά (2018, σ.3)).

Οι βασικοί τομείς εφαρμογής των οντολογιών είναι οι εξής : Σημασιολογικός Ιστός, Τεχνητή νοημοσύνη, Ανάπτυξη λογισμικού, Βιοϊατρική Πληροφορία, Βιβλιοθηκονομία και Αρχιτεκτονική της πληροφορίας σε μορφή γνώσης. Δίνοντας ένα γενικότερο πλαίσιο, θα μπορούσαμε να ορίσουμε τις οντολογίες ως τον πυλώνα κατάταξης και εξόρυξης δεδομένων, πληροφορίας ή γνώσης, οπουδήποτε αυτή διατίθεται σε μεγάλες ποσότητες (Τσουκαλά (2018, σ.2)).

Υπάρχουν δύο βασικοί τομείς όσον αφορά τις οντολογίες: ο πρώτος αφορά την κατασκευή οντολογιών και ο δεύτερος την αναζήτηση – εξόρυξη πληροφοριών από υπάρχουσες οντολογίες με την χρήση ερωτημάτων. Στην περίπτωση της κατασκευής οντολογιών και την κατάταξη και καταγραφή γνώσης και πληροφοριών, ο βασικός στόχος είναι η δημιουργία ενός λεξιλογίου καταγεγραμμένης πληροφορίας το οποίο θα αποτελεί πιστή αναπαράσταση γνώσης. Μέσω του ταξινομικού αυτού πλαισίου που λειτουργεί ως βάση δεδομένων, αντλούνται οι χρήσιμες πληροφορίες από τα αντίστοιχα πληροφοριακά συστήματα. Ακολούθως, σε ό, τι έχει να κάνει με την έρευνα και εξόρυξη πληροφοριών από υπάρχουσες οντολογίες, η αναζήτηση και κατάταξη της πληροφορίας γίνεται μέσω προτύπων ή γλωσσών τα οποία αποτελούν τα αντίστοιχα εργαλεία προς αυτήν την κατεύθυνση. Είναι απαραίτητο να σημειωθεί πως ανάλογα με την

μεθοδολογία και τον αντίστοιχο φορέα κατασκευής οντολογιών αντιστοιχίζεται η κάθε γλώσσα ανά περίπτωση (Τσουκαλά (2018, σ.3)). Περισσότερες λεπτομέρειες για τις γλώσσες και τα πρότυπα οντολογιών διατυπώνονται σε επόμενη ενότητα.

Συνήθως ο ορισμός οντολογιών για έναν τομέα ενδιαφέροντος γίνεται από ειδικούς του τομέα και είναι μια δύσκολη και χρονοβόρα εργασία. Η αυτόματη εξαγωγή οντολογιών από δομημένη πληροφορία έχει επιλύσει επιμέρους αυτό το πρόβλημα. Ωστόσο, αφού περισσότερη πληροφορία βρίσκεται σε αδόμητη μορφή διαθέσιμη στον παγκόσμιο ιστό, ένα ενδιαφέρον και πιο δύσκολο πρόβλημα είναι η αυτόματη εξαγωγή οντολογιών από έγγραφα σε φυσική γλώσσα.

1.2 Σκοπός – Στόχοι

Η παρούσα εργασία πραγματεύεται την πειραματική μελέτη της εφαρμογής *Text2Onto* για την αυτόματη εξαγωγή οντολογικής γνώσης από κείμενο σε φυσική γλώσσα. Η εφαρμογή *Text2Onto* είναι ένα από τα ελάχιστα εργαλεία εκμάθησης οντολογιών το οποίο έχει αναπτυχθεί για να υποστηρίξει τη δημιουργία οντολογιών από αδόμητο κείμενο. Πρώτος στόχος της εργασίας είναι η εξαγωγή συμπερασμάτων σχετικά με την έκταση των δυνατοτήτων και τους τρόπους εφαρμογής του εργαλείου στον τομέα των οντολογιών. Ένας ακόμα στόχος είναι να εξαχθούν συμπεράσματα για το αν μπορούν και σε τι βαθμό να αξιοποιηθούν οι κριτικές κινηματογραφικών ταινιών για τη δημιουργία οντολογίας.

1.3 Διάρθρωση της μελέτης

Η εργασία διαρθρώνεται σε συνολικά εννέα κεφάλαια. Το παρόν εισαγωγικό κεφάλαιο αναφέρεται στην προβληματική της εργασίας καθώς και στους στόχους της. Στο 2^{ο} κεφάλαιο γίνεται μια εισαγωγή στον χώρο του σημασιολογικού ιστού. Στο 3^{ο} κεφάλαιο δίνονται κάποια βασικά στοιχεία σε σχέση με τις οντολογίες και τις γλώσσες/πρότυπα οντολογιών (*RDF, RDFS, OWL*). Στο 4^{ο} κεφάλαιο γίνεται αναφορά στο περιβάλλον οντολογιών *Protégé* το οποίο χρησιμοποιείται στην πορεία των πειραμάτων. Το 5^{ο} κεφάλαιο αναφέρεται στην εφαρμογή εκμάθησης οντολογιών *Text2Onto*, στον τρόπο χρήσης της, καθώς και σε παραδείγματα περιπτώσεων που έχει χρησιμοποιηθεί. Τα πειράματα με την εφαρμογή *Text2Onto*, με είσοδο δεδομένων αδόμητου κειμένου από τον τομέα των κινηματογραφικών ταινιών και τηλεοπτικών σειρών, καθώς και τα αποτελέσματα που λαμβάνονται παρουσιάζονται στο 6^{ο} κεφάλαιο. Στο 7^{ο} κεφάλαιο γίνεται αναφορά στις διαδικασίες κώδικα που αναπτύχθηκαν σε

γλώσσα προγραμματισμού *Javascript* για την ποσοτική και ποιοτική ανάλυση των αποτελεσμάτων των πειραμάτων που εξάγονται από το *Text2Onto*. Με χρήση των παραπάνω διαδικασιών κώδικα στο 8^{ο} κεφάλαιο αναλύονται τα αποτελέσματα των πειραμάτων, η σύγκριση των αποτελεσμάτων μεταξύ διαφόρων πειραμάτων. Παρουσιάζονται γραφήματα που δημιουργήθηκαν με βάση την παραπάνω ποσοτική και ποιοτική ανάλυση των αποτελεσμάτων, γίνεται σχολιασμός τους και εξάγονται συμπεράσματα. Επίσης αναφέρεται η διαδικασία εισόδου των εξαγόμενων αποτελεσμάτων στο περιβάλλον οντολογιών *Protégé* καθώς και η σύγκριση των εξαγόμενων οντολογιών με ένα πρότυπο οντολογίας σχετικά με κινηματογραφικές ταινίες. Τέλος, στο 9^{ο} κεφάλαιο γίνεται μια τελική σύνοψη της όλης μελέτης.

2 Σημασιολογικός Ιστός

2.1 Περιγραφή

Ο Σημασιολογικός Ιστός έχει ως βάση του μια συλλογή από τεχνολογίες και μεθόδους μέσω των οποίων δίνεται η δυνατότητα στους υπολογιστές να αντιλαμβάνονται τη σημασία της πληροφορίας που διαχειρίζονται. Στο Διαδίκτυο κυκλοφορεί πολύ μεγάλος όγκος πληροφορίας. Η προσθήκη σημασίας σε αυτή την πληροφορία θα ενισχύσει κατά πολύ την ευφυή εκμετάλλευσή της μέσω της απελευθέρωσης μεγάλου πλήθους δυνατοτήτων. Οι τεχνολογίες και οι μέθοδοι του Σημασιολογικού Ιστού έχουν ως σκοπό η πληροφορία που παρέχεται στο χρήστη να είναι όσο πιο σχετική γίνεται με την αναζήτησή του. Ο Σημασιολογικός Ιστός περιλαμβάνει και τα πληροφοριακά συστήματα μέσω των οποίων εξάγεται πληροφορία από την ήδη υπάρχουσα. Εμπνευστής του Σημασιολογικού Ιστού είναι ο Tim Berners-Lee που υπήρξε και ο ιδρυτής της κοινοπραξίας *W3C* (*World Wide Web Consortium*) που έχει σαν στόχο την ανάπτυξη των προδιαγραφών, των συνεργασιών και της τεχνολογίας του Διαδικτύου και του Σημασιολογικού Ιστού (Γιάου (2013, σ.8)).

2.2 Έξυπνος Ιστός – «Χαζός» Ιστός

Έστω ότι κάποιος πραγματοποιεί στο Διαδίκτυο μια αναζήτηση σχετικά με ένα εθνικό πάρκο. Μπαίνοντας στον ιστότοπο του εθνικού πάρκου μεταξύ άλλων βλέπει και μια λίστα των ξενοδοχείων που έχουν υποκαταστήματα κοντά στο πάρκο. Σε αυτή τη λίστα βλέπει πως μία γνωστή αλυσίδα ξενοδοχείων έχει ένα υποκατάστημα στην περιοχή και έτσι αποφασίζει να κλείσει εκεί ένα δωμάτιο. Μπαίνοντας στην ιστοσελίδα της αλυσίδας αναζητά την τοποθεσία του ξενοδοχείου και προς έκπληξή του δεν βρίσκει να υπάρχει υποκατάστημα κοντά στο εθνικό πάρκο. Γιατί υπάρχει αυτή η αναντιστοιχία; Φαίνεται να είναι κάτι «χαζό» από τη στιγμή που στον ιστότοπο του εθνικού πάρκου, η αλυσίδα φαίνεται να έχει εκεί κοντά υποκατάστημα (Allemang, Hendler (2011, σ.2)).

Ας υποθέσουμε ότι κάποιος πρόκειται να παρακολουθήσει ένα συνέδριο σε μια πόλη. Πηγαίνει στην ιστοσελίδα μιας αλυσίδας ξενοδοχείων της προτίμησης του με σκοπό την εύρεση ξενοδοχείου στην περιοχή. Η βασική ερώτηση που προκύπτει άμεσα στον ενδιαφερόμενο είναι σχετικά με το ποιο είναι το πιο κοντινό ξενοδοχείο της αλυσίδας στο συνέδριο και σε πόση απόσταση βρίσκεται. Από τη στιγμή που δεν υπάρχει η αντίστοιχη πληροφορία στην ιστοσελίδα της αλυσίδας, ο ενδιαφερόμενος αναγκάζεται να πάρει απάντηση στην ερώτηση του μέσω κάποιου ιστότοπου που μπορεί

να υπολογίσει τις αποστάσεις αυτές, αφού πάρει τις διευθύνσεις του συνεδρίου και του ξενοδοχείου. Αφού ο ενδιαφερόμενος μπαίνει στη διαδικασία να αντιγράφει και να επικολλά διευθύνσεις από τη μια σελίδα στην άλλη και να σημειώνει τις αποστάσεις, αυτό που θα σκεφτεί ενστικτωδώς είναι το εξής: «Γιατί θα πρέπει ο ίδιος να αντιγράψει και να επικολλήσει όλες αυτές τις πληροφορίες από τη μια σελίδα στην άλλη και να μην υπάρχει κάποιος τρόπος όλα αυτά να γίνονται αυτόματα μέσω της ιστοσελίδας της αλυσίδας, ώστε να πάρει την πληροφορία που θέλει;» (Allemang, Hendler (2011, σ.2)).

Τα δύο αυτά παραδείγματα έχουν κάτι το κοινό. Καθένα από αυτά έχει μια αναπαράσταση δεδομένων των οποίων η παρουσίαση στο χρήστη φαίνεται «χαζή». Στη συγκεκριμένη περίπτωση, «χαζή» σημαίνει ασυνεπής, μη συγχρονισμένη και αποσυνδεδεμένη. Το ερώτημα που τίθεται στην προκειμένη περίπτωση είναι το τι θα χρειαστεί για να γίνουν ανάλογες εμπειρίες στο Διαδίκτυο πιο έξυπνες. Η λύση θα δοθεί μέσω εξυπνότερων εφαρμογών ή μήπως μέσω μιας πιο έξυπνης υποδομής Ιστού; (Allemang, Hendler (2011, σ.3)).

2.3 Έξυπνες εφαρμογές Διαδικτύου

Η δημιουργία έξυπνων και καινοτόμων εφαρμογών είναι ένα διαρκώς αυξανόμενο φαινόμενο στο Διαδίκτυο. Καθημερινά εμφανίζονται ιστότοποι με νέες δυνατότητες. Με βάση αυτό το φαινόμενο το ερώτημα που τίθεται είναι το πώς μπορεί η υποδομή του Παγκόσμιου Ιστού να κάνει αυτές τις εφαρμογές περισσότερο "έξυπνες"; Ο λόγος που επιδιώκεται η βελτίωση της υποδομής του Παγκόσμιου Ιστού είναι για να μπορέσουν οι έξυπνες εφαρμογές του Διαδικτύου να ανταποκριθούν στην αποστολή τους στο μέγιστο βαθμό σύμφωνα με τις προδιαγραφές τους. Ακόμα και οι εξυπνότερες και πιο διορατικές εφαρμογές είναι τόσο έξυπνες όσο τα δεδομένα που είναι διαθέσιμα τους το επιτρέπουν. Οι ασυνεπείς ή αντιφατικές είσοδοι θα εξακολουθήσουν να προκαλούν σύγχυση, αποσύνδεση, "χαζά" αποτελέσματα, ακόμη και από πολύ έξυπνες εφαρμογές. Επομένως μπορούμε να πούμε πως σε ό, τι έχει να κάνει με τον σχεδιασμό του Σημασιολογικού Ιστού η πρόκληση δεν είναι να δημιουργηθεί μια υποδομή Παγκόσμιου Ιστού που είναι όσο το δυνατόν εξυπνότερη, αλλά μια υποδομή που είναι όσο το δυνατόν πιο κατάλληλη για την εργασία ενσωμάτωσης πληροφοριών στον Ιστό (Allemang, Hendler (2011, σ.3)).

Ο στόχος του Σημασιολογικού Ιστού δεν είναι να κάνει τα δεδομένα έξυπνα, επειδή τα έξυπνα δεδομένα δεν είναι αυτό που χρειάζεται. Ο στόχος του είναι να βρεθούν τα δεδομένα που πρέπει στον τόπο που πρέπει, έτσι ώστε να μπορέσουν οι

έξυπνες εφαρμογές να εκτελεστούν σύμφωνα με τις δυνατότητες τους. Επομένως, το να τεθεί το ερώτημα του πώς μπορούμε να κάνουμε την υποδομή του Παγκόσμιου Ιστού εξυπνότερη δεν έχει κάποιο νόημα. Αντίθετα έχει νόημα το να τεθεί το ερώτημα του τι μπορεί να προσφέρει η υποδομή του Παγκόσμιου Ιστού για τη βελτίωση της συνέπειας και της διαθεσιμότητας των δεδομένων στον Παγκόσμιο Ιστό. (Allemang, Hendler (2011, σ.3)).

2.4 Ένας κατανεμημένος ιστός δεδομένων

Η υποδομή του Παγκόσμιου Ιστού αποτελείται από ένα κατανεμημένο δίκτυο ιστοσελίδων που μπορούν να αναφέρονται μεταξύ τους με συνδέσεις που ονομάζονται Ενιαίοι Εντοπιστές Πόρων (διευθύνσεις *URL*). Ο Σημασιολογικός Ιστός έχει ως κύριο στόχο του τη δημιουργία ενός κατανεμημένου ιστού στο επίπεδο των δεδομένων και όχι στο επίπεδο της παρουσίασης. Χρησιμοποιώντας την υποδομή του Παγκόσμιου Ιστού τα στοιχεία δεδομένων μπορούν να αναφέρονται μεταξύ τους, χρησιμοποιώντας παγκόσμιες αναφορές που ονομάζονται Ενιαίοι Προσδιορισμοί Πόρων (διευθύνσεις *URI*). Η υποδομή του Παγκόσμιου Ιστού παρέχει ένα μοντέλο δεδομένων με το οποίο μπορούν να διανεμηθούν πληροφορίες σε σχέση με μια συγκεκριμένη οντότητα επάνω από το Διαδίκτυο. Αυτή η διανομή θα επέτρεπε στα παραδείγματα των ξενοδοχείων του εθνικού πάρκου και του συνεδρίου που αναφέρθηκαν νωρίτερα, παρόλο που οι πληροφορίες διανέμονται σε ιστότοπους ελεγχόμενους από περισσότερους από έναν οργανισμό, να λειτουργήσουν σαν να είχαν να κάνουν με ένα μόνο οργανισμό. Το μοντέλο δεδομένων είναι ανεξάρτητο από τις εφαρμογές του Διαδικτύου που θα το χρησιμοποιήσουν καθώς αποτελεί μέρος της υποδομής του Παγκόσμιου Ιστού. Όταν μια αλυσίδα ξενοδοχείων δημοσιεύει πληροφορίες σχετικά με τα ξενοδοχεία και τις τοποθεσίες τους, δεν δημοσιεύει απλώς μια παρουσίαση αυτών των πληροφοριών που είναι ανθρώπινα αναγνώσιμη, αλλά αντίθετα μια διανεμητέα, αναγνώσιμη από μηχανή, περιγραφή των δεδομένων. Το μοντέλο δεδομένων που η υποδομή του Σημασιολογικού Ιστού χρησιμοποιεί για την απεικόνιση αυτού του κατανεμημένου ιστού δεδομένων ονομάζεται *RDF* (*Resource Description Framework*), (Allemang, Hendler (2011, σ.6)).

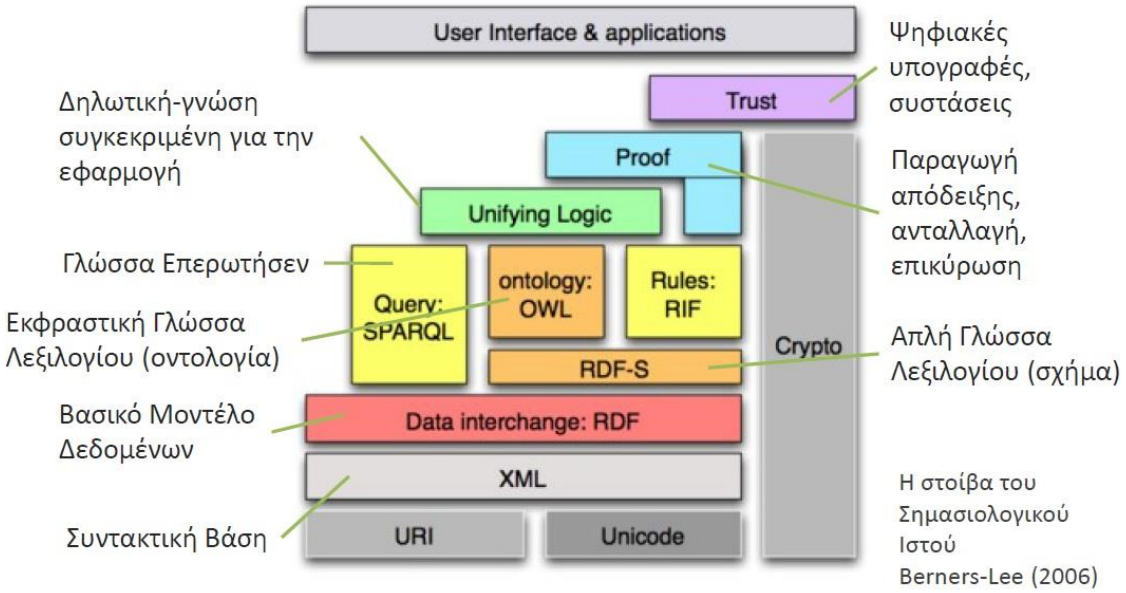
2.5 Σύνοψη

Ο Παγκόσμιος Ιστός σήμερα είναι ένα απείθαρχο μέρος που χαρακτηρίζεται από μεγάλη ποικιλία διαφορετικών πηγών, οργανισμών και στυλ πληροφοριών. Οι διάφορες αντιφάσεις και ασυνέπειες σε έγγραφα του Παγκόσμιου Ιστού μπορούν να

αντιμετωπιστούν από τη διαδικασία της ανθρώπινης παρατήρησης και της εφαρμογής της κοινής λογικής. Το ζητούμενο είναι πώς με μια μηχανή που συνδυάζει πληροφορίες μπορεί να έρθει κάποια τάξη στο χάος και πώς μπορεί κάποιος να έχει εμπιστοσύνη στις πληροφορίες που συγχωνεύονται από πολλαπλές πηγές. Η απείθαρχη φύση του Παγκόσμιου Ιστού επεκτείνεται σε υπερθετικό βαθμό στο επίπεδο του Σημασιολογικού Ιστού, με έναν πλούσιο όγκο διασυνδεδεμένων πληροφοριών χωρίς κανένα χάρτη, ευρετήριο ή καθοδήγηση (Allemang, Hendler (2011, σ.11)).

Επομένως το ερώτημα που τίθεται είναι το πώς μπορεί να καταστεί ένα τέτοιο μέρος χρήσιμο. Οι ειδικοί στον τομέα των οντολογιών έχουν ως αντικείμενο της εργασίας τους να αντιμετωπίσουν αυτήν την πρόκληση, έχοντας ως εργαλεία τις γλώσσες/πρότυπα των οντολογιών του Σημασιολογικού Ιστού, *RDF*, *RDFS*, *SPARQL* και *OWL*. Η τέχνη τους είναι να φτιάχνουν λογικούς, χρησιμοποιήσιμους και ανθεκτικούς πόρους πληροφορίας από το μέσο του κατανεμημένου ιστού δεδομένων (Allemang, Hendler (2011, σ.11)).

Με βάση τα παραπάνω η αρχιτεκτονική του Σημασιολογικού Ιστού κατά τον Tim Berners-Lee φαίνεται στην Εικόνα 2-1 που ακολουθεί.



Εικόνα 2-1: Η Αρχιτεκτονική του Σημασιολογικού Ιστού

3 Οντολογίες – Γλώσσες/Πρότυπα οντολογιών

3.1 Ορισμός οντολογιών

Στο πεδίο της επιστήμης της πληροφορικής ο όρος οντολογία ορίζεται ως το σύνολο του λεξιλογίου και των εννοιών που χρησιμοποιούνται για την περιγραφή και την αναπαράσταση της επίνοιας μιας θεματικής περιοχής. Με τον όρο επίνοια αναφερόμαστε στην απλοποιημένη μορφή του κόσμου που είναι επιθυμητό να αναπαρασταθεί για ένα συγκεκριμένο λόγο και αποτελείται από τις οντότητες που υπάρχουν στον κόσμο και τις μεταξύ τους σχέσεις. Εν προκειμένω η οντολογία συνδέει τη σημασιολογία του πραγματικού κόσμου με την σημασιολογία που μπορεί να επεξεργαστεί από ηλεκτρονικούς υπολογιστές (Γιάου (2013, σ.14)).

Τα βασικά συστατικά στοιχεία μιας οντολογίας είναι τα εξής:

- Κλάσεις (*Classes*): Αποτελούν την οργάνωση εννοιών που ανήκουν σε ένα κοινό πεδίο βάσει συγκεκριμένων προδιαγραφών.
- Σχέσεις (*Relations*): Είναι ο τρόπος με τον οποίο συσχετίζονται οι έννοιες μεταξύ τους.
- Συναρτήσεις (*Functions*): Είναι μία ειδική σχέση κατά την οποία το ν-οστό στοιχείο μιας σχέσης μπορεί να προσδιοριστεί από τα ν-1 προηγούμενα στοιχεία.
- Αξιώματα (*Axioms*): Αφορούν καταστάσεις που θεωρούνται ως δεδομένες χωρίς να απαιτείται περαιτέρω διερεύνηση.
- Στιγμιότυπα (*Instances*): Είναι συγκεκριμένα μοναδικά στοιχεία μίας κλάσης (Γιάου (2013, σ.14, 15)).

3.2 Γλώσσες/Πρότυπα Οντολογιών

Το πρότυπο *XML* (*eXtensible Markup Language*) που αποτελεί ένα από τα βασικά πρότυπα ανταλλαγής δεδομένων μεταξύ πληροφοριακών συστημάτων μέσω του Διαδικτύου, αποτέλεσε τη βάση για την ανάπτυξη γλωσσών/πρότυπων για τη δομημένη κωδικοποίηση της σημασιολογίας των δεδομένων. Οι σημαντικότερες γλώσσες/πρότυπα που αναπτύχθηκαν είναι οι *RDF, RDFS* και *OWL*. Αποτελούν όλες προτάσεις του *W3C* προς την κατεύθυνση της ανάπτυξης του Σημασιολογικού Ιστού (Γιάου (2013, σ.17)).

3.3 RDF (Resource Description Framework)

Τα *RDF, RDFS* και *OWL* είναι οι βασικές γλώσσες/πρότυπα του Σημασιολογικού Ιστού. Το *RDF* αποτελεί το πλαίσιο πάνω στο οποίο στηρίζονται και οι άλλες δύο γλώσσες/πρότυπα. Η πιο σημαντική του λειτουργία είναι η διαχείριση κατανεμημένων δεδομένων που αποτελεί ένα βασικό θέμα στον Σημασιολογικό Ιστό. Χρησιμοποιεί σε μεγάλο βαθμό γνωστά χαρακτηριστικά της υποδομής του Παγκόσμιου Ιστού τα οποία και επεκτείνει δημιουργώντας έτσι μια βάση για ένα κατανεμημένο δίκτυο δεδομένων (Allemang, Hendler (2011, σ.27)).

Ο Παγκόσμιος Ιστός αποτελείται από ιστοσελίδες που περιέχουν δεδομένα και είναι συνδεδεμένες μεταξύ τους. Στον Σημασιολογικό Ιστό τα δεδομένα, η πληροφορία και οι γνώσεις σχετικά με οτιδήποτε στον κόσμο αναφέρονται ως πόροι. Το *RDF* που αποτελεί τη βασική τεχνολογία του Σημασιολογικού Ιστού περιέχει στην πλήρη ονομασία του τη λέξη πόρος καθώς αυτή είναι Πλαίσιο Περιγραφής Πόρων (***R**esource **D**escription **F**ramework*). Σε ένα δίκτυο πληροφοριών οποιοσδήποτε μπορεί να συνεισφέρει στη γνώση σχετικά με έναν πόρο. Άλλωστε το Διαδίκτυο αναπτύχθηκε με τέτοιο πρωτοφανή ρυθμό λόγω αυτού του γεγονότος. (Allemang, Hendler (2011, σ.27, 28)).

Μια πολύ συχνή μορφή αναπαράστασης των δεδομένων είναι σε μορφή πίνακα. Η αναπαράσταση αυτή χρησιμοποιείται από το *RDF* για τη δημιουργία του βασικού δομικού του στοιχείου που ονομάζεται τριπλέτα. Οι σειρές του πίνακα αναφέρονται στα στοιχεία που περιγράφονται ενώ οι στήλες αντίστοιχα αναφέρονται στις ιδιότητες αυτών των στοιχείων. Τα κελιά του πίνακα δίνουν τις συγκεκριμένες τιμές γι' αυτές τις ιδιότητες. Το βασικό δομικό στοιχείο του *RDF*, η τριπλέτα, δημιουργείται με βάση τα τρία αυτά στοιχεία. Το αναγνωριστικό για τη σειρά καλείται υποκείμενο της τριπλέτας, το αναγνωριστικό για τη στήλη καλείται κατηγορούμενο της τριπλέτας, ενώ η τιμή του κελιού ονομάζεται αντικείμενο της τριπλέτας. Ένα παράδειγμα μιας τριπλέτας του *RDF* είναι το εξής: *mov:TomCruise mov:PlayedIn mov:TopGun*. Όταν περισσότερες από μία τριπλέτες αναφέρονται στο ίδιο στοιχείο, μερικές φορές είναι βολικό να τις αναπαραστήσουμε ως κατευθυνόμενο γράφο στον οποίο κάθε τριπλέτα είναι μια ακμή από το υποκείμενο προς το αντικείμενο του, με το κατηγορούμενο ως ετικέτα της ακμής (Allemang, Hendler (2011, σ.28, 31)).

Το *RDF* χρησιμοποιείται ως ένας τρόπος διανομής δεδομένων από διάφορες πηγές. Για την αξιοποίηση των δεδομένων απαιτείται συγχώνευση αυτών των πηγών.

Αυτό αποτελεί μια σχετικά απλή διαδικασία. Όπως ειπώθηκε και παραπάνω οι τριπλέτες που είναι τα βασικά δομικά στοιχεία του *RDF* μπορούν να αναπαρασταθούν ως γράφοι, όπου η συγχώνευση πληροφορίας από δύο γράφους γίνεται με τον σχηματισμό ενός γράφου από τις τριπλέτες του κάθε μεμονωμένου γράφου σε συνδυασμό (Allemang, Hendler (2011, σ.32)).

Ένα βασικό πρόβλημα που προκύπτει από τη συγχώνευση των γράφων είναι το πότε ένας κόμβος σε ένα γράφο είναι ο ίδιος ως κόμβος σε έναν άλλο γράφο. Η επίλυση αυτού του ζητήματος γίνεται με τη χρήση των ενιαίων προσδιορισμών πόρων *URI* (***U****niform* ***R****esource* ***I****dentifier*) που αποτελούν θεμελιώδη τεχνολογία Παγκόσμιου Ιστού την οποία το *RDF* χρησιμοποιεί για τη επίλυση του συγκεκριμένου προβλήματος. Ο κάθε πόρος που υπάρχει στο Διαδίκτυο έχει ένα μοναδικό χαρακτηριστικό ταυτοποίησης που δίνεται μέσω του *URI*. Η σύνταξη και τη μορφή των *URIs* έχουν τη μορφή των πολύ γνωστών στους περισσότερους χρήστες του Διαδικτύου διευθύνσεων *URL* (***U****niform* ***R****esource Locator*) που στην ουσία αποτελούν μια υποπερίπτωση των *URIs*. Ο τρόπος με τον οποίο επιλύεται το πρόβλημα ταυτότητας που προκύπτει από τη συγχώνευση γράφων στο *RDF* είναι πως ένας κόμβος από έναν γράφο συγχωνεύεται με έναν κόμβο από άλλον γράφο, εάν έχουν το ίδιο *URI* (Allemang, Hendler (2011, σ.33-35)).

Η χρήση του *URI* ως πρότυπο παγκοσμίων αναγνωριστικών επιτρέπει μια παγκόσμια αναφορά για οποιοδήποτε σύμβολο. Ο οργανισμός τυποποίησης *W3C* χρησιμοποιεί αυτήν την ιδιότητα του *URI* για να προσδιορίσει την έννοια ορισμένων όρων στα πρότυπα τυποποίησης. Τα πρότυπα του *W3C* παρέχουν ορισμούς για όρους όπως *type, subClassOf, Class, inverseOf* κτλ. Η εφαρμογή των προτύπων αυτών έχει παγκόσμιο εύρος σε ολόκληρο τον Σημασιολογικό Ιστό, έτσι ώστε μέσω των *URIs* να αναφέρονται σε αυτές τις δεσμευμένες λέξεις με τον ίδιο τρόπο που αναφέρονται σε οποιοδήποτε άλλο πόρο του Σημασιολογικού Ιστού (Allemang, Hendler (2011, σ.37)).

Παρακάτω δίνονται οι τυπικοί χώροι ονομάτων του *W3C* για *RDF, RDFS, OWL*:

- ***rdf***: Υποδεικνύει τα αναγνωριστικά που χρησιμοποιούνται για το *RDF*. Το παγκόσμιο *URI* για το χώρο ονομάτων *rdf* είναι *http: //www.w3.org/1999/02/22-rdf-syntax-ns#*.
- ***rdfs***: Υποδεικνύει τα αναγνωριστικά που χρησιμοποιούνται για τη γλώσσα *RDF Schema, RDFS*. Το παγκόσμιο *URI* για το χώρο ονομάτων *rdfs* είναι *http://www.w3.org/2000/01/rdf-schema#*.

- ***owl***: Υποδεικνύει τα αναγνωριστικά που χρησιμοποιούνται για τη γλώσσα *Ontology Web Language*, *OWL*. Το παγκόσμιο *URI* για το χώρο ονομάτων *owl* είναι *http://www.w3.org/2002/07/owl#* (Allemang, Hendler (2011, σ.38)).

Τα παραπάνω *URIs* κάνουν εμφανή και την αντιστοίχιση μεταξύ ενός *URI* και μιας διεύθυνσης *URL*. Οποιοδήποτε *URI* σε έναν από αυτούς τους χώρους ονομάτων (π.χ. *http://www.w3.org/2000/01/rdf-schema#subclassOf*) αναφέρεται σε ένα συγκεκριμένο όρο, για τον οποίο το *W3C* κάνει σχετικές δηλώσεις, στο συγκεκριμένο παράδειγμα, στο πρότυπο *RDFS* (Allemang, Hendler (2011, σ.38)).

Η αναπαράσταση του *RDF* σε τριπλέτες της μορφής υποκείμενο-κατηγορούμενο-αντικείμενο ή η αναπαράστασή του με τη μορφή γράφου είναι χρήσιμες για την παρουσίαση του *RDF* σε βιβλία και έγγραφα, δεν είναι όμως κατάλληλες όταν θέλουμε να δημοσιεύσουμε δεδομένα σε *RDF* στον Παγκόσμιο Ιστό. Τα δεδομένα στον Παγκόσμιο Ιστό αναπαρίστανται σε *HTML* ή γενικότερα σε *XML*. Για το λόγο αυτό το *W3C* συνέστησε τη χρήση του *RDF* σε μια μορφή *XML* που ονομάζεται *RDF/ XML*. Παρακάτω βλέπουμε ένα παράδειγμα ενός τέτοιου αρχείου (Allemang, Hendler (2011, σ.44, 46)):

```
<rdf:RDF
    xmlns:mfg="http://www.WorkingOntologist.com/Examples/Chapter3/
    Manufacturing#"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntaxns#">
    <mfg:Product
        rdf:about="http://www.WorkingOntologist.com/Examples/Chapter3/
        Manufacturing#Product1">
        <mfg:Available>23</mfg:Available>
        <mfg:Division>Manufacturing support</mfg:Division>
        <mfg:ProductLine>Paper machine</mfg:ProductLine>
        <mfg:SKU>FB3524</mfg:SKU>
        <mfg:ModelNo>ZX-3</mfg:ModelNo>
        <mfg:ManufactureLocation>Sacramento</mfg:Manufacture
        Location>
    </mfg:Product>
    <mfg:Product
        rdf:about="http://www.WorkingOntologist.com/Examples/Chapter3/
```

```
        Manufacturing#Product2”>
        <mfg:SKU>KD5243</mfg:SKU>
        <mfg:Division>Manufacturing support</mfg:Division>
        <mfg:ManufactureLocation>Sacramento</mfg:Manufacture
        Location>
        <mfg:Available>4</mfg:Available>
        <mfg:ModelNo>ZX-3P</mfg:ModelNo>
        <mfg:ProductLine>Paper machine</mfg:ProductLine>
    </mfg:Product>
</rdf:RDF>
```

3.4 RDFS (Resource Description Framework Schema)

Η *RDFS* (*Resource Description Framework Schema*) όπως δηλώνεται και στο όνομά της είναι η γλώσσα σχήματος για το *RDF*. Χρησιμοποιεί τις τριπλέτες του *RDF* για την περιγραφή των κλάσεων, των συσχετίσεων μεταξύ κλάσεων (υποκλάσεις), των ιδιοτήτων αντικειμένων που προσδιορίζουν σχέσεις μεταξύ ατόμων των κλάσεων και των συσχετίσεων μεταξύ ιδιοτήτων (υπο-ιδιότητες). Η *RDFS* περιέχει ένα απλό και κομψό σχήμα για την κληρονομικότητα. Όλες οι πληροφορίες του σχήματος της *RDFS* (κλάσεις, υποκλάσεις, υπο-ιδιότητες, πεδίο ορισμού και πεδίο τιμών, κλπ) εκφράζονται σε τριπλέτες. Ένα παράδειγμα μιας τριπλέτας της *RDFS* είναι το εξής: *mov:Comedy rdfs:subClassOf mov:Movie*. (Allemang, Hendler (2011, σ.151)).

Ένα πολύ βασικό στοιχείο που καθορίζει τη σημασιολογία στην *RDFS* είναι ο μηχανισμός συμπερασμάτων. Ο συγκεκριμένος μηχανισμός ορίζει πως η έννοια κάθε δομικού στοιχείου της *RDFS* δίνεται από το σύνολο των συμπερασμάτων που μπορούν να εξαχθούν από αυτό. Μέσω αυτού του μηχανισμού ορίζονται οι έννοιες των υποκλάσεων και των υπο-ιδιοτήτων. Οι κανόνες του μηχανισμού συμπερασμάτων είναι απλοί αλλά οι δηλώσεις πολύ ισχυρές. Η *RDFS* επίσης περιέχει τα δομικά στοιχεία *rdfs:domain* και *rdfs:range* που περιγράφουν σχέσεις μεταξύ κλάσεων και ιδιοτήτων. (Allemang, Hendler (2011, σ.151)).

Παρότι η *RDFS* αποτελείται από ένα μικρό σύνολο δομικών στοιχείων και απλών κανόνων, χρησιμοποιώντας το μηχανισμό συμπερασμάτων μπορεί να επιλύσει ένα μεγάλο πλήθος ζητημάτων ενσωμάτωσης και συγχώνευσης διαφορετικών πηγών. Με αυτόν τον τρόπο μπορούμε να κρίνουμε αν τα αποτελέσματα είναι επιθυμητά ή όχι. Η συνδυασμένη χρήση των δομικών στοιχείων της *RDFS* είναι ιδιαίτερα

αποτελεσματική στον καθορισμό του πώς πληροφορίες με διαφορετική δομή μπορεί να χρησιμοποιηθούν μαζί με ομοιόμορφο τρόπο (Allemang, Hendler (2011, σ.151)).

3.5 OWL (Web Ontology Language)

Ο Σημασιολογικός Ιστός έχει ως βάση του την αναπαράσταση των δεδομένων μέσω του πλαισίου της *RDF* και μέσω των προσαρμοσμένων σχημάτων ετικετών της *XML*. Οι δυνατότητες που έχει η *RDF* εξαντλούνται στην περιγραφή μοντέλων δεδομένων, στην μεταξύ τους σχέση καθώς και στην αναπαράσταση των κλάσεων και ιδιοτήτων. Επίσης υπάρχει αδυναμία επιβολής σημασιολογικών περιορισμών μέσω της *XML*. Για την κάλυψη των παραπάνω αδυναμιών χρειάζεται μία οντολογική γλώσσα που θα περιγράφει τη σημασιολογία της πληροφορίας των σελίδων του Διαδικτύου καθώς και τις συσχετίσεις μεταξύ των διαφόρων όρων. (Γιάου (2013, σ.24)).

Τις παραπάνω απαιτήσεις ήρθε να καλύψει η *OWL* (*Web Ontology Language*). Πρόκειται για μια σημασιολογική γλώσσα σήμανσης και χρησιμοποιείται για τη δημιουργία και τη μεταφορά οντολογιών μέσω του Διαδικτύου. Αναπαριστά τις έννοιες όσο και τις μεταξύ τους συσχετίσεις και δίνει δυνατότητες επεξεργασίας του περιεχόμενου της πληροφορίας. Υποστηρίζει διατύπωση πλούσιας σημασιολογίας με επιπλέον λεξιλόγιο για την περιγραφή κλάσεων και ιδιοτήτων. Έχει ως βάση τις *RDF* και *RDFS* και έχει εκτεταμένες δυνατότητες σε σχέση με αυτές. Η σύνταξη της βασίζεται στην *XML* (Γιάου (2013, σ.24)).

Τα βασικότερα στοιχεία της γλώσσας είναι τα εξής:

- Κλάσεις (*Classes*): Αποτελούν τα βασικότερα συστατικά της γλώσσας. Κάθε άτομο που περιγράφεται με την *OWL* είναι μέλος της κλάσης *owl:Thing* και κάθε κλάση που δηλώνεται από έναν χρήστη είναι αυτόματα υποκλάση της *owl:Thing*. Υποστηρίζεται η ιεραρχία κλάσεων με κάθε υποκλάση να αποτελεί μία εξειδίκευση μίας γενικότερης κλάσης. Η δήλωση μίας κλάσης περιλαμβάνει την ονομασία της και μία σειρά από περιορισμούς που καθορίζουν τη μορφή των ατόμων της.
- Άτομα (*Individuals*): Ένα άτομο είναι μέλος μιας κλάσης και δηλώνεται σαν τέτοιο. Οι κλάσεις αντιστοιχούν στα σύνολα των στοιχείων του τομέα αναφοράς ενώ τα άτομα αντιστοιχούν στις πραγματικές οντότητες που μπορούν να ομαδοποιηθούν σε αυτές τις κλάσεις.
- Ιδιότητες (*Properties*): Οι ιδιότητες διακρίνονται σε δύο τύπους: τις ιδιότητες τύπου δεδομένων (*datatypes*) που προσδιορίζουν σχέσεις μεταξύ ατόμων κλάσεων και δεδομένων (*RDF literals, XML schema datatypes*) και τις ιδιότητες αντικειμένων

(*object properties*) που προσδιορίζουν σχέσεις μεταξύ ατόμων των κλάσεων. Καταγράφονται σαν σχέσεις ανάμεσα στις οποίες χρησιμοποιείται μία μέθοδος για τον περιορισμό τους. Η μέθοδος αυτή προσδιορίζεται από το πεδίο ορισμού (*domain*) και το πεδίο τιμών (*range*). Μια ιδιότητα μπορεί να αποτελεί και εξειδίκευση μιας άλλης ιδιότητας. Έστω για παράδειγμα μία κλάση Α με πεδίο ορισμού την κλάση Β και πεδίο τιμών την κλάση Γ. Η δήλωση αυτή οδηγεί στη συσχέτιση ατόμων της Β με άτομα της Γ (Γιάου (2013, σ.26, 27)).

Τα χαρακτηριστικά που μπορεί να έχει μια ιδιότητα είναι τα εξής:

- Μεταβατικότητα (*Transitive Property*): Αν μια ιδιότητα *P* είναι μεταβατική τότε για κάθε *x, y* και *z* θα ισχύει πάντα ότι αν *P(x, y)* και *P(y, z)* τότε και *P(x, z)*.
- Συμμετρικότητα (*Symmetric Property*): Αν μια ιδιότητα *P* είναι συμμετρική τότε για κάθε *x* και *y* θα ισχύει πάντα ότι *P(x, y)* αν και μόνο αν *P(y, x)*.
- Λειτουργικότητα (*Functional Property*): Αν μια ιδιότητα *P* είναι λειτουργική τότε για όλα τα *x, y* και *z* θα ισχύει πάντα ότι αν *P(x, y)* και *P(x, z)* τότε *y = z*.
- Αντιστροφή (*Inverse Of*): Αν μια ιδιότητα *P1* είναι αντίστροφη μιας *P2* τότε για όλα τα *x* και *y* θα είναι *P1(x, y)* αν και μόνο αν *P2(y, x)*.
- Αντίστροφη Λειτουργικότητα (*Inverse Functional Property*): Αν μια ιδιότητα *P* είναι αντίστροφα λειτουργική τότε για όλα τα *x, y, z* θα ισχύει ότι αν *P(y, x)* και *P(z, x)* θα είναι και *y = z* (Γιάου (2013, σ.27)).

Το πεδίο τιμών (*range*) των ιδιοτήτων μπορεί να περιοριστεί. Οι περιορισμοί καταγράφονται εντός του *owl:Restriction*. Οι τρόποι που μπορεί να γίνει αυτό είναι οι εξής:

- Όλες οι τιμές από (*All Values From*): Ορίζει ότι για κάθε άτομο μιας κλάσης, η οποία έχει άτομα μιας συγκεκριμένης ιδιότητας, όλες οι δυνατές τιμές αυτής της ιδιότητας αποτελούν μέλη της κλάσης που προσδιορίζεται από το *owl:allValuesFrom*.
- Μερικές τιμές από (*Some Values From*): Ορίζει ότι για κάθε άτομο μιας κλάσης, η οποία έχει άτομα μιας συγκεκριμένης ιδιότητας, η ιδιότητα αυτή θα έχει τιμή ένα ή περισσότερα μέλη της κλάσης που προσδιορίζεται από το *owl:allValuesFrom*.
- Πληθικότητα (*Cardinality*): Ο περιορισμός καθορίζει την πληθικότητα των στοιχείων σε ένα σύνολο. Με το *owl:maxCardinality* καθορίζεται το ανώτερο όριο πληθικότητας ενώ με το *owl:minCardinality* μπορούμε να καθορίσουμε το κατώτερο επιτρεπόμενο όριο.

- Έχει τιμή (*Has Value*): Χρησιμοποιείται για τον προσδιορισμό κλάσεων που βασίζονται σε μία συγκεκριμένη τιμή μιας ιδιότητας. Τα άτομα – μέλη μιας τέτοιας κλάσης έχουν τιμή στην ιδιότητα αυτή ίδια με αυτή του *hasValue* (Γιάου (2013, σ.27, 28)).

Πέρα από τα τυπικά χαρακτηριστικά η *OWL* εμπεριέχει μηχανισμούς επαύξησης της αποδοτικότητας στην ανάπτυξη των οντολογιών. Οι μηχανισμοί που χρησιμοποιούνται στην κατεύθυνση αυτή είναι:

- Ισοδυναμίες ανάμεσα σε κλάσεις και ιδιότητες: Υπάρχει η δυνατότητα να δημιουργηθεί μία οντολογία σαν σύνθεση δύο άλλων όταν υπάρχει κλάση ή ιδιότητα της μίας οντολογίας που έχει αντιστοιχία με κάποια της άλλης οντολογίας. Στις περιπτώσεις αυτές χρησιμοποιείται η ιδιότητα *owl:equivalentClass* για να δηλωθεί ότι δύο κλάσεις έχουν ακριβώς ίδια στιγμιότυπα και η *owl:equivalentProperty* αντίστοιχα για ιδιότητες.
- Διαφοροποίηση ατόμων: Προσδιορίζει τιμές που είναι αμοιβαία ξεχωριστές ώστε να δηλωθεί η διαφοροποίηση οντοτήτων. Έτσι η έκφραση *owl:AllDifferent* σε συνδυασμό με την *owl:distinctMembers* καθορίζει ένα σύνολο ατόμων που όλα είναι ξεχωριστά μεταξύ τους (Γιάου (2013, σ.28, 29)).

Οι μηχανισμοί αυτοί συμπληρώνονται από αντίστοιχους που χρησιμοποιούνται για τη σύνθεση οντολογιών. Έτσι διακρίνονται οι παρακάτω δυνατότητες:

- Τελεστές συνόλων όπως είναι οι: Τομή (*Intersection*), Ένωση (*Union*), Συμπλήρωμα (*Complement*).
- Απαριθμημένες Κλάσεις (*Enumerated Classes*): Υπάρχει η δυνατότητα καθορισμού μίας κλάσης με την παράθεση των μελών της.
- Κλάσεις με μη κοινά στοιχεία (*Disjoint Classes*): Για τη δήλωση συνόλων κλάσεων με κανένα κοινό στοιχείο χρησιμοποιείται το *owl:disjointWith* για να δηλωθεί ότι κανένα άτομο που αποτελεί μέλος μιας κλάσης δεν μπορεί την ίδια στιγμή να αποτελεί και άτομο μιας άλλης κλάσης (Γιάου (2013, σ.29)).

Τέλος υπάρχει η περίπτωση διαφορετικών εκδόσεων μίας οντολογίας όπου μια οντολογία μπορεί να εξελιχθεί και να εμπλουτιστεί σε σχέση με μια προηγούμενη μορφή της. Η σύνδεση με την προηγούμενη έκδοση γίνεται μέσω της ιδιότητας *owl:priorVersion*. Οι ιδιότητες *owl:backwardCompatibleWith* και *owl:incompatibleWith* έχουν να κάνουν με τη συμβατότητα μεταξύ των εκδόσεων. Πληροφορίες σχετικά με την

παρούσα έκδοση της οντολογίας δίνονται μέσω της ιδιότητας *owl:versionInfo* (Γιάου (2013, σ.29)).

Αναλυτικότερες πληροφορίες σχετικά με τις γλώσσες/πρότυπα *RDF, RDFS, OWL* ξεφεύγουν από τους σκοπούς της συγκεκριμένης εργασίας και θα πρέπει να αναζητηθούν σε αντίστοιχα συγγράμματα.

4 Το περιβάλλον οντολογιών Protégé

4.1 Περιγραφή

Το *Protégé* είναι ένα πλούσιο σε λειτουργίες περιβάλλον οντολογιών με πλήρη υποστήριξη της *OWL*. Υποστηρίζει τη δημιουργία και την επεξεργασία μιας ή περισσοτέρων οντολογιών σε έναν ενιαίο χώρο εργασίας μέσω ενός πλήρως προσαρμόσιμου περιβάλλοντος χρήστη. Τα εργαλεία απεικόνισης επιτρέπουν τη διαλογική πλοήγηση των σχέσεων οντολογίας. Η προηγμένη υποστήριξη επεξήγησης βοηθά στην ανίχνευση ασυνεπειών. Διαθέτει λειτουργίες που περιλαμβάνουν συγχώνευση οντολογιών, μετακίνηση αξιωμάτων μεταξύ οντολογιών, μετονομασία πολλαπλών οντοτήτων και πολλά άλλα. Η αρχιτεκτονική του *Protégé* μπορεί να προσαρμοστεί για να χτίσει τόσο απλές όσο και σύνθετες εφαρμογές βασιζόμενες σε οντολογίες. Οι προγραμματιστές μπορούν να ενσωματώσουν την έξοδο του *Protégé* με συστήματα κανόνων ή άλλους μηχανισμούς επίλυσης προβλημάτων για να κατασκευάσουν ένα ευρύ φάσμα έξυπνων συστημάτων.

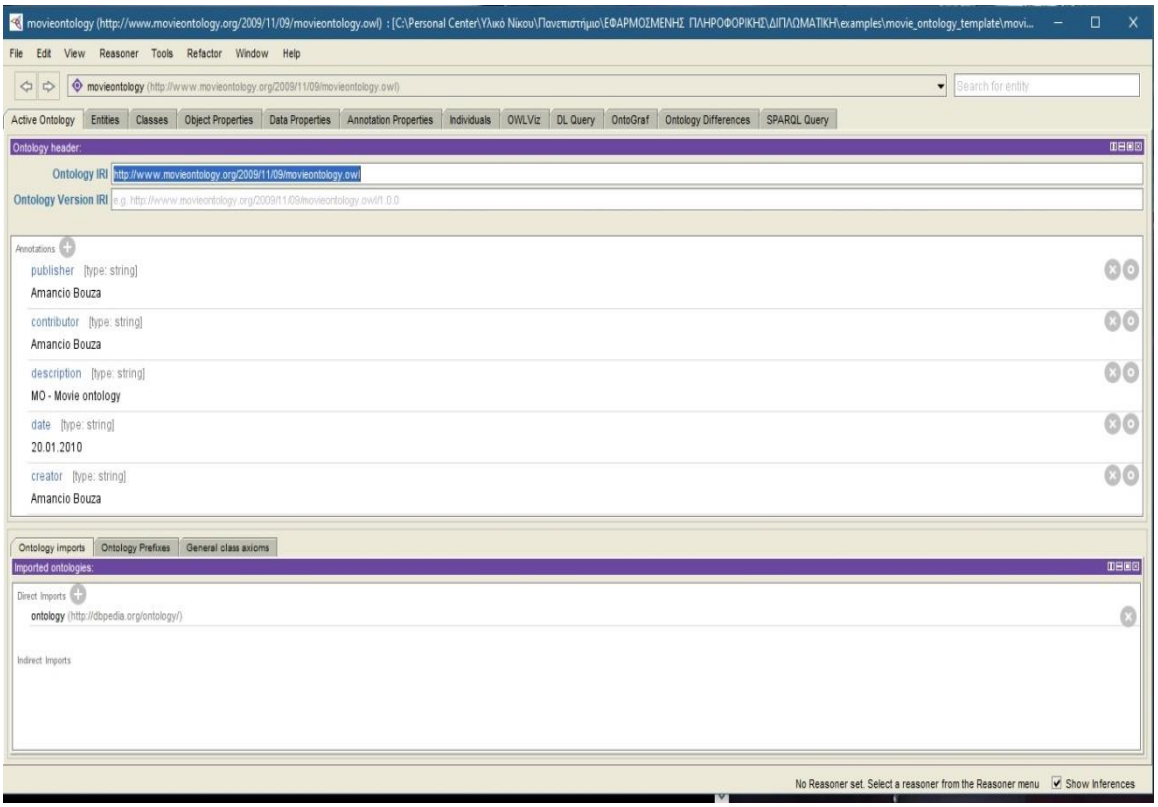
4.2 Περιβάλλον εργασίας

Το περιβάλλον εργασίας του *Protégé* είναι οργανωμένο σε καρτέλες εντός ενός κοινού παραθύρου. Οι καρτέλες είναι οι εξής:

- ***Active Ontology***
- ***Entities***
- ***Classes***
- ***Object Properties***
- ***Data Properties***
- ***Annotation Properties***
- ***Individuals***
- ***OWL Viz***
- ***DL Query***
- ***Onto Graf***
- ***Ontology Differences***
- ***SPARQL Query***

Στη συνέχεια δίνεται μια σύντομη περιγραφή για τις πιο βασικές από τις παραπάνω καρτέλες. Έχει επιλεγεί ως παράδειγμα ένα πρότυπο οντολογίας σχετικά με κινηματογραφικές ταινίες (Amancio Bouza (2010)).

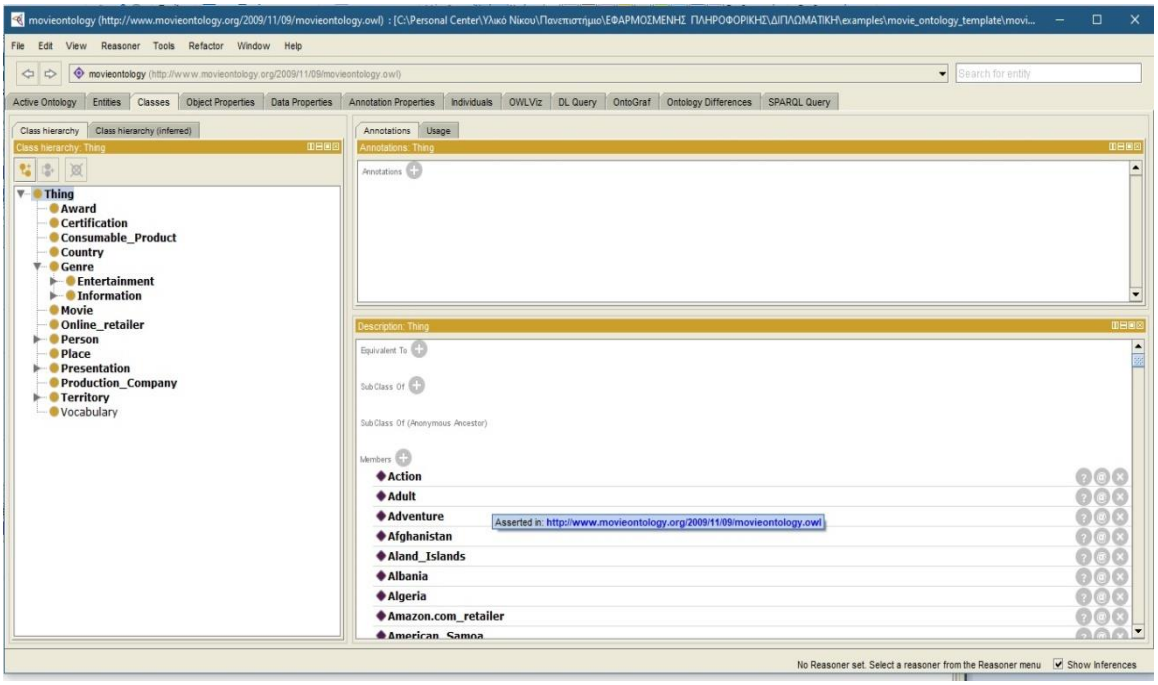
Αρχικά η καρτέλα "*Active Ontology*" επιτρέπει να οριστούν πληροφορίες σχετικά με την οντολογία. Εδώ ορίζεται το *URI* της οντολογίας, μπορούν να προστεθούν και να επεξεργαστούν σχόλια σχετικά με την οντολογία, και οι χώροι ονομάτων μπορούν να ρυθμιστούν μέσω αυτής της καρτέλας. Η καρτέλα "*Active Ontology*" φαίνεται στην Εικόνα 4-1 που ακολουθεί (Matthew Horridge (2011, σ.13)).



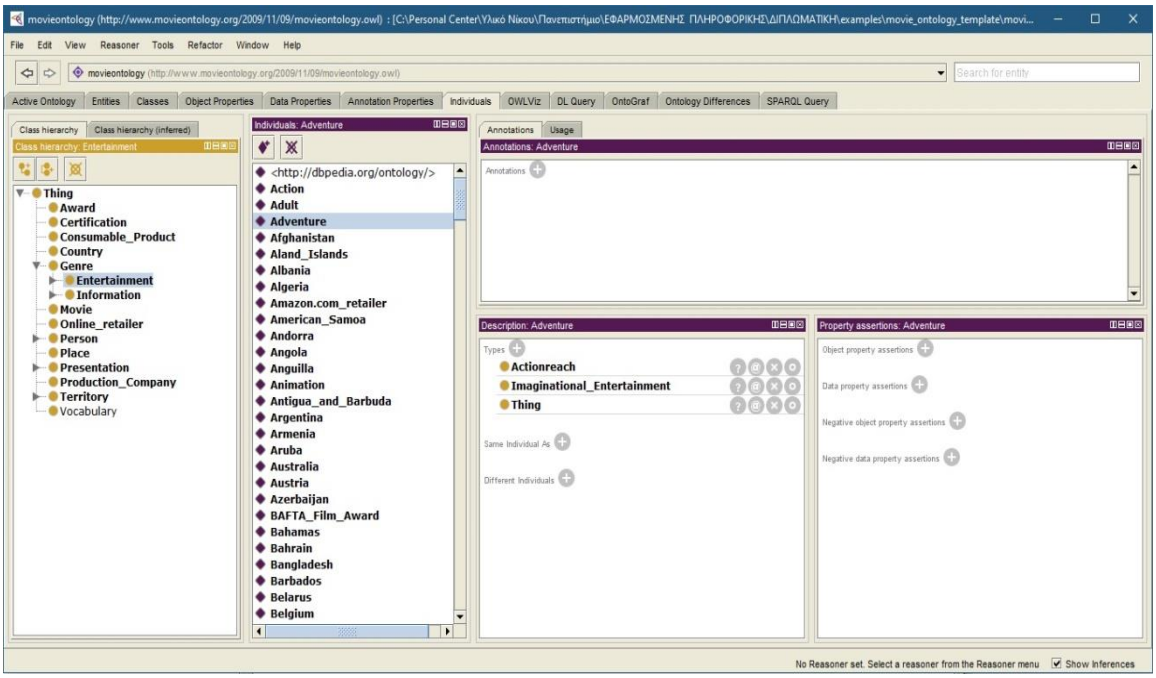
Εικόνα 4-1: Η καρτέλα "Active Ontology"

Όπως αναφέρθηκε στο προηγούμενο κεφάλαιο τα κύρια δομικά στοιχεία μιας οντολογίας *OWL* είναι οι κλάσεις. Στο *Protégé*, η δημιουργία και η επεξεργασία των κλάσεων πραγματοποιείται χρησιμοποιώντας την καρτέλα "*Classes*" που εμφανίζεται παρακάτω στην Εικόνα 4-2. Στην αριστερή πλευρά φαίνεται η ιεραρχική δομή των κλάσεων της οντολογίας. Μια κενή οντολογία περιέχει μια κλάση που ονομάζεται *Thing*. Όλες οι κλάσεις είναι υποκατηγορίες της κλάσεως *Thing* (Matthew Horridge (2011, σ.15)).

Όπως επίσης αναφέρθηκε στην ανάλυση των βασικών στοιχείων της *OWL* τα μέλη μια κλάσης ονομάζονται άτομα. Η *OWL* επιτρέπει να ορίζουμε άτομα και να αναθέτουμε ιδιότητες σχετικά με αυτά. Για τη δημιουργία και την επεξεργασία ατόμων στο *Protégé* χρησιμοποιείται η καρτέλα "*Individuals*" που φαίνεται πιο κάτω στην Εικόνα 4-3 (Matthew Horridge (2011, σ.89)).



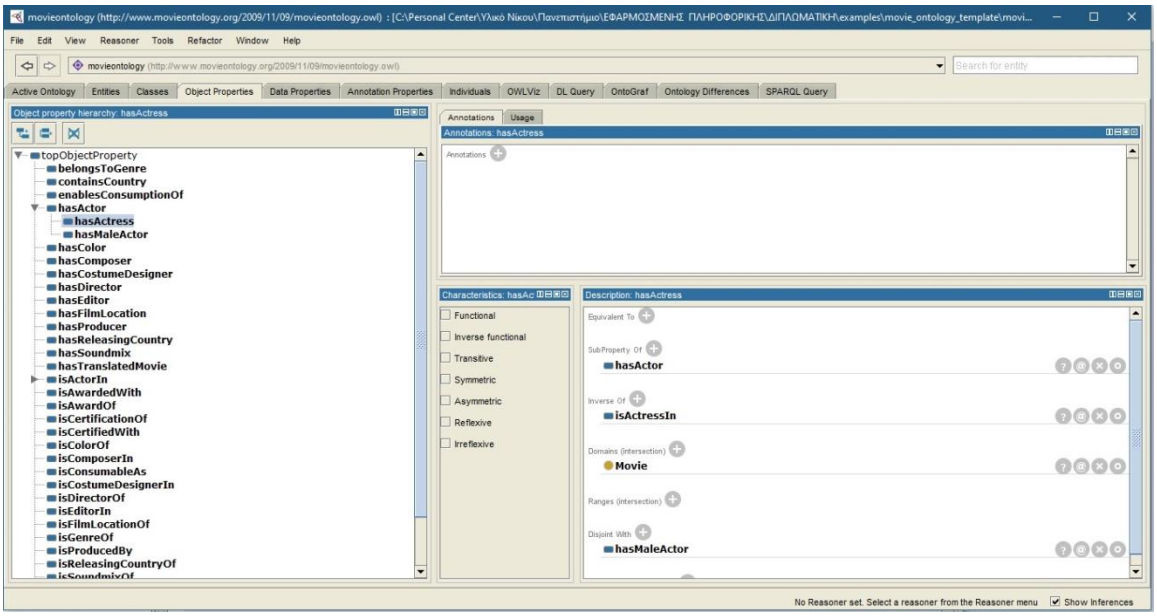
Εικόνα 4-2: Η καρτέλα "Classes"



Εικόνα 4-3: Η καρτέλα "Individuals"

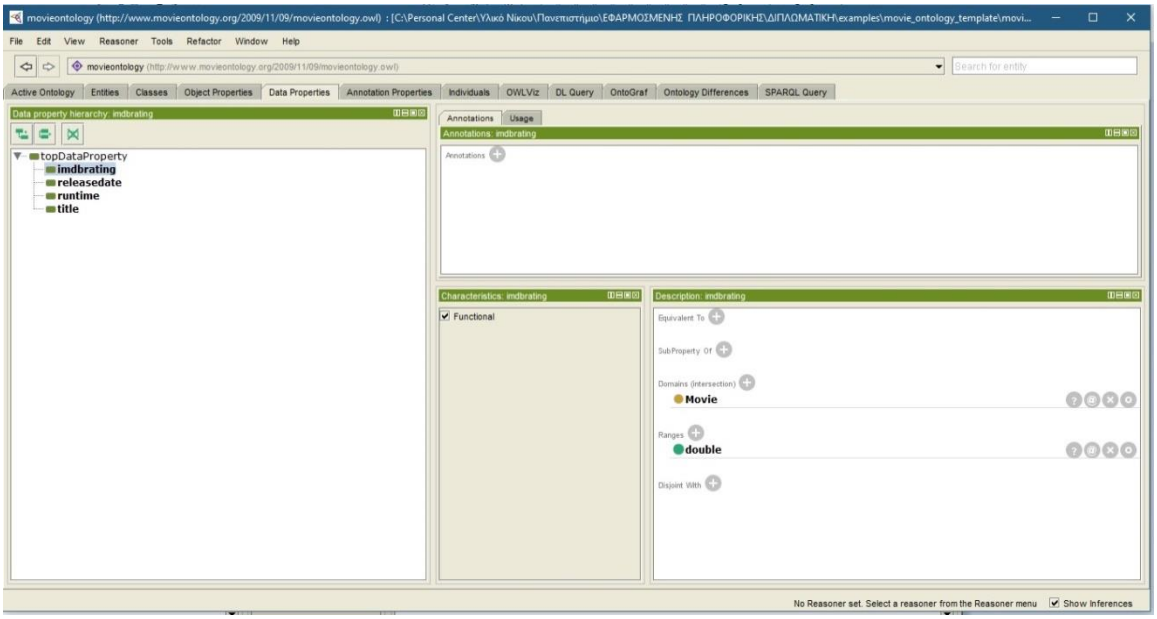
Οι ιδιότητες της *OWL* αντιπροσωπεύουν σχέσεις. Υπάρχουν τρεις τύποι ιδιοτήτων: οι ιδιότητες αντικειμένων, οι ιδιότητες τύπου δεδομένων και οι ιδιότητες σχολιασμού. Όπως αναφέρθηκε και στο προηγούμενο κεφάλαιο, οι ιδιότητες αντικειμένων είναι σχέσεις μεταξύ δύο ατόμων. Συνδέουν ένα άτομο μια κλάσης με ένα άλλο άτομο μιας κλάσης. Η δημιουργία και η επεξεργασία των ιδιοτήτων αντικειμένων

πραγματοποιείται χρησιμοποιώντας την καρτέλα "*Object Properties*" που εμφανίζεται παρακάτω στην Εικόνα 4-4 (Matthew Horridge (2011, σ.23)).



Εικόνα 4-4: Η καρτέλα "Object Properties"

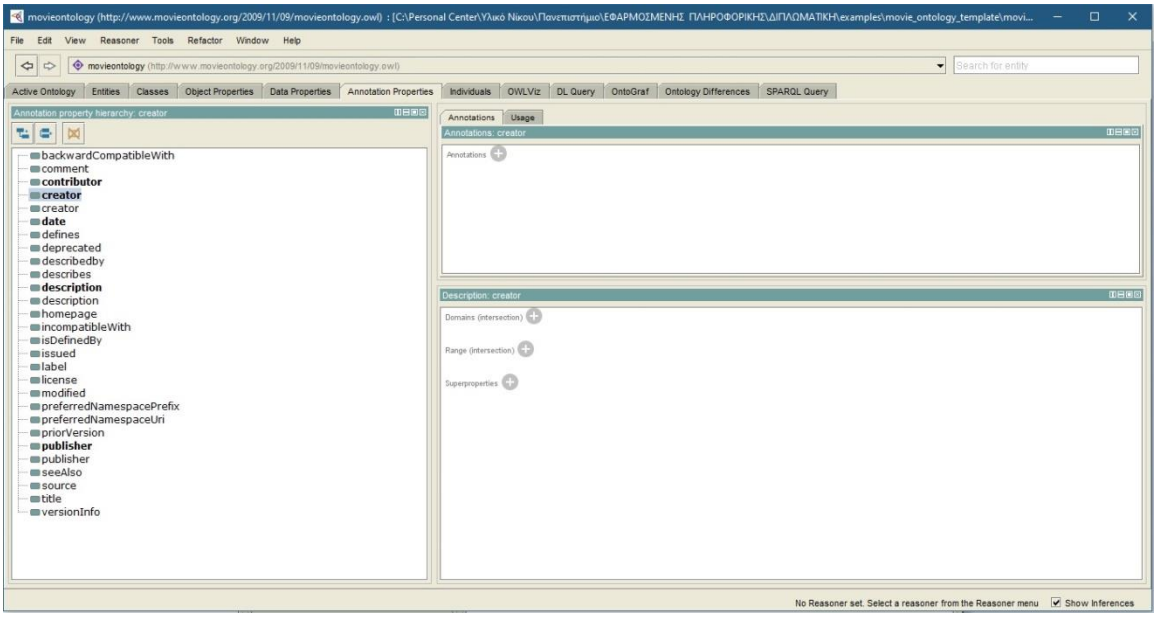
Οι ιδιότητες τύπου δεδομένων συνδέουν ένα άτομο μιας κλάσης με μια τιμή *XML Schema Datatype* ή ένα *rdf literal*. Περιγράφουν δηλαδή τις σχέσεις μεταξύ ατόμων και τιμών δεδομένων. Για τη δημιουργία και την επεξεργασία των ιδιοτήτων τύπου δεδομένων χρησιμοποιείται η καρτέλα "*Data Properties*" που φαίνεται πιο κάτω στην Εικόνα 4-5 (Matthew Horridge (2011, σ.76)).



Εικόνα 4-5: Η καρτέλα "Data Properties"

Οι ιδιότητες σχολιασμού μπορούν να χρησιμοποιηθούν για την προσθήκη πληροφοριών (μεταδεδομένα - δεδομένα σχετικά με τα δεδομένα) σε κλάσεις, άτομα,

ιδιότητες αντικειμένων/τύπων δεδομένων και την ίδια την οντολογία (τεχνικά την επικεφαλίδα οντολογίας). Αυτές οι πληροφορίες μπορούν να λάβουν τη μορφή ελεγκτικών ή συντακτικών πληροφοριών. Για παράδειγμα, σχόλια, ημερομηνία δημιουργίας, συγγραφέας ή αναφορές σε πόρους. Η διαχείριση των ιδιοτήτων σχολιασμού πραγματοποιείται χρησιμοποιώντας την καρτέλα "*Annotation Properties*" που εμφανίζεται παρακάτω στην Εικόνα 4-6 (Matthew Horridge (2011, σ.94)).



Εικόνα 4-6: Η καρτέλα "Annotation Properties"

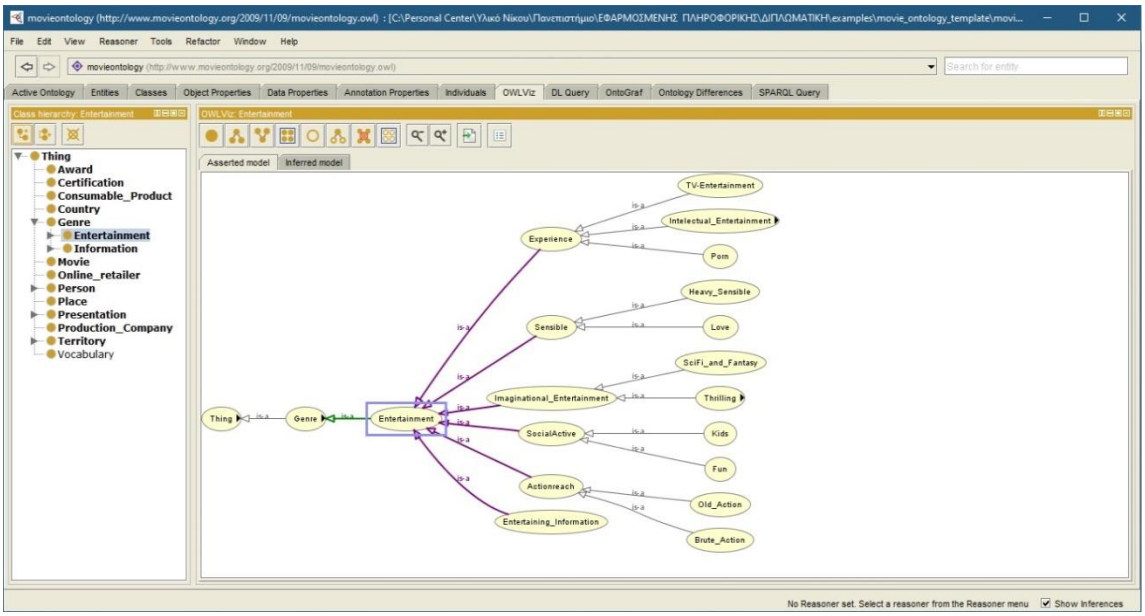
4.3 Το εργαλείο Reasoner

Ένα από τα βασικά εργαλεία στο *Protégé* είναι το εργαλείο *Reasoner*. Στο *Protégé* η ιεραρχία των κλάσεων που κατασκευάζεται με χειροκίνητο τρόπο ονομάζεται δηλωμένη ιεραρχία. Η ιεραρχία των κλάσεων που υπολογίζεται αυτόματα από το εργαλείο *Reasoner* ονομάζεται συναγόμενη ιεραρχία. Μια από τις κύριες υπηρεσίες που προσφέρει το εργαλείο, είναι να ελέγξει εάν μια κλάση είναι ή όχι υποκλάση μιας άλλης κλάσης. Με τον τρόπο αυτό υπολογίζεται η συναγόμενη ιεραρχία κλάσεων της οντολογίας. Μια άλλη υπηρεσία που προσφέρει είναι ο έλεγχος συνέπειας. Με βάση την περιγραφή μιας κλάσης μπορεί να ελεγχθεί εάν είναι ή όχι δυνατό για την κλάση να έχει ένα ή περισσότερα άτομα. Μια κλάση θεωρείται ασυνεπής αν δεν είναι δυνατόν να έχει κανένα άτομο. (Matthew Horridge (2011, σ.48)).

Μέσω της καρτέλα "*OWL Viz*" δίνεται η δυνατότητα μιας οπτικής απεικόνισης της ιεραρχίας των κλάσεων, με διαλογική πλοήγηση των σχέσεων οντολογίας,

επιτρέποντας τη σύγκριση της δηλωμένης ιεραρχίας κλάσεων και της συναγόμενης ιεραρχίας. Η καρτέλα "*OWL Viz*" εμφανίζεται παρακάτω στην Εικόνα 4-7.

Πιο αναλυτική περιγραφή σχετικά με το περιβάλλον οντολογιών *Protégé* ξεφεύγει από τους σκοπούς της συγκεκριμένης εργασίας και θα πρέπει να αναζητηθεί στο εγχειρίδιο της εφαρμογής.



Εικόνα 4-7: Η καρτέλα "OWL Viz"

5 Η εφαρμογή εκμάθησης οντολογιών Text2Onto

5.1 Σύντομη εισαγωγή

Όπως είδαμε και σε προηγούμενα κεφάλαια οι οντολογίες αποτελούν μια τεχνολογία μοντελοποίησης δεδομένων προσανατολισμένη στη σημασιολογία επάνω σε έναν τομέα ενδιαφέροντος, με σκοπό να εξυπηρετηθούν οι ολοένα αυξανόμενες ανάγκες ανταλλαγής γνώσεων και ενοποίησης. Βασικό στοιχείο της τεχνολογίας μιας οντολογίας είναι πως καθίσταται μηχανικά επεξεργάσιμη επιτρέποντας την ανταλλαγή των δεδομένων της, τα οποία σχολιάζονται σημασιολογικά, μεταξύ διαφορετικών εφαρμογών. Μια οντολογία μπορεί να αναπαριστά γνώση προερχόμενη από διάφορα είδη δεδομένων. Όλα τα παραπάνω κάνουν την οικοδόμηση μιας οντολογίας ενός μεγάλου όγκου δεδομένων από τους ειδικούς του τομέα των οντολογιών μια αρκετά δύσκολη, επίπονη και χρονοβόρα διαδικασία. Για την υποστήριξη του χρήστη στην κατασκευή των οντολογιών έχουν αναπτυχθεί διάφορα εργαλεία όπως το *TextToOnto*, το σύστημα *ASIUM*, το *Mo'k Workbench*, το *OntoLearn* ή το *OntoLT*. Ωστόσο, όλα τα παραπάνω εργαλεία έχουν αρκετές αδυναμίες (Philip Cimiano, Johanna Volker (2005, σ.2)).

Στον Σημασιολογικό Ιστό σήμερα μπορεί κάποιος να συναντήσει διάφορες γλώσσες οντολογικής αναπαράστασης. Μερικές από τις πιο διαδεδομένες οντολογικές γλώσσες είναι οι *RDFS*, *OWL* και *F-Logic*. Ένα από τα βασικά προβλήματα των οντολογικών εργαλείων που αναφέρθηκαν παραπάνω είναι πως το καθένα είναι προσανατολισμένο προς ένα συγκεκριμένο οντολογικό μοντέλο που δεν μπορεί εύκολα να μετατραπεί εύκολα σε άλλα οντολογικά μοντέλα. Αυτό είναι ένα μειονέκτημα, καθώς τα εργαλεία εκμάθησης οντολογιών θα πρέπει να είναι ανεξάρτητα από ένα συγκεκριμένο μοντέλο οντολογίας προκειμένου να είναι ευρέως εφαρμόσιμα και χρησιμοποιούμενα. Η εφαρμογή *Text2Onto* σχεδιάστηκε με στόχο την υπέρβαση του συγκεκριμένου προβλήματος, αναπαριστώντας τις οντολογικές δομές όχι σε μια συγκεκριμένη γλώσσα οντολογίας αλλά με τη μορφή των αποκαλούμενων θεμελιακών στοιχείων μοντελοποίησης. Στη συνέχεια αυτά τα θεμελιακά στοιχεία μπορούν να μεταφραστούν μέσω συγκεκριμένου μηχανισμού της εφαρμογής σε οποιαδήποτε από τις οντολογικές γλώσσες που αναφέρθηκαν πιο πάνω (Philip Cimiano, Johanna Volker (2005, σ.3)).

Μια επιπλέον αδυναμία των εργαλείων εκμάθησης οντολογιών που αναπτύχθηκαν αρχικά είναι πως δεν αναπτύχθηκε ιδιαίτερα η αλληλεπίδραση με τους τελικούς χρήστες μέσω μιας εύχρηστης διεπαφής χρήστη κάνοντας αυτά τα εργαλεία εύχρηστα μόνο τους γλωσσολόγους ή τους ειδικούς της εκμάθησης μηχανής. Τέλος τα εργαλεία αυτά παρουσιάζονται μη αποδοτικά όσον αφορά τις αλλαγές που γίνονται στο υποκείμενο σώμα των δεδομένων εισόδου καθώς μόλις αυτό αλλάξει θα πρέπει να επανεξετάσουν πλήρως την οντολογία που έχει δημιουργηθεί (Philip Cimiano, Johanna Volker (2005, σ.3)).

Το *Text2Onto* είναι ένας πλήρης επανασχεδιασμός της εφαρμογής *TextToOnto*, μιας σουίτας εργαλείων για την εκμάθηση οντολογιών από αδόμητο κείμενο. Το *TextToOnto* υλοποιεί ποικίλους αλγορίθμους για διαφορετικές εργασίες εκμάθησης οντολογιών, όπως την εξόρυξη όρων, την ταξινομική κατασκευή καθώς και τις σχέσεις μεταξύ εννοιών. Ο σχεδιασμός του *TextToOnto* έδωσε βάση στην αλγοριθμική ραχοκοκαλιά με αποτέλεσμα η αλληλεπίδραση με τον χρήστη να έχει περάσει σε δεύτερη μοίρα και να απουσιάζει μια εύχρηστη διεπαφή (Johanna Volker, York Sure (2005, σ.23)).

Η δημιουργία της εφαρμογής *Text2Onto* είχε ως στόχο την επίλυση των παραπάνω προβλημάτων. Καταρχάς η εφαρμογή παρέχει στον χρήστη ένα εύχρηστο περιβάλλον εργασίας. Η αρχιτεκτονική του *Text2Onto* έχει ως θεμέλιο ένα μοντέλο που ονομάζεται Πιθανοτικό Οντολογικό Μοντέλο (***P**robabilistic **O**ntology **M**odel* - *POM*). Το *POM* εξάγει τα αποτελέσματα της οντολογικής εκμάθησης προσδίδοντάς τους μια τιμή πιθανότητας. Επίσης μέσω της τεχνικής της ανακάλυψης αλλαγών οδηγούμενων από τα δεδομένα, η οποία είναι υπεύθυνη για τον εντοπισμό αλλαγών στο σώμα των δεδομένων εισόδου, το *Text2Onto* τροποποιεί τα οντολογικά αποτελέσματα του Πιθανοτικού Οντολογικού Μοντέλου χωρίς να επαναϋπολογιστεί η οντολογία, για ολόκληρη τη συλλογή των δεδομένων εισόδου (Philip Cimiano, Johanna Volker (2005, σ.3)).

Η ανάθεση τιμών πιθανοτήτων στα αποτελέσματα βελτιώνει τη διαδικασία της οντολογικής εκμάθησης παρουσιάζοντας στον χρήστη τα αποτελέσματα ταξινομημένα σύμφωνα με τη βεβαιότητα του συστήματος. Επίσης για καθένα από τα στοιχεία των οντολογικών αποτελεσμάτων του Πιθανοτικού Οντολογικού Μοντέλου αποθηκεύεται ένας δείκτης που δίνει την αντιστοίχιση με τα τμήματα της συλλογής δεδομένων εισόδου από τα οποία προέρχεται το στοιχείο, βελτιώνοντας έτσι τη δυνατότητα κατανόησης της

δημιουργίας μιας έννοιας, στιγμιότυπου ή σχέσης στην προκύπτουσα οντολογία (Philip Cimiano, Johanna Volker (2005, σ.3)).

Μέσω της τεχνικής της ανακάλυψης αλλαγών οδηγούμενων από τα δεδομένα η αποδοτικότητα της εφαρμογής αυξάνεται καθώς δεν υπάρχει ανάγκη επεξεργασίας ολόκληρης της συλλογής των δεδομένων εισόδου όταν αυτή αλλάζει. Επίσης υπάρχει η δυνατότητα παρακολούθησης των αλλαγών της οντολογίας μετά την τελευταία αλλαγή στη συλλογή των δεδομένων εισόδου και έτσι ο χρήστης έχει τη δυνατότητα ανίχνευσης της εξέλιξης της οντολογίας σε σχέση με αλλαγές που πραγματοποιούνται στο υποκείμενο σώμα των δεδομένων εισόδου (Philip Cimiano, Johanna Volker (2005, σ.3)).

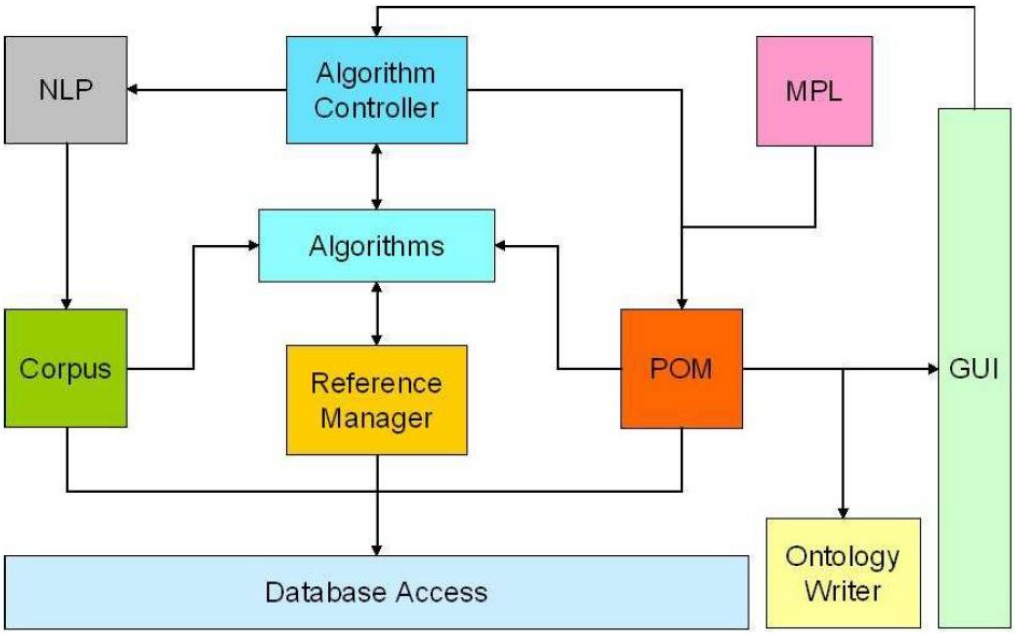
5.2 Η αρχιτεκτονική του Text2Onto

Όπως ειπώθηκε και στην προηγούμενη ενότητα, το βασικό θεμελιακό στοιχείο της αρχιτεκτονικής του *Text2Onto* είναι το Πιθανοτικό Οντολογικό Μοντέλο (*POM*) το οποίο είναι υπεύθυνο για την εξαγωγή και αποθήκευση των αποτελεσμάτων των διαφορετικών αλγορίθμων εκμάθησης οντολογιών. Ένα επίσης πολύ βασικό εργαλείο της εφαρμογής είναι ο ελεγκτής αλγορίθμων (*Algorithm Controller*), ο σκοπός του οποίου είναι α) η ενεργοποίηση της γλωσσικής προεπεξεργασίας των δεδομένων (*Natural Language Processing - NLP*), β) η εκτέλεση των αλγορίθμων οντολογικής εκμάθησης και γ) η εφαρμογή των αιτημάτων αλλαγής των αλγορίθμων στο *POM* (Johanna Volker, York Sure (2005, σ.16)).

Η εκτέλεση κάθε αλγορίθμου αποτελείται από τρία στάδια. Σε πρώτη φάση ο αλγόριθμος ενημερώνεται για τις πρόσφατες αλλαγές στο σώμα των δεδομένων κειμένου (*Corpus*). Στη συνέχεια, σε δεύτερη φάση αυτές οι αλλαγές μεταβιβάζονται στην αποθήκη αναφοράς που αποθηκεύει εγγραφές που αφορούν τη σχέση μεταξύ της οντολογίας και των δεδομένων εισόδου (π.χ. δείκτες σε όλες τις εμφανίσεις μιας έννοιας). Τέλος, σε τρίτη φάση, τα αιτήματα για αλλαγές στο *POM* παράγονται από το ενημερωμένο περιεχόμενο της αποθήκης αναφοράς (*Reference Manager - Database Access*) (Johanna Volker, York Sure (2005, σ.16)).

Οι αλγόριθμοι της εφαρμογής ταξινομούνται σύμφωνα με δύο διαφορετικές παραμέτρους. Ο πρώτος τρόπος ταξινόμησης είναι σύμφωνα με το είδος των θεμελιακών στοιχείων μοντελοποίησης που παράγουν ενώ ο δεύτερος σύμφωνα με τη μέθοδο που χρησιμοποιείται για την εξαγωγή οντολογικών στιγμιότυπων των σχετικών θεμελιακών στοιχείων από το κείμενο. Οι δεδομένοι αλγόριθμοι για το καθένα είδος θεμελιακών στοιχείων μοντελοποίησης (*Modeling Primitive Library – MPL*), μπορούν να

παραμετροποιηθούν ώστε να συνδυαστούν τα αποτελέσματά τους προκειμένου να αποκτηθεί μια πιο αξιόπιστη πιθανότητα για κάθε θεμελιακό στοιχείο. Η αρχιτεκτονική του *Text2Onto* φαίνεται σχηματικά στην εικόνα 5-1 που ακολουθεί. Τα βασικά συστατικά στοιχεία είναι η μηχανή γλωσσικής προεπεξεργασίας των δεδομένων (*NLP engine*), οι αλγόριθμοι (*Algorithms*), ο ελεγκτής αλγορίθμων (*Algorithm Controller*) και το Πιθανοτικό Οντολογικό Μοντέλο (*POM*) (Johanna Volker, York Sure (2005, σ.16)).



Εικόνα 5-1: Η αρχιτεκτονική του Text2Onto

5.3 Το Πιθανοτικό Οντολογικό Μοντέλο

Το Πιθανοτικό Οντολογικό Μοντέλο (*POM*) του *Text2Onto* εξάγει τα αποτελέσματα οντολογικής εκμάθησης με τη μορφή συγκεκριμένων θεμελιακών στοιχείων μοντελοποίησης. Τα θεμελιακά στοιχεία μοντελοποίησης ορίζονται από τη βιβλιοθήκη θεμελιακών στοιχείων μοντελοποίησης (*Modeling Primitive Library – MPL*) του *Text2Onto* και είναι ανεξάρτητα από μια συγκεκριμένη γλώσσα αναπαράστασης οντολογίας. Η συγκεκριμένη μοντελοποίηση είναι ευέλικτη και επεκτάσιμη καθώς η προσθήκη νέων θεμελιακών στοιχείων δεν επηρεάζει το υπάρχον πλαίσιο. Τα θεμελιακά στοιχεία μοντελοποίησης μπορούν να μεταφραστούν σε διάφορες γλώσσες οντολογικής αναπαράστασης όπως *RDFS, OWL* και *F-Logic* μέσω των επονομαζόμενων συγγραφέων οντολογίας (*Ontology Writers*) (Philip Cimiano, Johanna Volker (2005, σ.5)).

Τα θεμελιακά στοιχεία μοντελοποίησης που χρησιμοποιούνται στο *Text2Onto* φαίνονται στον πίνακα 5-1 που ακολουθεί (Peter Haase, Johanna Volker (2005, σ.5)).

Θεμελιακό στοιχείο	**Επεξήγηση**	***OWL***
Concept	Μια έννοια (κλάση) *C*. Παράδειγμα: *movie, actor*	*C*
Instance	Ένα στιγμιότυπο *a*. Παράδειγμα: *tom cruise, hannibal*	*a*
Similarity	Ισοδυναμία μεταξύ των εννοιών *C1* και *C2*. Παράδειγμα: *similarity (film, movie)*	$C1 \equiv C2$
SubclassOf	Κληρονομικότητα κλάσεων *C1, C2*. Παράδειγμα: *subclassof (comedy, movie)*	$C1 \subset C2$
InstanceOf	Στιγμιότυπο έννοιας *a, C*. Παράδειγμα: *instanceof(tom cruise, actor)*	*C(a)*
Relation	Μια σχέση *R* μεταξύ των εννοιών *C1* και *C2*. Παράδειγμα: *hasDirector(movie, director)*	$C1 \subset \forall R.C2$
SubtopicOf	Επιμέρους θέμα *C1, C2* Παράδειγμα: *subtopicof(horror_movie, movie)*	$C1 \subset C2$

Πίνακας 5-1: Θεμελιακά στοιχεία του Πιθανοτικού Οντολογικού Μοντέλου

5.4 Ανακάλυψη αλλαγών οδηγούμενων από τα δεδομένα

Νωρίτερα στην εισαγωγή του παρόντος κεφαλαίου έγινε αναφορά εκτός από το Πιθανοτικό Οντολογικό Μοντέλο και σε ένα δεύτερο τρόπο οντολογικής εκμάθησης που εισάγει το *Text2Onto*, την τεχνική ανακάλυψης αλλαγών οδηγούμενων από τα δεδομένα. Οι οδηγούμενες από τα δεδομένα αλλαγές δημιουργούνται από τροποποιήσεις στο υποκείμενο σώμα δεδομένων εισόδου αναπαριστώντας τις γνώσεις που διαμορφώνονται από μια οντολογία. Επομένως, η ανακάλυψη αλλαγών οδηγούμενων από τα δεδομένα παρέχει μεθόδους για την αυτόματη ή την ημιαυτόματη προσαρμογή μιας οντολογίας σύμφωνα με τις τροποποιήσεις που εφαρμόζονται στο υποκείμενο σύνολο δεδομένων (Philip Cimiano, Johanna Volker (2005, σ.6)).

Ένα βασικό όφελος της τεχνικής ανακάλυψης αλλαγών οδηγούμενων από τα δεδομένα είναι πως δίνεται η δυνατότητα στον χρήστη να παρακολουθεί τις αλλαγές που προκύπτουν στην οντολογία από την τελευταία αλλαγή στη συλλογή δεδομένων εισόδου, ώστε να μπορεί να εντοπίσει την εξέλιξη της οντολογίας σε σχέση με τις

αλλαγές στο υποκείμενο σώμα δεδομένων εισόδου. Επίσης, δεν υπάρχει πλέον η ανάγκη επεξεργασίας ολόκληρης της συλλογής εγγράφων όταν αυτή αλλάζει, οδηγώντας έτσι σε αυξημένη αποδοτικότητα (Philip Cimiano, Johanna Volker (2005, σ.6)).

5.5 Γλωσσική προεπεξεργασία των δεδομένων

Όπως αναφέρθηκε και στο κεφάλαιο σχετικά με την αρχιτεκτονική του *Text2Onto* η πρώτη λειτουργία του ελεγκτή αλγορίθμων (*Algorithm Controller*) είναι η ενεργοποίηση της γλωσσικής προεπεξεργασίας των δεδομένων (*Natural Language Processing - NLP*). Η εξαγωγή οντολογιών από αδόμητο κείμενο σε πολλά από τα υπάρχοντα περιβάλλοντα εκμάθησης οντολογιών πραγματοποιείται είτε μέσω τεχνικών μάθησης μηχανής είτε μέσω γλωσσικής ανάλυσης. Το *Text2Onto* συνδυάζει προσεγγίσεις μάθησης μηχανής με βασική γλωσσική επεξεργασία, όπως διαχωρισμός λέξεων και προτάσεων ή λημματοποίηση και ρηχή γραμματική ανάλυση. Στη βασική του έκδοση το *Text2Onto* υποστηρίζει την εκμάθηση οντολογιών μόνο από κείμενα σε αγγλική γλώσσα. Υπάρχει και η έκδοση της εφαρμογής που υποστηρίζει εκμάθηση οντολογιών από κείμενα και στην Ισπανική γλώσσα (Philip Cimiano, Johanna Volker (2005, σ.7, 8)).

5.6 Αλγόριθμοι

Σε αυτή την ενότητα αναφέρονται εν συντομία οι αλγόριθμοι που χρησιμοποιούνται για την εξαγωγή των θεμελιακών στοιχείων μοντελοποίησης από το σώμα των δεδομένων κειμένου. Οι αλγόριθμοι χρησιμοποιούν ορισμένα κριτήρια για την αξιολόγηση της πιθανότητας των εξαγόμενων όρων. Μετά την εκτέλεση των αλγορίθμων για κάθε όρο, σε κάθε θεμελιακό στοιχείο, οι τιμές των μετρήσεων κανονικοποιούνται σε [0 ... 1] και χρησιμοποιούνται ως αντίστοιχη πιθανότητα στο *POM*. Οι αλγόριθμοι ανά θεμελιακό στοιχείο είναι οι εξής (Johanna Volker, York Sure (2005, σ.20)):

- ***Concept***
  - ***EntropyConceptExtraction***
  - ***ExampleConceptExtraction***
  - ***RTFConceptExtraction***
  - ***TFIDFConceptExtraction***
- ***Instance***
  - ***ExampleInstanceExtraction***

- ***TFIDFInstanceExtraction***

- ***Similarity***
  - ***ContextSimilarityExtraction***
- ***SubclassOf***
  - ***PatternConceptClassification***
  - ***VerticalRelationsConceptClassification***
  - ***WordNetConceptClassification***
- ***InstanceOf***
  - ***ContextInstanceClassification***
  - ***PatternInstanceClassification***
- ***Relation***
  - ***SubcatRelationExtraction***
- ***SubtopicOf***
  - ***SubtopicOfRelationConversion***
  - ***SubtopicOfRelationExtraction***

Ξεκινώντας από το θεμελιακό στοιχείο *Concep*t, ο αλγόριθμος *EntropyConceptExtraction* υπολογίζει την εντροπία που είναι ένας συνδυασμός του *C-value* (δείκτης συσχέτισης ανάμεσα σε μια μονάδα και στο σύνολο, μεθόδου βασισμένη στη συχνότητα, ευαίσθητη σε όρους πολλαπλών λέξεων) και του *NC-value* (Δείκτες ενσωμάτωσης πληροφοριών από τα συμφραζόμενα που δείχνουν τη συσχέτιση ανάμεσα σε μια μονάδα και στο σύνολο), (Sonam Mittal, Nupur Mittal (2013, σ.6)).

Ο αλγόριθμος *RTFConceptExtraction* υπολογίζει τη σχετική συχνότητα όρου (*Relative Term Frequency*) που λαμβάνεται διαιρώντας την απόλυτη συχνότητα όρου (πόσες φορές εμφανίζεται ένας όρος ‘τ’ στο έγγραφο ‘δ’) του όρου ‘τ’ στο έγγραφο ‘δ’, με τη μέγιστη απόλυτη συχνότητα όρου (ο αριθμός των φορών που οποιοσδήποτε όρος εμφανίζεται το μέγιστο αριθμό φορών στο έγγραφο ‘δ’) του εγγράφου ‘δ’ (Sonam Mittal, Nupur Mittal (2013, σ.6)).

Ο αλγόριθμος *TFIDFConceptExtraction* εξάγει τις πιο σημαντικές έννοιες σε σχέση με ένα δεδομένο τομέα ενδιαφέροντος. Η συνάφεια κάθε έννοιας καθορίζεται με τον υπολογισμό της μέσης τιμής $TF*IDF$ σε σχέση με ολόκληρη τη συλλογή εγγράφων (Johanna Volker, Denny Vrandecic, York Sure (2005, σ.14)). Ο όρος *TF* (*Term Frequency*) δίνει τη συχνότητα των όρων στη συλλογή των εγγράφων, ενώ ο όρος *IDF* (*Inverse Document Frequency*) λαμβάνεται διαιρώντας τον συνολικό αριθμό εγγράφων

με τον αριθμό των εγγράφων που περιέχουν τον όρο και στη συνέχεια τη λήψη του λογαρίθμου αυτού του πηλίκου (Sonam Mittal, Nupur Mittal (2013, σ.6)).

Συνεχίζοντας με το θεμελιακό στοιχείο *Instance*, για τον αλγόριθμο *TFIDFInstanceExtraction* ισχύουν τα αντίστοιχα με προηγουμένως. Ο αλγόριθμος εξάγει τα πιο σημαντικά στιγμιότυπα σε σχέση με ένα δεδομένο τομέα ενδιαφέροντος. Η συνάφεια κάθε στιγμιότυπου καθορίζεται με τον υπολογισμό της μέσης τιμής *TF∗IDF* σε σχέση με ολόκληρη τη συλλογή εγγράφων (Johanna Volker, Denny Vrandecic, York Sure (2005, σ.11)).

Για το θεμελιακό στοιχείο *Similarity*, ο μοναδικός διαθέσιμος αλγόριθμος είναι ο *ContextSimilarityExtraction*. Ακολουθώντας την υπόθεση ότι οι όροι ή οι έννοιες είναι ισοδύναμες, στο βαθμό στον οποίο μοιράζονται παρόμοια συντακτικά πλαίσια, ο αλγόριθμος υπολογίζει την ομοιότητα μεταξύ των όρων με βάση τα χαρακτηριστικά σχετικά με τα συμφραζόμενα, που εξάγονται από το σώμα των εγγράφων, σύμφωνα με τα οποία, το πλαίσιο των όρων ποικίλλει από απλές λέξεις έως γλωσσικά χαρακτηριστικά που εξάγονται με την τεχνική του *shallow parsing*, η οποία χρησιμοποιείται στην επεξεργασία φυσικής γλώσσας. Αυτή η ομοιότητα με βάση το σώμα εγγράφων λαμβάνεται ως πιθανότητα για την ισοδυναμία των εν λόγω εννοιών (Johanna Volker, York Sure (2005, σ.22)).

Ως προς το θεμελιακό στοιχείο *SubclassOf*, ο αλγόριθμος *PatternConceptClassification* βασίζεται στο ταίριασμα με ένα προκαθορισμένο σύνολο προτύπων, ανεξάρτητο από τον τομέα ενδιαφέροντος. Ο αλγόριθμος *VerticalRelationsConceptClassification* εφαρμόζει μια απλή γλωσσική ευρετική. Τέλος ο αλγόριθμος *WordNetConceptClassification* αξιοποιεί τη δομή της εφαρμογής *WordNet*, με στόχο τη λήψη αποδείξεων για σχέσεις κλάσης-υποκλάσης (Johanna Volker, Denny Vrandecic, York Sure (2005, σ.11)).

Ακολούθως για το θεμελιακό στοιχείο *InstanceOf*, ο αλγόριθμος *ContextInstanceClassification* προκειμένου να αντιστοιχίσει στιγμιότυπα που εμφανίζονται στο σώμα των εγγράφων με τη σωστή έννοια της οντολογίας, χρησιμοποιεί μια προσέγγιση που βασίζεται στην ομοιότητα, εξάγοντας διανύσματα σχετικά με τα συμφραζόμενα για στιγμιότυπα και έννοιες από τη συλλογή κειμένων και εκχωρώντας στιγμιότυπα στις έννοιες που αντιστοιχούν στο διάνυσμα της υψηλότερης ομοιότητας σε σχέση με το δικό τους διάνυσμα (Johanna Volker, York Sure (2005, σ.22)). Ο αλγόριθμος *PatternInstanceClassification* βασίζεται στο ταίριασμα με ένα

προκαθορισμένο σύνολο προτύπων, ανεξάρτητο από τον τομέα ενδιαφέροντος (Johanna Volker, Denny Vrandecic, York Sure (2005, σ.12)).

Σε ό, τι έχει να κάνει με το θεμελιακό στοιχείο *Relation*, ο μοναδικός διαθέσιμος αλγόριθμος είναι ο *SubcatRelationExtraction*. Ο συγκεκριμένος αλγόριθμος χρησιμοποιεί μια γλωσσική προσέγγιση. Χρησιμοποιώντας την τεχνική του *shallow parsing*, που χρησιμοποιείται στην επεξεργασία φυσικής γλώσσας, λαμβάνονται συντακτικά πλαίσια, βάσει των οποίων για κάθε φράση ρήματος, βρίσκει το υποκείμενο, το αντικείμενο και τη σχετική τους θέση (φιλτράροντας τα ουσιαστικά και τα ρήματα από την πρόταση) και στη συνέχεια τα αφαιρεί από την πρόταση και ετοιμάζει τη σχέση (Sonam Mittal, Nupur Mittal (2013, σ.6)).

Τέλος για το θεμελιακό στοιχείο *SubtopicOf,* ο αλγόριθμος *SubtopicOfRelationConversion* δημιουργεί σχέσεις υπο-θέματος λαμβάνοντας υπόψη ταξινομικές και μη ταξινομικές σχέσεις που είχαν προηγουμένως εξαχθεί από το σώμα εγγράφων. Ο αλγόριθμος *SubtopicOfRelationExtraction* εξετάζει την εμφάνιση εννοιών στο σώμα. Υποθέτει ότι ένα υπο-θέμα τείνει να εμφανίζεται σε ένα υποσύνολο αυτών των εγγράφων που σχετίζονται με το θέμα στο οποίο υπάγεται (Johanna Volker, Denny Vrandecic, York Sure (2005, σ.12)).

5.7 Εγκατάσταση και παραμετροποίηση

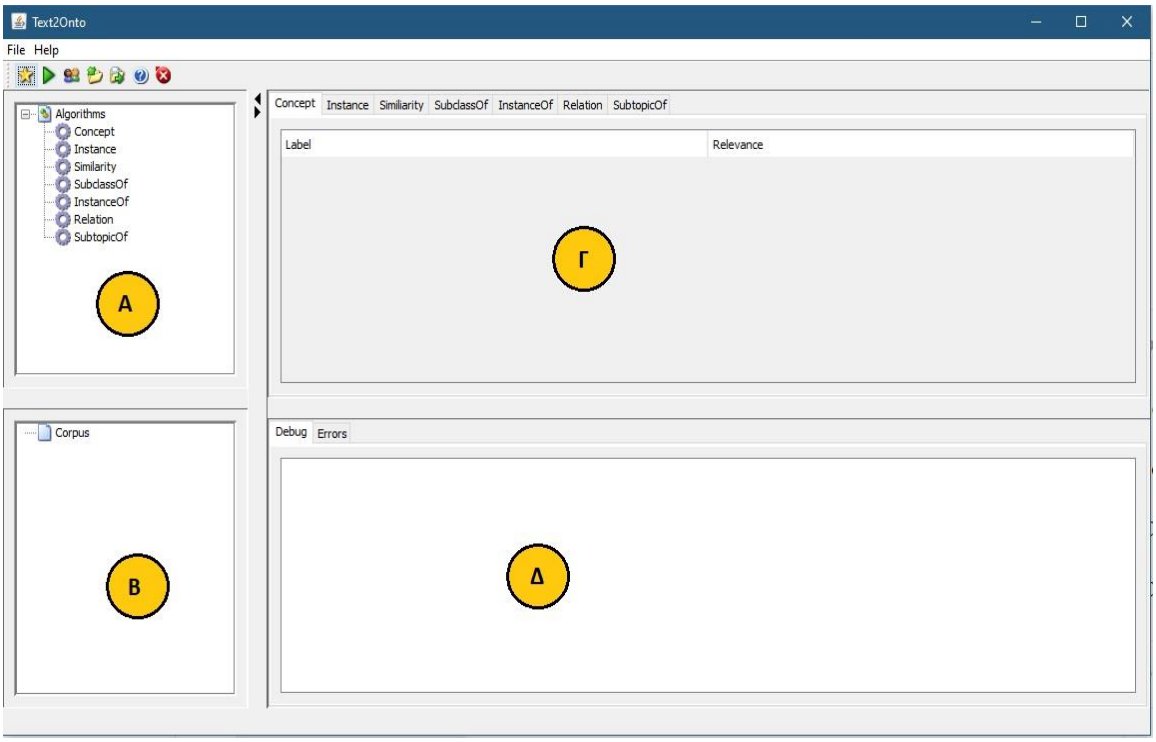
Σε αυτή την ενότητα δίνονται εν συντομία τα βήματα που χρειάστηκαν για την εγκατάσταση και την παραμετροποίηση της εφαρμογής *Text2Onto*. Τα βήματα που ακολουθήθηκαν είναι τα εξής:

1. Μέσω του συνδέσμου <*https://code.google.com/archive/p/text2onto/downloads*> επιλέχθηκε η έκδοση 17-09-2007. Αρχικά είχε επιλεγεί η έκδοση 09-11-2007 αλλά παρατηρήθηκε μετά τις πρώτες χρήσεις καθώς και μετά από σχετική αναζήτηση ότι η συγκεκριμένη έκδοση παρουσιάζει πρόβλημα ως προς την εξαγωγή των οντολογιών σε αρχεία *owl*, γι' αυτό και τελικά επιλέχθηκε η πρώτη έκδοση που δεν παρουσίασε το εν λόγω πρόβλημα.
2. Εγκατάσταση της εφαρμογής *GATE* έκδοση 4.0 και όχι νεώτερης έκδοσης καθώς δεν συνεργάζονται σωστά με το *Text2Onto*. <*https://gate.ac.uk/download/*>
3. Εγκατάσταση της εφαρμογής *WordNet* έκδοση 2.0 και όχι νεώτερης έκδοσης καθώς δεν συνεργάζονται σωστά με το *Text2Onto*.
   <*https://wordnet.princeton.edu/download/old-versions*>

4. Εγκατάσταση της *Java 6*. Παρατηρήθηκε ότι η εφαρμογή λειτουργεί σωστά μόνο με τις εκδόσεις 5 και 6 της *Java* και όχι με παλαιότερες ή νεώτερες εκδόσεις της.
5. Εγκατάσταση του *Text2Onto* και παραμετροποίηση συγκεκριμένων αρχείων σύμφωνα με τις οδηγίες που υπάρχουν εσωτερικά στο πακέτο εγκατάστασης. Χρήσιμες φάνηκαν και οι οδηγίες στους παρακάτω συνδέσμους. <*https://www.youtube.com/watch?v=B9tXt6y-fLo*> <*https://ryadyo.wordpress.com/2012/02/16/things-to-remember-while-installing-text2onto/*>

5.8 Το περιβάλλον εργασίας

Το περιβάλλον εργασίας του *Text2Onto* αποτελείται από διαφορετικά τμήματα παραμετροποίησης της διαδικασίας εκμάθησης οντολογιών και παρουσίασης των αποτελεσμάτων. Παρακάτω στην εικόνα 5-2 φαίνονται τα βασικά τμήματα του περιβάλλοντος εργασίας της εφαρμογής.

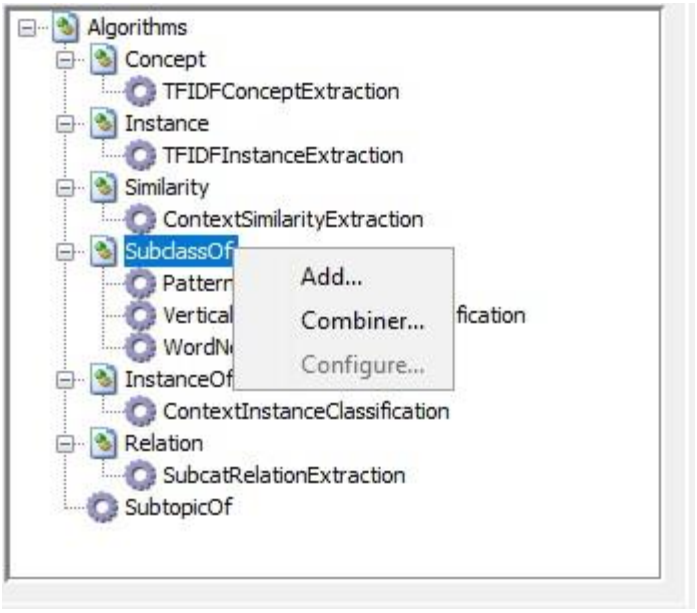


Εικόνα 5-2: Το περιβάλλον εργασίας του Text2Onto

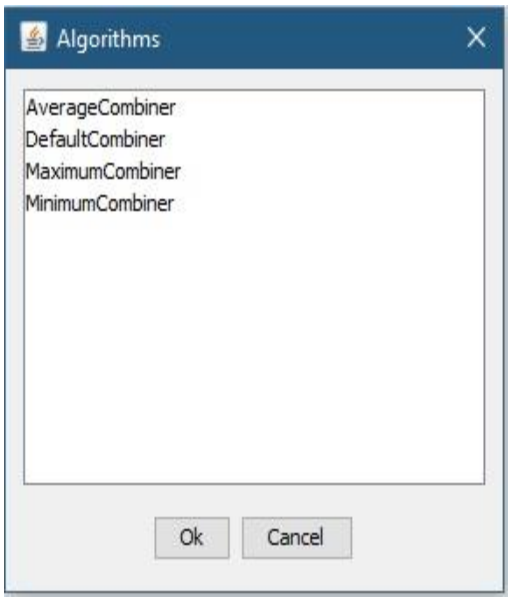
Επάνω αριστερά (Α) υπάρχει το τμήμα του ελεγκτή αλγορίθμων, το οποίο μπορεί να χρησιμοποιηθεί για τη ρύθμιση μιας ροής εργασίας επιλέγοντας κατάλληλους αλγόριθμους για τις διάφορες εργασίες εκμάθησης οντολογιών. Ο χρήστης μπορεί να επιλέξει ανάμεσα σε έναν αριθμό προκαθορισμένων στρατηγικών για το συνδυασμό των αποτελεσμάτων αυτών των αλγορίθμων, όπως φαίνεται και στην εικόνα 5-3. Σε

περίπτωση που επιλεγεί ως στρατηγική ένας συνδυασμός αλγορίθμων υπάρχουν οι επιλογές συνδυαστών αλγορίθμων που φαίνονται στην εικόνα 5-4 (*Average Combiner*, *Default Combiner*, *Maximum Combiner*, *Minimum Combiner*). Για παράδειγμα αν ο χρήστης κάνει την επιλογή *Average Combiner* κάθε φορά που ένα συγκεκριμένο οντολογικό στοιχείο εξάγεται ταυτόχρονα από διαφορετικούς αλγόριθμους, αυτό το είδος συνδυασμού ενημερώνει την τιμή της πιθανότητας για το συγκεκριμένο στοιχείο, με τον μέσο όρο από τις τιμές πιθανότητας που υπολογίζονται από όλους τους αλγόριθμους.

Στην κάτω αριστερή γωνία (B) βρίσκεται το τμήμα της υποκείμενης συλλογής εγγράφων το οποίο επιτρέπει στο χρήστη να δημιουργήσει ένα σώμα καθορίζοντας τα έγγραφα κειμένου από τα οποία πρέπει να εξαχθεί η οντολογία. Τα έγγραφα μπορεί να είναι σε μορφή απλού κειμένου ή και *HTML* και *PDF* έγγραφα. Ο χρήστης με δεξί κλικ επάνω στην ένδειξη *Corpus* μπορεί να εισάγει όσα έγγραφα επιθυμεί, να τα διαγράψει καθώς και να τα εμφανίσει σε μορφή αναδυόμενου παραθύρου.



Εικόνα 5-3: Το τμήμα ελεγκτή αλγορίθμων



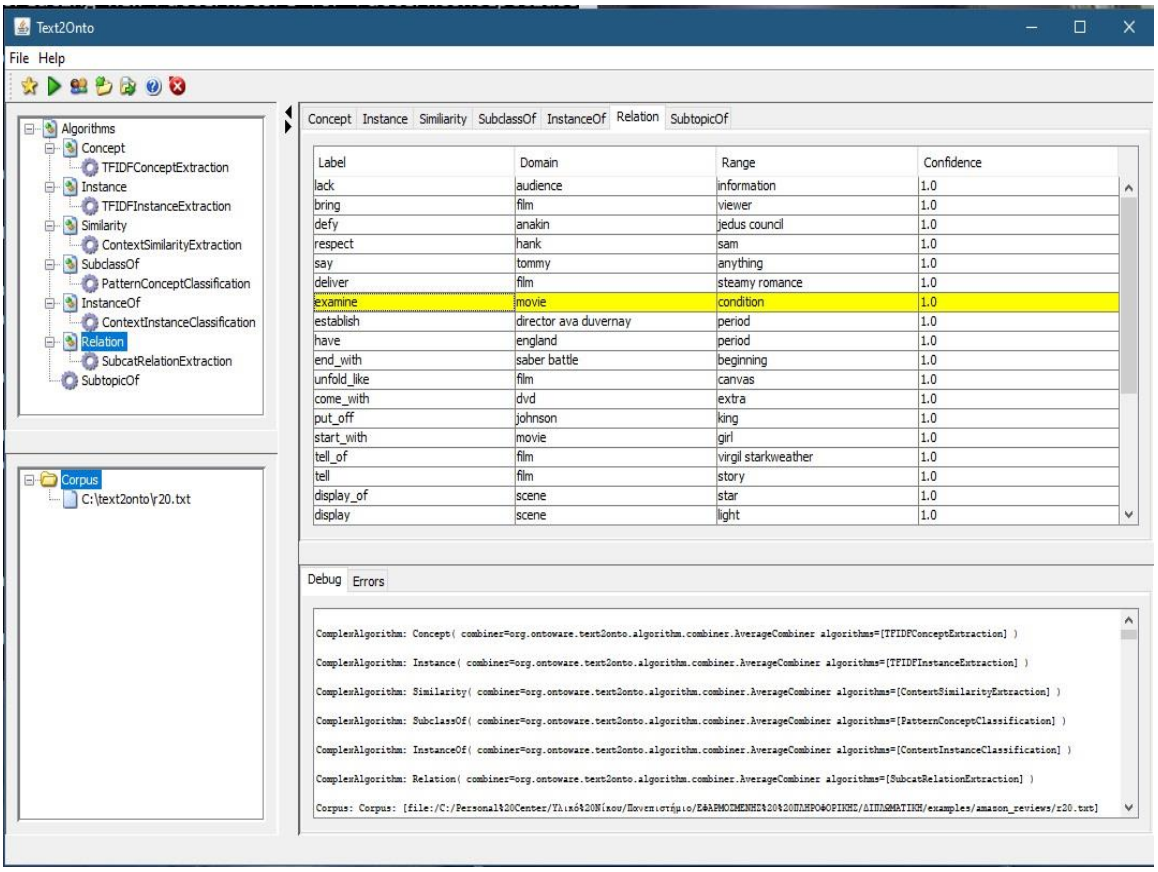
Εικόνα 5-4: Οι συνδυαστές αλγορίθμων

Επάνω δεξιά (Γ) βρίσκεται το τμήμα του Πιθανοτικού Οντολογικού Μοντέλου (*POM*), όπου και φαίνονται τα αποτελέσματα της τρέχουσας διαδικασίας εκμάθησης οντολογίας. Υπάρχουν διαφορετικές καρτέλες, μία για κάθε τύπο θεμελιακού στοιχείου μοντελοποίησης που εξάγονται από το υποκείμενο σώμα των εγγράφων. Τέλος, κάτω από το τμήμα των αποτελεσμάτων (Δ) υπάρχει το τμήμα για τον εντοπισμό σφαλμάτων εξόδου των αποτελεσμάτων και μηνυμάτων σφάλματος του περιβάλλοντος εργασίας.

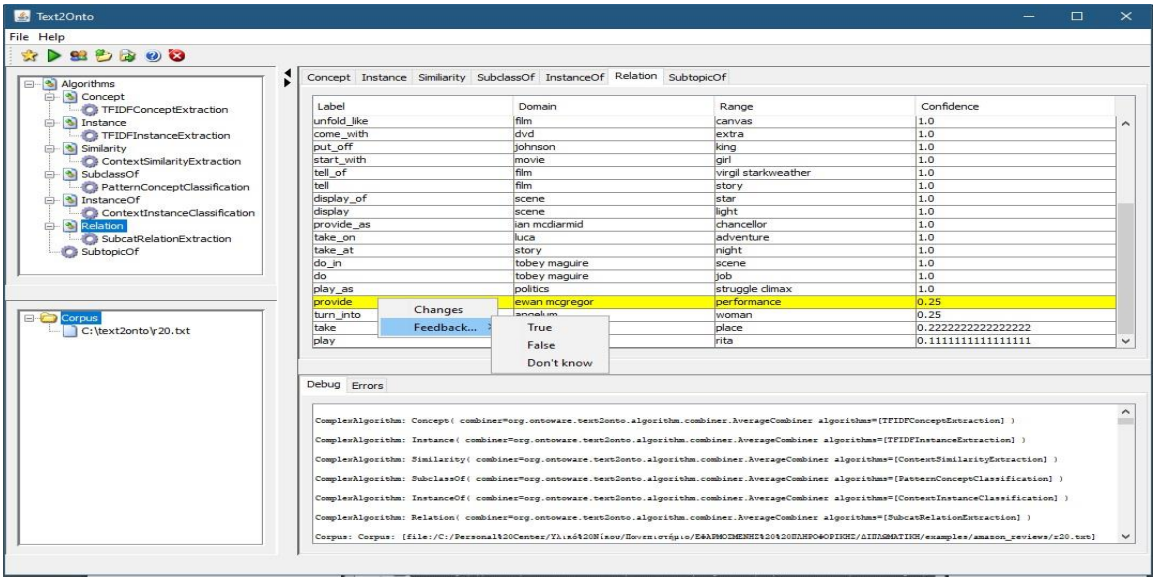
Ξεκινώντας τη διαδικασία ο χρήστης επιλέγει, όπως ειπώθηκε παραπάνω, τα έγγραφα κειμένου για τη δημιουργία του υποκείμενου σώματος δεδομένων και στη συνέχεια τους κατάλληλους συνδυασμούς αλγορίθμων για κάθε θεμελιακό στοιχείο από το τμήμα ελεγκτή αλγορίθμων. Για να ξεκινήσει η διαδικασία εκμάθησης οντολογίας, ο χρήστης μπορεί να επιλέξει *File* → *Run* από το κύριο μενού ή απλά να πατήσει το κατάλληλο κουμπί της γραμμής εργαλείων. Επιπλέον, το μενού *File* επιτρέπει επίσης την έναρξη μιας νέας συνεδρίας μάθησης οντολογίας μέσω της επιλογής *New*, καθώς και μέσω της επιλογής *Export* την εξαγωγή των αποτελεσμάτων του Πιθανοτικού Οντολογικού Μοντέλου σε συγκεκριμένο πρότυπο οντολογιών (*OWL, RDFS, KAON*). Οι υπόλοιπες επιλογές του μενού (*Save, Load, Import*) παρατηρήθηκε ότι δεν λειτουργούν στην παρούσα έκδοση της εφαρμογής και πιθανώς να υπάρχουν για χρήση σε μελλοντική της αναβάθμιση.

Μόλις μια οντολογία εξαχθεί από το υποκείμενο σώμα των δεδομένων, τα θεμελιακά στοιχεία μοντελοποίησης του Πιθανοτικού Οντολογικού Μοντέλου εμφανίζονται στον χρήστη όπως φαίνεται παρακάτω στην εικόνα 5-5. Σε αντίστοιχη

στήλη εμφανίζεται και η τιμή πιθανότητας του κάθε στοιχείου. Τα αποτελέσματα μπορούν επίσης να φιλτραριστούν, δίνοντας ανατροφοδότηση σε καθένα από αυτά για το αν είναι σκόπιμο να συμπεριληφθούν στην εξαγόμενη οντολογία. Αυτό γίνεται με δεξί κλικ σε κάθε στοιχείο ξεχωριστά και στη συνέχεια ορίζοντας την κατάλληλη ανατροφοδότηση όπως φαίνεται στην εικόνα 5-6 (*True*, *False*, *Don't know*).

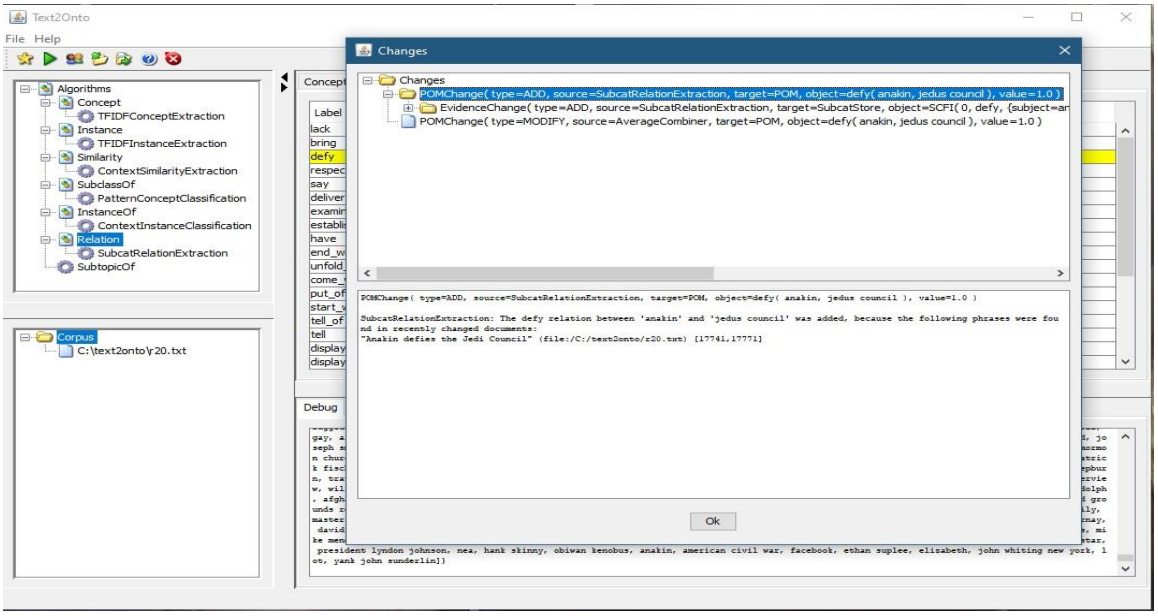


Εικόνα 5-5: Τα αποτελέσματα του Πιθανοτικού Οντολογικού Μοντέλου



Εικόνα 5-6: Η ανατροφοδότηση αποτελεσμάτων από το χρήστη

Ένας μέγιστος βαθμός ανιχνευσιμότητας δίνεται από το γεγονός ότι ο χρήστης όχι μόνο μπορεί να δει το ιστορικό αλλαγών οποιουδήποτε στοιχείου οντολογίας, αλλά και να πάρει μια εξήγηση φυσικής γλώσσας για όλες τις αποφάσεις μοντελοποίησης του συστήματος όπως φαίνεται παρακάτω στην εικόνα 5-7. Αυτό γίνεται μέσω της επιλογής *Changes* και πάλι με δεξί κλικ σε κάθε στοιχείο ξεχωριστά. Όπως φαίνεται και στην εικόνα ο χρήστης μπορεί να δει για κάθε στοιχείο βάσει ποιου αλγορίθμου ή συνδυασμού αλγορίθμων προέκυψε αυτό και από ποιο ακριβώς σημείο εγγράφου του υποκείμενου σώματος του συνόλου των εγγράφων αντλήθηκε η σχετική πληροφορία.



Εικόνα 5-7: Αλλαγές στα στοιχεία οντολογίας

5.9 Παραδείγματα εφαρμογών

Σε αυτήν την ενότητα γίνεται αναφορά σε ορισμένα παραδείγματα χρήσης της εφαρμογή *Text2Onto* και κάποιων βασικών συμπερασμάτων που έχουν προκύψει για την έκταση των δυνατοτήτων της στον τομέα της εκμάθησης οντολογιών.

Ένα παράδειγμα εφαρμογής του *Text2Onto* αφορά τον τομέα της διαχείρισης γνώσης και πιο συγκεκριμένα ένα σενάριο εφαρμογής που έχει να κάνει με μία ψηφιακή βιβλιοθήκη. Σε αυτό περιγράφεται η σχετική διαδικασία αποτίμησης της ποιότητας των αποτελεσμάτων που εξάγονται από το *Text2Onto* (Johanna Volker, Denny Vrandecic, York Sure (2005)).

Ένα δεύτερο παράδειγμα αναφέρεται στη χρήση της εφαρμογής *Text2Onto* σε ένα πλαίσιο ανάπτυξης οντολογίας στον τομέα της ασφάλειας κατασκευών της κατασκευαστικής βιομηχανίας (Han-Hsiang Wang, Frank Boukamp (2008)).

Το τελευταίο αναφέρεται σε ένα πείραμα χρήσης της εφαρμογής *Text2Onto*. Σε αυτό τα πειραματικά δεδομένα που χρησιμοποιούνται ως είσοδος αδόμητου κειμένου στην εφαρμογή προέρχονται από άρθρα που αφορούν τον τομέα της κατασκευής και εργαλείων εκμάθησης οντολογιών. Η οντολογία που παράγεται αυτόματα από την εφαρμογή συγκρίνεται με αντίστοιχη, παραγόμενη χειρωνακτικά χωρίς τη χρήση εργαλείου για την εξαγωγή των σχετικών συμπερασμάτων σχετικά με τα αποτελέσματα που μας δίνει το *Text2Onto* (Sonam Mittal, Nupur Mittal (2013)).

6 Πειράματα με την εφαρμογή εκμάθησης οντολογιών Text2Onto

6.1 Επιλογή του τομέα ενδιαφέροντος

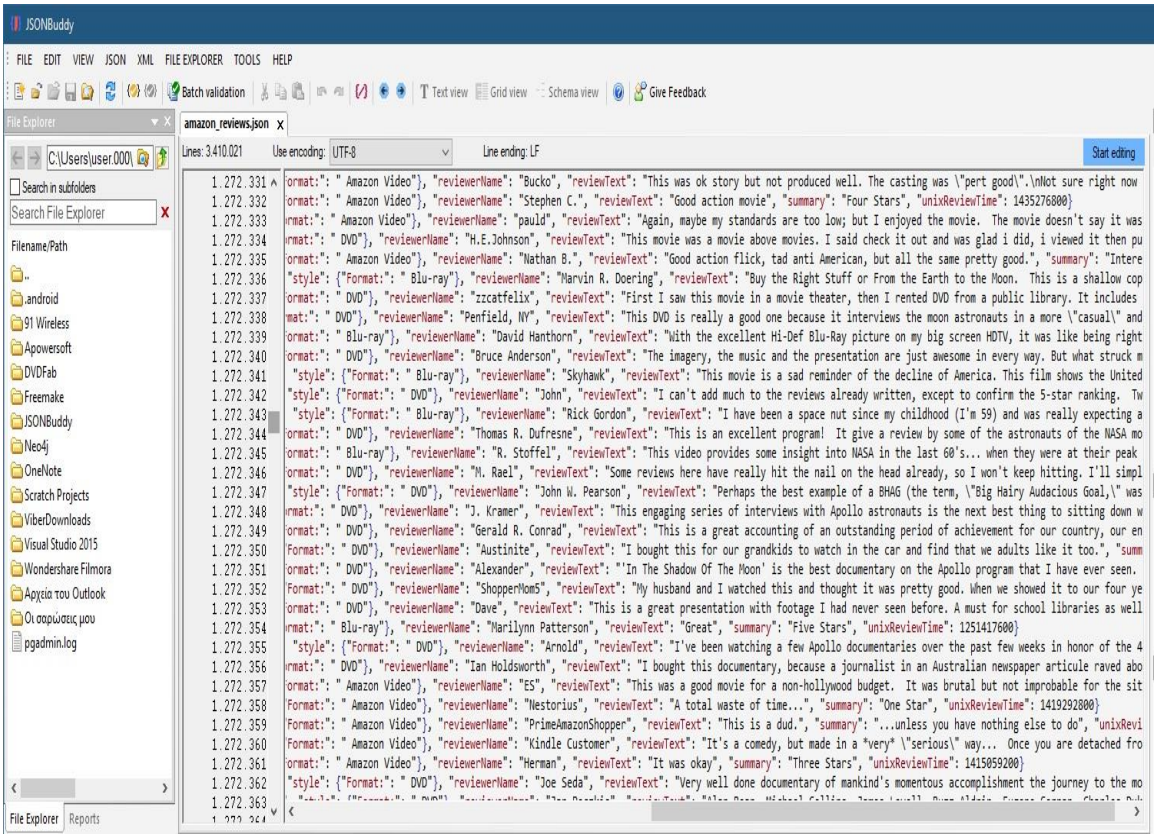
Όπως αναφέρθηκε και στο εισαγωγικό κεφάλαιο της εργασίας ο βασικός της σκοπός είναι η πειραματική μελέτη της εφαρμογής εκμάθησης οντολογιών *Text2Onto* για την αυτόματη εξαγωγή οντολογικής γνώσης σε γλώσσα *OWL* από αρχεία κειμένου, έτσι ώστε να εξαχθούν συμπεράσματα σχετικά με την έκταση των δυνατοτήτων και τους τρόπους εφαρμογής του εργαλείου στον τομέα των οντολογιών.

Ο αρχικός προβληματισμός είχε να κάνει με τον τομέα από τον οποίο θα αντληθούν τα δεδομένα σε μορφή κειμένου και εν συνεχεία το είδος τους. Θεωρήθηκε σκόπιμο να επιλεχθεί ένας τομέας διαδεδομένος σε ένα ευρύ φάσμα ανθρώπων, γι' αυτό και επιλέχθηκε ο τομέας των κινηματογραφικών ταινιών και τηλεοπτικών σειρών. Τα δεδομένα που επιλέχθηκαν έχουν να κάνουν με κριτικές και στοιχεία κινηματογραφικών ταινιών και αντλήθηκαν από δύο βασικές πηγές: Η πρώτη πηγή είναι η εταιρεία *Amazon*, η μεγαλύτερη στον κόσμο από την άποψη του κύκλου εργασιών από την πώληση αγαθών και υπηρεσιών μέσω του Διαδικτύου και μια από τις πρώτες που βασίστηκε στο Διαδίκτυο για την παροχή των υπηρεσιών της με γνώμονα τις πραγματικές πωλήσεις καταναλωτικών αγαθών. Η δεύτερη πηγή είναι η *Internet Movie Database* (Διαδικτυακή βάση δεδομένων ταινιών), πιο γνωστή με τη συντομογραφία *IMDb*, η οποία είναι διαδικτυακή βάση δεδομένων με πληροφορίες για ηθοποιούς, ταινίες, τηλεοπτικά προγράμματα, παρουσιαστές της τηλεόρασης, βιντεοπαιχνίδια και συντελεστές παραγωγής ταινιών ή προγραμμάτων.

Τα δεδομένα σε μορφή κειμένου που αντλήθηκαν από τις δύο αυτές πηγές χωρίστηκαν σε δύο βασικές υποκατηγορίες από την καθεμία πηγή, οπότε τελικά προέκυψαν για τα πειράματα που θα ακολουθούσαν, τέσσερις βασικές κατηγορίες με αρχεία δεδομένων. Οι τέσσερις αυτές κατηγορίες περιγράφονται αναλυτικά παρακάτω.

Αρχικά η πρώτη κατηγορία δεδομένων η οποία θα έχει την κωδική ονομασία *amazon_random* προέκυψε με την άντληση δεδομένων σχετικών με κριτικές κινηματογραφικών ταινιών και τηλεοπτικών σειρών από την *Amazon*. Η άντληση τους έγινε από τον σύνδεσμο <*https://nijianmo.github.io/amazon/index.html*>. Τα δεδομένα που αντλήθηκαν ήταν σε μορφή *JSON* αρχείου με κριτικές *5-core*, δηλαδή όπου καθένας από τους χρήστες και τα στοιχεία έχουν πέντε κριτικές το καθένα. Από το παραπάνω

αρχείο αυτό που ενδιέφερε κατά βάση για τα πειράματα ήταν το πεδίο *reviewText* το οποίο και περιείχε την κριτική του χρήστη για μια συγκεκριμένη ταινία ή τηλεοπτική σειρά. Λόγω του μεγάλου όγκου του αρχείου (2,32 *GB*) επιλέχθηκε η εφαρμογή *JSONBuddy* που είναι κατάλληλη για το άνοιγμα και την επεξεργασία μεγάλων αρχείων σε μορφή *JSON* <*https://www.json-buddy.com*>. Η μορφή του αρχείου, μέσω της εφαρμογής *JSONBuddy*, φαίνεται παρακάτω στην εικόνα 6-1.

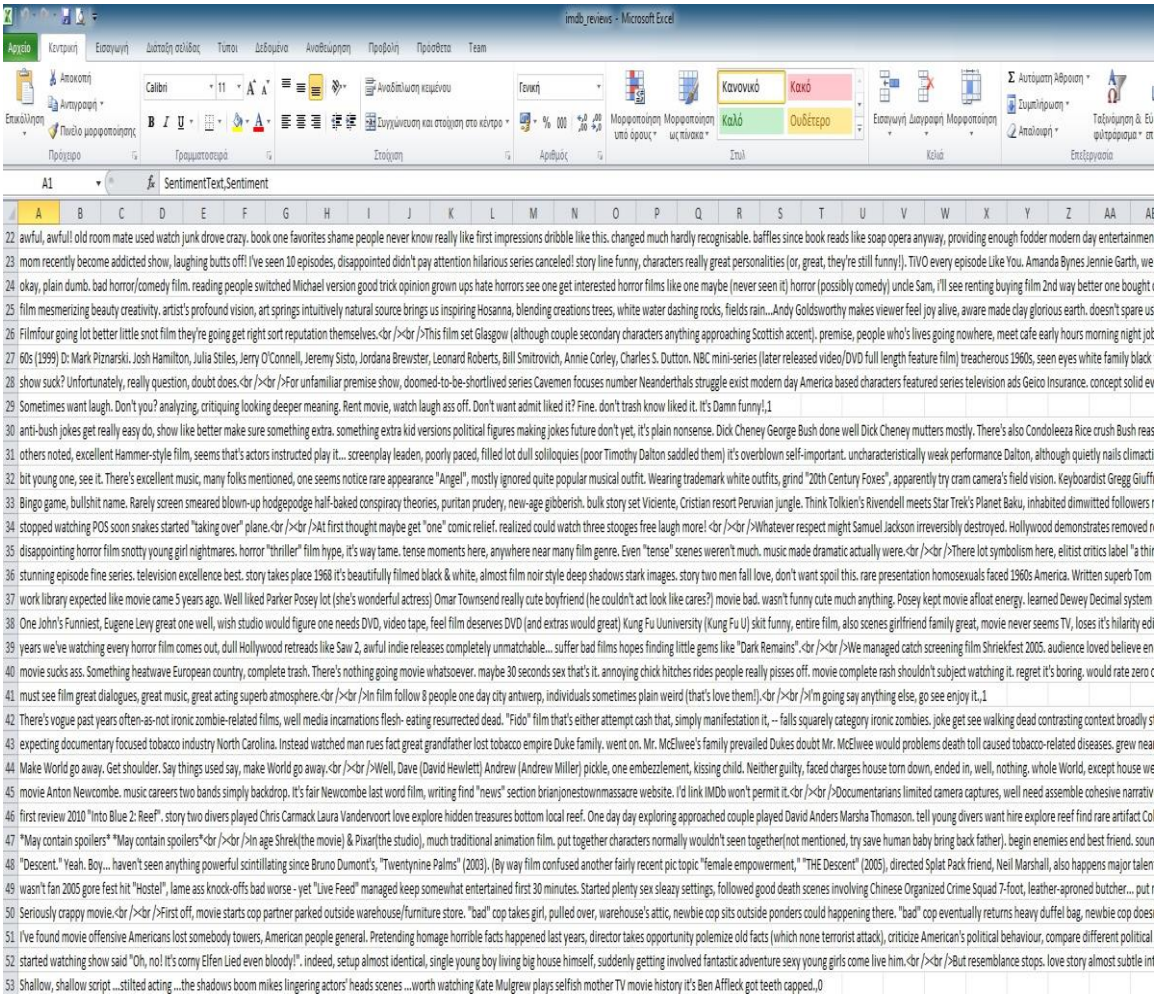


Εικόνα 6-1: Το αρχείο κριτικών κινηματογραφικών ταινιών και τηλεοπτικών σειρών από την Amazon

Η δεύτερη κατηγορία δεδομένων η οποία θα έχει την κωδική ονομασία *amazon_same* προέκυψε ως παραλλαγή της πρώτης. Ενώ δηλαδή στην πρώτη κατηγορία τα αρχεία δεδομένων με τις κριτικές που αντλήθηκαν έγιναν με τυχαίο τρόπο, στη δεύτερη κατηγορία αντλήθηκαν δεδομένα κριτικών μόνο για τέσσερις συγκεκριμένα τυχαίες ταινίες. Ο λόγος ήταν, όπως θα περιγραφεί και παρακάτω, να παρατηρηθεί αν θα υπήρχε έντονη διαφοροποίηση προς το καλύτερο της εξαγόμενης οντολογίας καθώς τα δεδομένα εισόδου θα είχαν μεγαλύτερη συνοχή και ομοιότητα αφού θα προέρχονταν από κριτικές που αφορούσαν τις ίδιες ταινίες.

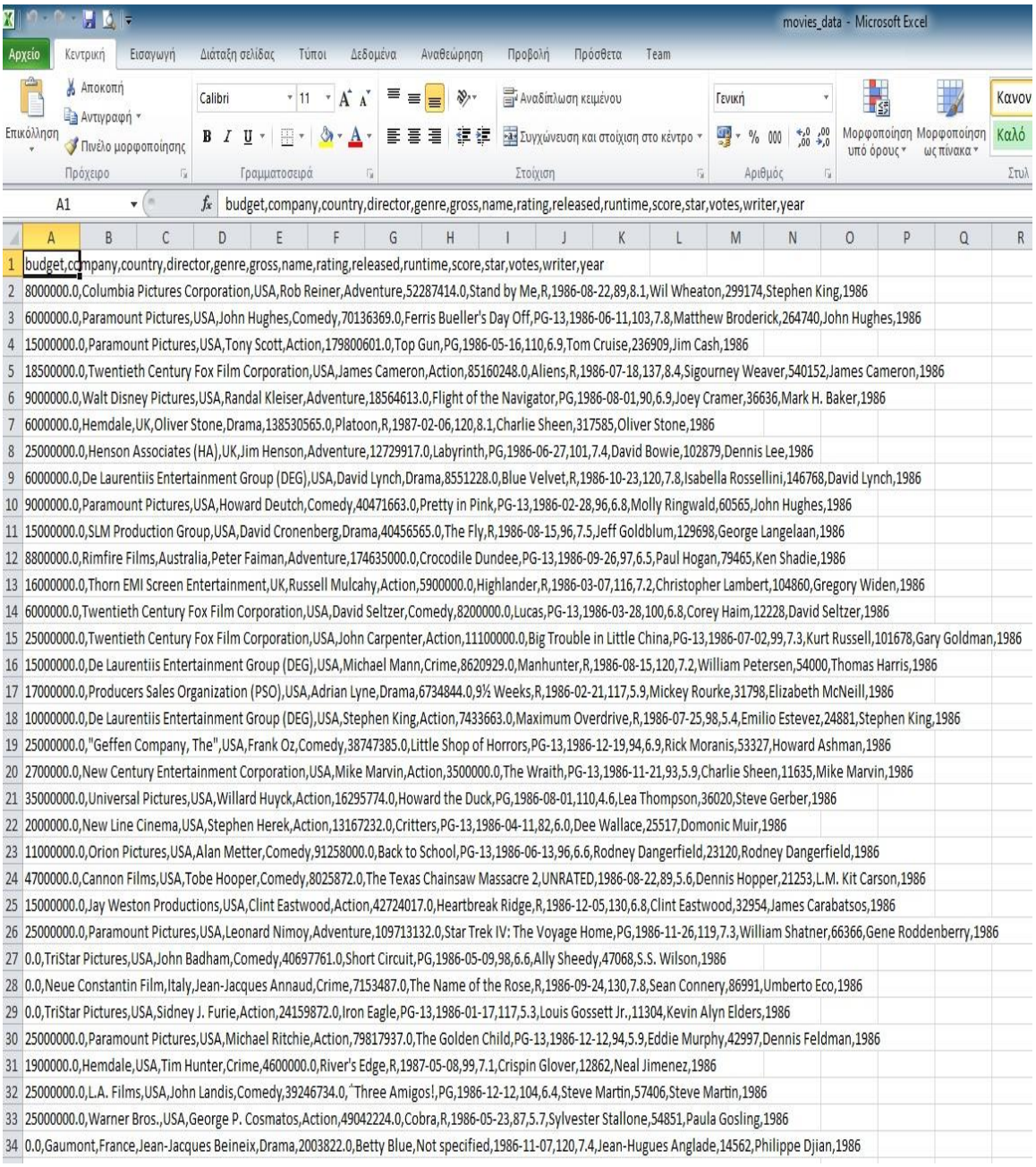
Για την τρίτη κατηγορία δεδομένων η οποία θα έχει την κωδική ονομασία *imdb_reviews* ο στόχος ήταν να επιλεγεί μια διαφορετική πηγή άντλησης κριτικών

κινηματογραφικών ταινιών. Έτσι επιλέχθηκε τα δεδομένα να αντληθούν από την *IMDb*. Η άντληση τους έγινε από το σύνδεσμο <*https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews*>. Τα δεδομένα που αντλήθηκαν ήταν σε μορφή *csv* αρχείου. Η μορφή του αρχείου φαίνεται παρακάτω στην εικόνα 6-2.



Εικόνα 6-2: Το αρχείο κριτικών κινηματογραφικών ταινιών από την IMDb

Τέλος ως τέταρτη κατηγορία δεδομένων η οποία θα έχει την κωδική ονομασία *imdb_movies_data* ο στόχος ήταν να επιλεγεί ένα διαφορετικό από τις κριτικές είδος δεδομένων. Έτσι επιλέχθηκαν, με πηγή άντλησης των δεδομένων πάλι την *IMDb*, δεδομένα σχετικά με στοιχεία κινηματογραφικών ταινιών (όνομα ταινίας, ηθοποιοί, σκηνοθέτης, σεναριογράφος, έτος κυκλοφορίας κτλ). Τα δεδομένα που αντλήθηκαν ήταν σε μορφή *csv* αρχείου. Η άντληση τους έγινε από τον σύνδεσμο <*https://www.kaggle.com/danielgrijalvas/movies*>. Η μορφή του αρχείου φαίνεται παρακάτω στην εικόνα 6-3.



Εικόνα 6-3: Το αρχείο στοιχείων κινηματογραφικών ταινιών από την IMDb

6.2 Η συλλογιστική των πειραμάτων

Αφού επιλέχθηκαν οι κατηγορίες των δεδομένων που περιγράφηκαν στην προηγούμενη ενότητα το επόμενο βήμα αφορά τη συλλογιστική πραγματοποίησης των πειραμάτων με την εφαρμογή. Οι κατευθύνσεις διαφοροποίησης των πειραμάτων είναι δύο: η πρώτη αφορά τη διαφοροποίηση ως προς τον αριθμό των δεδομένων εισόδου. Η δεύτερη, τη διαφοροποίηση ως προς την επιλογή και το συνδυασμό των διαθέσιμων αλγορίθμων για κάθε θεμελιακό στοιχείο.

Ως προς τα δεδομένα εισόδου, σε ό, τι έχει να κάνει με τις κατηγορίες δεδομένων *amazon_rando*m και *imdb_reviews*, επιλέχθηκε να δημιουργηθούν αρχεία με 20, 100 και 1000 κριτικές αντίστοιχα ώστε να παρατηρηθούν οι διαφοροποιήσεις στα αποτελέσματα της οντολογίας σε σχέση με τη μεταβολή του αριθμού των κριτικών. Τα παραπάνω αρχεία είναι αρχεία κειμένου (*.txt*) και εξήχθησαν από τα αρχεία κριτικών σε μορφή *JSON* και *CSV* που αναφέρθηκαν στην προηγούμενη ενότητα. Η εξαγωγή των κριτικών έγινε με τυχαίο τρόπο. Στα πειράματα χρησιμοποιήθηκε ως είσοδος στο υποκείμενο σώμα των εγγράφων αρχικά το αρχείο των 20. Στο επόμενο βήμα πειράματος προστέθηκε στο υποκείμενο σώμα και το αρχείο των 100 ενώ στο τελευταίο βήμα προστέθηκε και το αρχείο των 1000 κριτικών.

Όσον αφορά την κατηγορία δεδομένων *amazon_same,* όπως και στην κατηγορία *amazon_random* επιλέχθηκε να εξαχθεί ένα αρχείο με 1000 κριτικές, όχι όμως με τυχαίο τρόπο αλλά με κριτικές που αφορούν μόνο τέσσερις συγκεκριμένες ταινίες. Επιλέχθηκε μόνο ένα αρχείο με 1000 κριτικές γιατί ο βασικός στόχος των πειραμάτων αυτής της κατηγορίας ήταν η σύγκριση με τα πειράματα της πρώτης κατηγορίας και όχι όσον αφορά τη μεταβολή στον αριθμό των κριτικών. Καθώς τα δεδομένα εισόδου προέρχονται από κριτικές που αφορούν τις ίδιες ταινίες, άρα υπάρχει και μια σχετική επανάληψη ως προς το περιεχόμενο των κριτικών, κρίθηκε σκόπιμο να παρατηρηθούν οι διαφοροποιήσεις των αποτελεσμάτων που αφορούν τα θεμελιακά στοιχεία της εξαγόμενης οντολογίας σε σχέση και με τον αριθμό τους αλλά και τις πιθανότητες τους, αντίστοιχα με αυτά της κατηγορίας *amazon_random* όπου οι κριτικές επιλέχθηκαν με τυχαίο τρόπο.

Τέλος, ως προς την κατηγορία δεδομένων *imdb_movies_data* όπου τα δεδομένα αποτελούν στοιχεία και όχι κριτικές κινηματογραφικών ταινιών, επιλέχθηκε να εξαχθεί από το αρχικό αρχείο, ένα αρχείο με στοιχεία από 1000 ταινίες. Το αρχείο αυτό χρησιμοποιείται ως είσοδος για τα πειράματα αυτής της κατηγορίας, με σκοπό τη σύγκριση των αποτελεσμάτων της εξαγόμενης οντολογίας με αυτά των προηγούμενων κατηγοριών.

Η δεύτερη κατεύθυνση διαφοροποίησης των πειραμάτων έχει να κάνει, όπως ειπώθηκε πιο πάνω, με την στρατηγική επιλογής και συνδυασμού των διαθέσιμων αλγορίθμων για κάθε θεμελιακό στοιχείο. Ως προς αυτήν την κατεύθυνση επιλέχθηκε να γίνουν πέντε διαφορετικά είδη πειραμάτων. Για τα πρώτα τέσσερα είδη πειραμάτων χρησιμοποιήθηκε από ένας αλγόριθμος για κάθε θεμελιακό στοιχείο. Καθώς οι

αλγόριθμοι που είναι διαθέσιμοι στην εφαρμογή για το καθένα θεμελιακό στοιχείο είναι από ένας έως τέσσερις, κατά τη σειρά εκτέλεσης αυτών των τεσσάρων πειραμάτων, όπου υπήρχε η δυνατότητα, χρησιμοποιούνταν στο επόμενο πείραμα, ανά θεμελιακό στοιχείο, ο επόμενος διαθέσιμος αλγόριθμος, ενώ αν δεν υπήρχε άλλος χρησιμοποιούνταν ο τελευταίος διαθέσιμος από το προηγούμενο πείραμα. Στο πέμπτο είδος πειραμάτων έγινε ένας συνδυασμός (με *Average Combiner*) όλων των διαθέσιμων αλγορίθμων ανά θεμελιακό στοιχείο.

Σε αυτό το σημείο θα πρέπει να σημειωθεί ότι το θεμελιακό στοιχείο μοντελοποίησης *SubtopicOf* δεν συμπεριλήφθηκε στα πειράματα λόγω πανομοιότυπων αποτελεσμάτων με το θεμελιακό στοιχείο *SubclassOf*, σε συνάρτηση με το γεγονός ότι η *OWL*, η οποία είναι και η γλώσσα που εξάγονται τα αποτελέσματα των πειραμάτων ως *owl* αρχεία, δεν υποστηρίζει το τελευταίο στοιχείο. Οι αλγόριθμοι που χρησιμοποιούνται για καθένα από τα πέντε είδη πειραμάτων ανά θεμελιακό στοιχείο συνοψίζονται παρακάτω, στον πίνακα 6-1.

Πείραμα 01	
Concept	*EntropyConceptExtraction*
Instance	*ExampleInstanceExtraction*
Similarity	*ContextSimilarityExtraction*
SubclassOf	*PatternConceptClassification*
InstanceOf	*ContextInstanceClassification*
Relation	*SubcatRelationExtraction*
Πείραμα 02	
Concept	*ExampleConceptExtraction*
Instance	*ExampleInstanceExtraction*
Similarity	*ContextSimilarityExtraction*
SubclassOf	*VerticalRelationsConceptClassification*
InstanceOf	*ContextInstanceClassification*
Relation	*SubcatRelationExtraction*
Πείραμα 03	
Concept	*RTFConceptExtraction*
Instance	*TFIDFInstanceExtraction*
Similarity	*ContextSimilarityExtraction*
SubclassOf	*WordNetConceptClassification*

InstanceOf	*PatternInstanceClassification*
Relation	*SubcatRelationExtraction*
Πείραμα 04	
Concept	*TFIDFConceptExtraction*
Instance	*TFIDFInstanceExtraction*
Similarity	*ContextSimilarityExtraction*
SubclassOf	*WordNetConceptClassification*
InstanceOf	*PatternInstanceClassification*
Relation	*SubcatRelationExtraction*
Πείραμα 05	
Concept	*EntropyConceptExtraction*
	ExampleConceptExtraction
	RTFConceptExtraction
	TFIDFConceptExtraction
Instance	*ExampleInstanceExtraction*
	TFIDFInstanceExtraction
Similarity	*ContextSimilarityExtraction*
SubclassOf	*PatternConceptClassification*
	VerticalRelationsConceptClassification
	WordNetConceptClassification
InstanceOf	*ContextInstanceClassification*
	PatternInstanceClassification
Relation	*SubcatRelationExtraction*

Πίνακας 6-1: Οι αλγόριθμοι των πέντε ειδών πειραμάτων ανά θεμελιακό στοιχείο

Με βάση τις δύο κατευθύνσεις διαφοροποίησης των πειραμάτων που περιγράφηκαν παραπάνω, ως προς τον αριθμό των δεδομένων εισόδου και ως προς την επιλογή και συνδυασμό των αλγορίθμων για κάθε θεμελιακό στοιχείο, στη συνέχεια πραγματοποιήθηκαν με την εφαρμογή πειράματα που να συνδυάζουν όλες τις περιπτώσεις και των δύο κατευθύνσεων. Τα πειράματα αυτά πραγματοποιήθηκαν και για τις τέσσερις κατηγορίες δεδομένων που αναλύθηκαν στην προηγούμενη ενότητα.

Όσον αφορά την ονομασία των πειραμάτων θεωρήθηκε σκόπιμο να έχει μία κωδικοποίηση που να προέρχεται από τις δύο κατευθύνσεις διαφοροποίησης. Η κωδικοποίηση αυτή έχει τη μορφή ***rX_expY_Z***, όπου το *X* να παίρνει τιμές 20, 100, 1000

ανάλογα με την ποσότητα των δεδομένων (κριτικών, στοιχείων ταινιών), όπου το *Y* να παίρνει τιμές 01, 02, 03, 04, 05 ανάλογα με το είδος πειραμάτων που φαίνονται στον πίνακα 6-1 και όπου *Z* να παίρνει τιμές *amazon_random, amazon_same, imdb_reviews, imdb_movies_data* ανάλογα με την κατηγορία των δεδομένων.

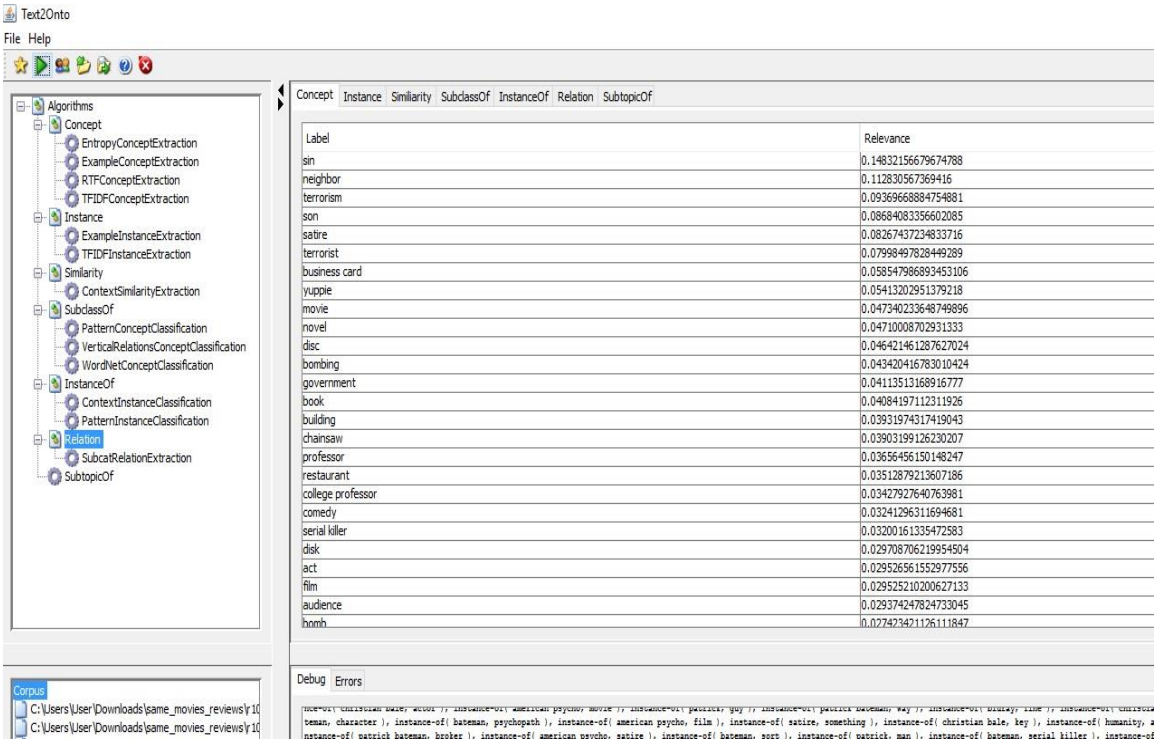
6.3 Εκτέλεση των πειραμάτων

Σύμφωνα με τη συλλογιστική που αναπτύχθηκε στην προηγούμενη ενότητα πραγματοποιήθηκαν τα πέντε είδη πειραμάτων για τις τέσσερις κατηγορίες δεδομένων που αναφέρθηκαν στην ενότητα 6.1. Συνολικά πραγματοποιήθηκαν σαράντα πειράματα. Τα αποτελέσματα των θεμελιακών στοιχείων μοντελοποίησης του Πιθανοτικού Οντολογικού Μοντέλου για το καθένα από αυτά εξήχθησαν σε μορφή *owl* αρχείων.

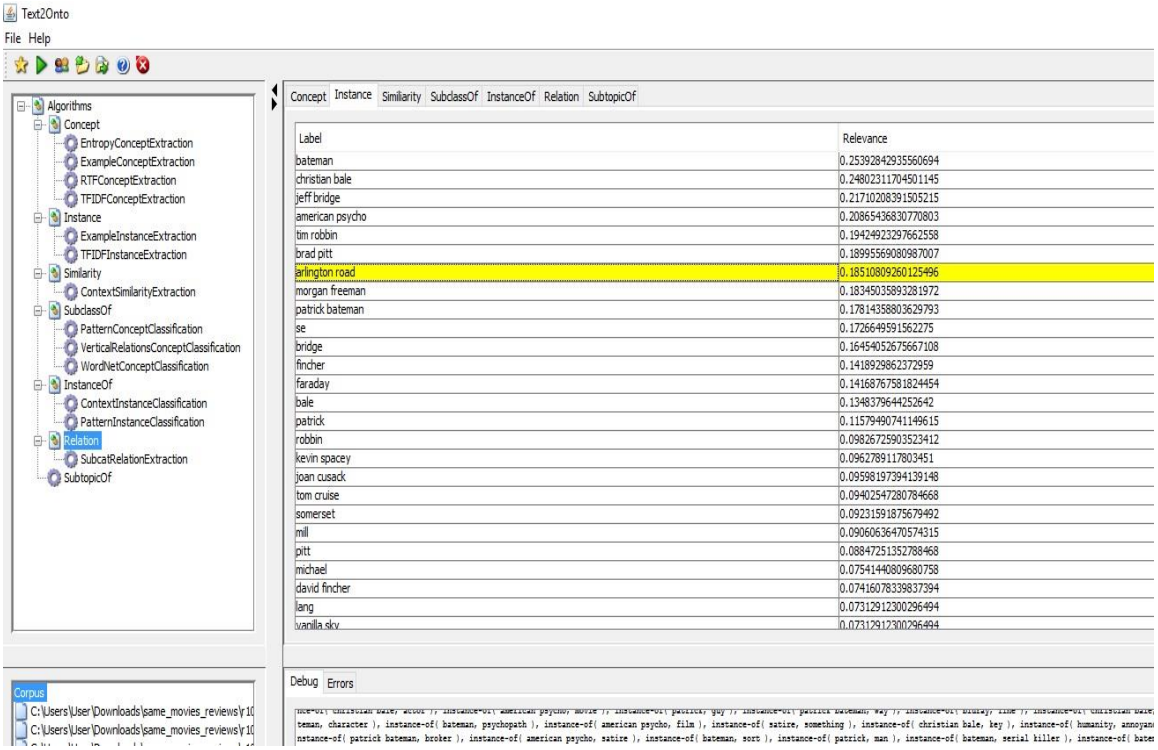
Πιο αναλυτικά για κατηγορίες δεδομένων *amazon_rando*m και *imdb_reviews* που υπήρχε και διαφοροποίηση των δεδομένων ως προς την ποσότητα τους, εκτελέστηκαν από δεκαπέντε πειράματα σε κάθε κατηγορία, πέντε για κάθε ποσότητα δεδομένων της τάξης των 20, 100 και 1000. Για τις κατηγορίες δεδομένων *amazon_same* και *imdb_movies_data* εκτελέστηκαν από πέντε πειράματα για την ποσότητα δεδομένων της τάξης των 1000.

Ενδεικτικά αναλύεται παρακάτω η εκτέλεση του πειράματος *r1000_exp05_amazon_same* που αφορά την κατηγορία δεδομένων με κριτικές ταινιών για τέσσερις συγκεκριμένες ταινίες με πηγή την *Amazon*. Αρχικά από το τμήμα της υποκείμενης συλλογής εγγράφων έχουμε εισάγει το σώμα των δεδομένων εισόδου σε μορφή εγγράφων κειμένου *txt*. Συγκεκριμένα τα δεδομένα των 1000 κριτικών έχουν σπάσει σε δέκα *tx*t αρχεία των 100 κριτικών το καθένα. Στη συνέχεια από το τμήμα του ελεγκτή αλγορίθμων επιλέγουμε για κάθε θεμελιακό στοιχείο όλους τους διαθέσιμους αλγόριθμους. Επιλέγουμε να τρέξουμε το πείραμα με ένα συνδυασμό αλγορίθμων, χρησιμοποιώντας ως συνδυαστή αλγορίθμων τον *Average Combiner*. Για να τρέξουμε το πείραμα επιλέγουμε το κουμπί *Run* από τη γραμμή εργαλείων. Αφού ολοκληρωθεί η διαδικασία της εξαγωγής αποτελεσμάτων οντολογίας από την εφαρμογή, στο τμήμα του Πιθανοτικού Οντολογικού Μοντέλου (*POM*), φαίνονται τα αποτελέσματα της εξαγόμενης οντολογίας του πειράματος για κάθε θεμελιακό στοιχείο στις αντίστοιχες καρτέλες. Τα στοιχεία οντολογίας εξάγονται ως αποτελέσματα για το κάθε θεμελιακό στοιχείο και εμφανίζονται στη μορφή δύο στηλών, μια με το όνομα του στοιχείου και μια με την τιμή της πιθανότητάς του. Τέλος μέσω της επιλογής *File → Export* γίνεται εξαγωγή των αποτελεσμάτων του Πιθανοτικού Οντολογικού Μοντέλου σε μορφή *owl*

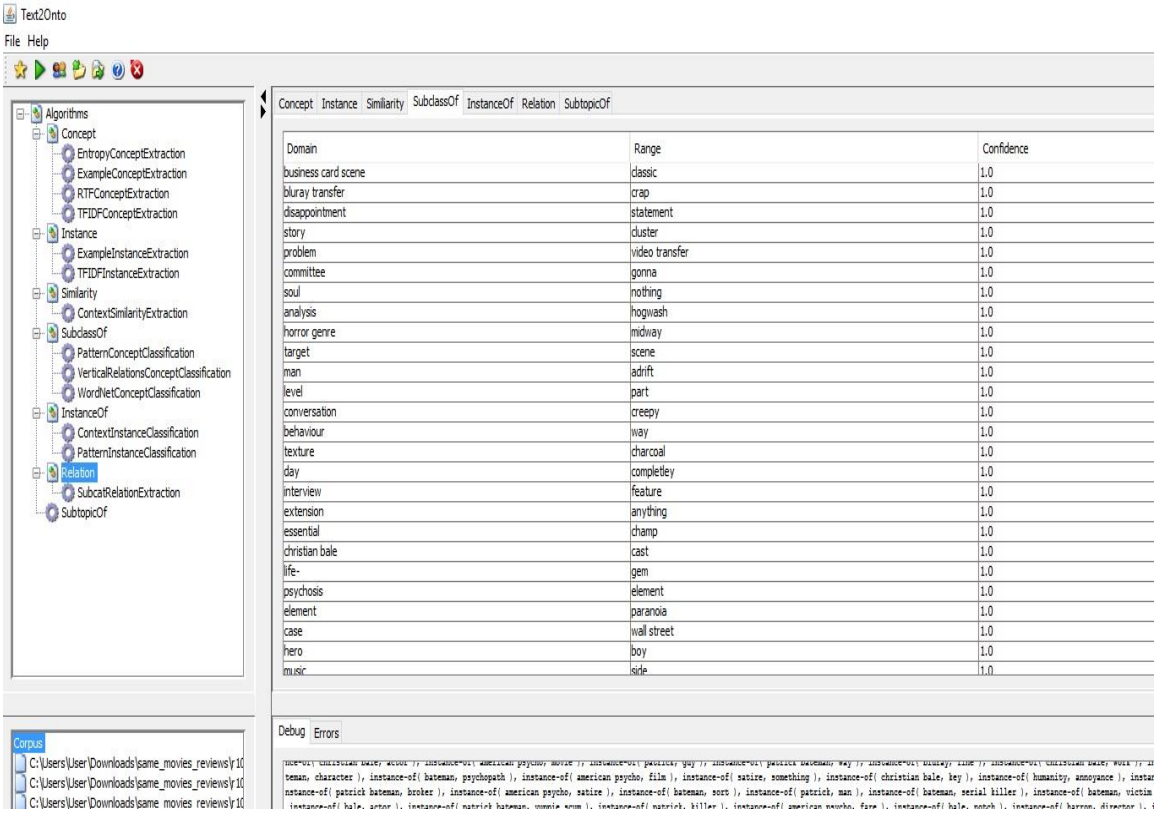
αρχείου. Τα αποτελέσματα της εκτέλεσης του πειράματος φαίνονται παρακάτω στις εικόνες 6-4 έως 6-8 για κάθε θεμελιακό στοιχείο αντίστοιχα.



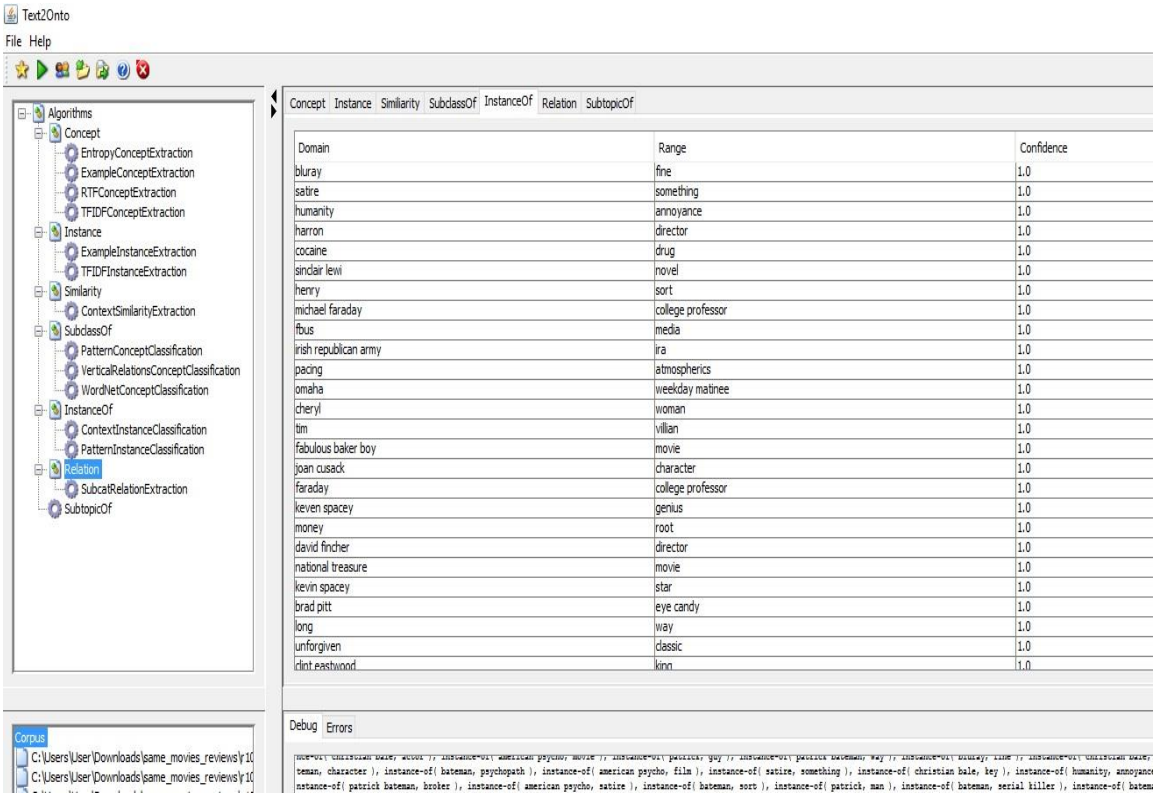
Εικόνα 6-4: Αποτελέσματα του θεμελιακού στοιχείου Concept για το πείραμα r1000_exp05_amazon_same



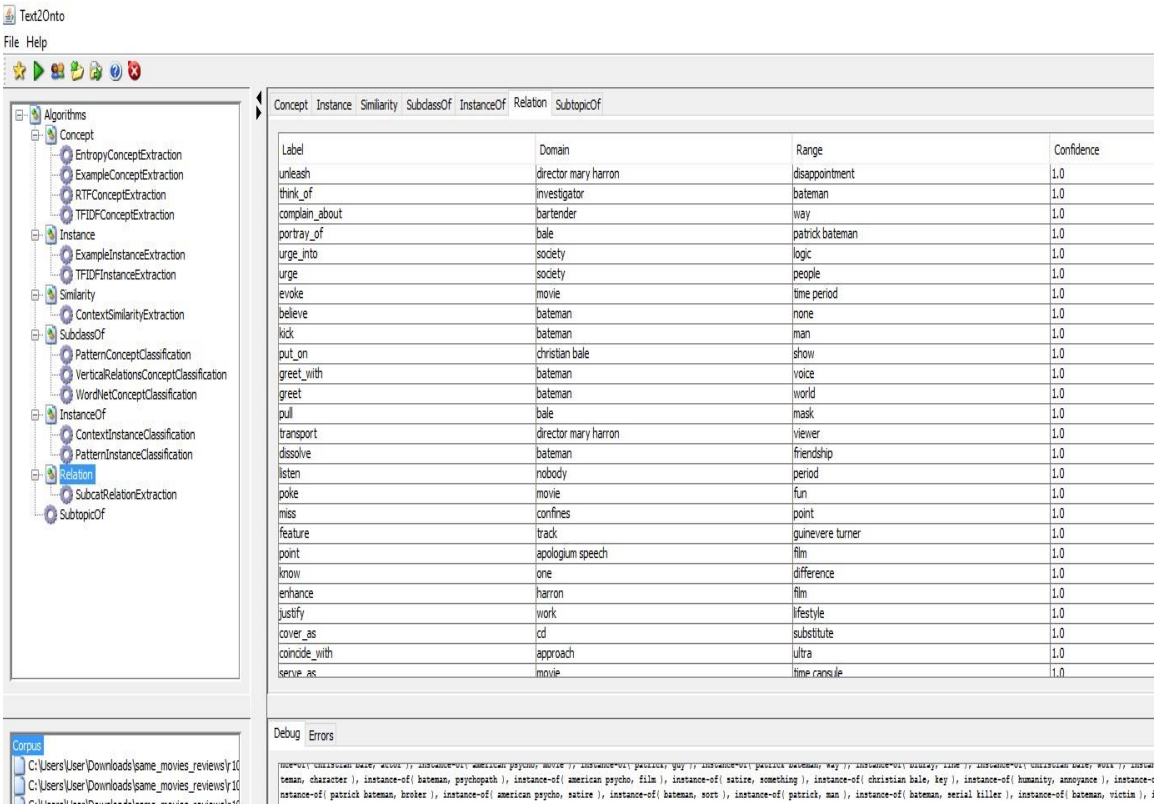
Εικόνα 6-5: Αποτελέσματα του θεμελιακού στοιχείου Instance για το πείραμα r1000_exp05_amazon_same



Εικόνα 6-6: Αποτελέσματα του θεμελιακού στοιχείου SubclassOf για το πείραμα r1000_exp05_amazon_same



Εικόνα 6-7: Αποτελέσματα του θεμελιακού στοιχείου InstanceOf για το πείραμα r1000_exp05_amazon_same



Εικόνα 6-8: Αποτελέσματα του θεμελιακού στοιχείου Relation για το πείραμα r1000_exp05_amazon_same

Όσον αφορά την εκτέλεση και τα αποτελέσματα των πειραμάτων ορισμένες γενικές παρατηρήσεις είναι οι παρακάτω:

- Για την εκτέλεση των πειραμάτων χρησιμοποιήθηκε laptop με επεξεργαστή *Intel Core i7-2630QM CPU @2.00GHz*, μνήμη *RAM 8.00GB* και λειτουργικό σύστημα *Windows 10 Home 64bit*.
- Ο χρόνος εκτέλεσης των πειραμάτων αυξανόταν εκθετικά με την αύξηση της ποσότητας των δεδομένων. Στην περίπτωση που τα δεδομένα ήταν στοιχεία ταινιών ο χρόνος εκτέλεσης ήταν σημαντικά μικρότερος σε σχέση με τις κριτικές ταινιών για όλες τις τάξεις των δεδομένων.
- Όσον αφορά τις κριτικές στην περίπτωση που ήταν της τάξης των 20 κριτικών, ο χρόνος εκτέλεσης των πειραμάτων ήταν λίγα λεπτά της ώρας. Στην περίπτωση των 100 κριτικών ήταν γύρω στα 30 λεπτά της ώρας. Στην περίπτωση των 1000 κριτικών παρατηρήθηκε ότι όταν αυτές ήταν όλες σε ένα αρχείο εισόδου η εκτέλεση του πειράματος ξεπέρασε τις 15 ώρες, ενώ όταν κατακερματίστηκαν σε 10 αρχεία εισόδου των 100 κριτικών, ο χρόνος εκτέλεσης ήταν ανάμεσα σε 2 και 3 ώρες.

- Σε όλα τα πειράματα που πραγματοποιήθηκαν, το Πιθανοτικό Οντολογικό Μοντέλο της εφαρμογής δεν έδωσε αποτελέσματα για το θεμελιακό στοιχείο *Similarity* για το οποίο μοναδικός διαθέσιμος αλγόριθμος ήταν ο *ContextSimilarityExtraction*.
- Όσον αφορά όλες τις εφαρμογές των πειραμάτων 01,02, όπου για το θεμελιακό στοιχείο *InstanceOf* χρησιμοποιήθηκε ο αλγόριθμος *ContextInstanceClassification*, το Πιθανοτικό Οντολογικό Μοντέλο της εφαρμογής δεν έδωσε αποτελέσματα.
- Στην κατηγορία δεδομένων *imdb_movies_data*, στο πείραμα 01, όπου για το θεμελιακό στοιχείο *SubclassOf* χρησιμοποιήθηκε ο αλγόριθμος *PatternConceptClassification*, το Πιθανοτικό Οντολογικό Μοντέλο της εφαρμογής δεν έδωσε αποτελέσματα.
- Σε ορισμένα πειράματα, σε αποτελέσματα που αφορούσαν τα θεμελιακά στοιχεία *Concept* και *Instance*, το πεδίο τιμής της πιθανότητας των στοιχείων τους δεν είχε αριθμητικές τιμές, αλλά την τιμή *NaN*.
- Σε όλες τις υπόλοιπες περιπτώσεις των πειραμάτων, το Πιθανοτικό Οντολογικό Μοντέλο της εφαρμογής έδωσε αποτελέσματα για όλα τα θεμελιακά στοιχεία μοντελοποίησης.
- Σε ό, τι έχει να κάνει με τα θεμελιακά στοιχεία *Concept* και *Instance*, παρατηρήθηκε σε όλες τις περιπτώσεις των πειραμάτων ότι οι τιμές πιθανότητας των στοιχείων οντολογίας τους είναι αρκετά χαμηλές σε σχέση και με τα υπόλοιπα θεμελιακά στοιχεία.

7 Ανάπτυξη διαδικασιών κώδικα για την ποσοτική και ποιοτική ανάλυση των αποτελεσμάτων των πειραμάτων

7.1 Συνοπτική περιγραφή

Όπως αναφέρθηκε και στην προηγούμενη ενότητα τα αποτελέσματα των θεμελιακών στοιχείων μοντελοποίησης του Πιθανοτικού Οντολογικού Μοντέλου, στα πειράματα που πραγματοποιήθηκαν, εξήχθησαν σε γλώσσα οντολογίας *OWL* με τη μορφή *owl* αρχείων. Λόγω του μεγάλου όγκου των αποτελεσμάτων των πειραμάτων και των αντίστοιχων *owl* αρχείων κρίθηκε αναγκαίο να αναπτυχθούν ορισμένες διαδικασίες κώδικα ώστε να διευκολυνθεί η ποσοτική και ποιοτική ανάλυση των αποτελεσμάτων των πειραμάτων. Για την ανάπτυξη των διαδικασιών χρησιμοποιήθηκε η γλώσσα προγραμματισμού *Javascript* ενώ η αντίστοιχη έξοδος των αποτελεσμάτων δόθηκε σε απλή *HTML* μορφή και μέσω της κονσόλας ενός προγράμματος περιήγησης.

Οι βασικές κατευθύνσεις της ανάλυσης των αποτελεσμάτων των πειραμάτων, για τις οποίες αναπτύχθηκαν και οι αντίστοιχες διαδικασίες, συνοπτικά ήταν οι εξής:

- Μέτρηση του αριθμού των στοιχείων της οντολογίας για κάθε θεμελιακό στοιχείο ενός *owl* αρχείου που πήραμε ως έξοδο από την εφαρμογή *Text2Onto*.
- Μαζική διαγραφή από ένα *owl* αρχείο των στοιχείων της οντολογίας, όπου η τιμή του πεδίου πιθανότητάς τους είναι κάτω από ένα κατώφλι πιθανότητας το οποίο ορίζεται για κάθε θεμελιακό στοιχείο αντίστοιχα.
- Σύγκριση των *owl* αρχείων που έχουν εξαχθεί από δύο διαφορετικά πειράματα, ως προς τον αριθμό των κοινών στοιχείων της οντολογίας για κάθε θεμελιακό στοιχείο αντίστοιχα.
- Ποσοτική ομαδοποίηση των στοιχείων της οντολογίας ενός *owl* αρχείου, ανάλογα με την τιμή του πεδίου πιθανότητάς τους σύμφωνα με κάποια προκαθορισμένα επίπεδα πιθανότητας για κάθε θεμελιακό στοιχείο αντίστοιχα.
- Χρησιμοποίηση των εξαγόμενων *owl* αρχείων ως είσοδο στο περιβάλλον οντολογιών *Protégé* για δυνατότητα περαιτέρω επεξεργασίας της εξαγόμενης οντολογίας.
- Σύγκριση ενός *owl* αρχείου που πήραμε ως έξοδο από το *Text2Onto* με ένα πρότυπο *owl* αρχείο μιας οντολογίας σχετικά με κινηματογραφικές ταινίες ως προς τον αριθμό των κοινών στοιχείων των οντολογιών για κάθε θεμελιακό στοιχείο αντίστοιχα.

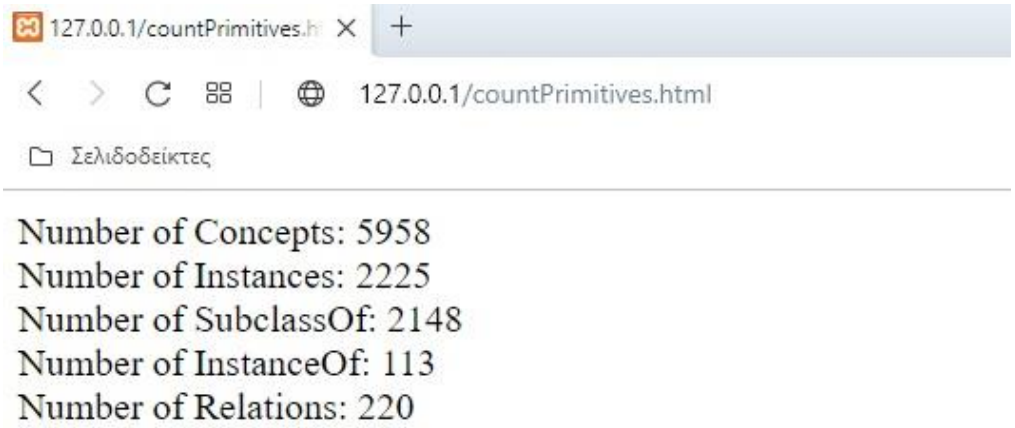
Για τις παραπάνω κατευθύνσεις αναπτύχθηκαν συνολικά έξι διαδικασίες οι οποίες περιγράφονται συνοπτικά στις επόμενες ενότητες. Οι διαδικασίες αυτές είναι οι εξής:

- **Μέτρησης θεμελιακών στοιχείων**
- **Διαγραφής θεμελιακών στοιχείων**
- **Σύγκρισης θεμελιακών στοιχείων**
- **Πιθανοτικής ομαδοποίησης θεμελιακών στοιχείων**
- **Μετατροπής θεμελιακών στοιχείων**
- **Σύγκρισης με πρότυπο οντολογίας**

7.2 Η διαδικασία μέτρησης θεμελιακών στοιχείων

Μια πρώτη κατεύθυνση όσον αφορά την ανάλυση των αποτελεσμάτων των πειραμάτων, όπως ειπώθηκε πιο πάνω, ήταν η μέτρηση του αριθμού των στοιχείων της οντολογίας για κάθε θεμελιακό στοιχείο των *owl* αρχείων που εξάγονται από το *Text2Onto* ως αποτελέσματα των πειραμάτων.

Για το σκοπό αυτό αναπτύχθηκε η διαδικασία μέτρησης θεμελιακών στοιχείων. Ο κώδικας περιέχει δύο συναρτήσεις: τη συνάρτηση *loadOwlDoc* η οποία φορτώνει το *owl* αρχείο και καλεί μια νέα συνάρτηση για την επεξεργασία του. Η συνάρτηση αυτή είναι η *processOwl* η οποία επεξεργάζεται το *DOM* (***D**ocument **O**bject **M**odel*) του *owl* αρχείου για να μετρήσει τον αριθμό των στοιχείων οντολογίας για κάθε θεμελιακό στοιχείο αντίστοιχα. Στην εικόνα 7-1 που ακολουθεί φαίνονται ενδεικτικά τα αποτελέσματα όπως εμφανίζονται στο πρόγραμμα περιήγησης.



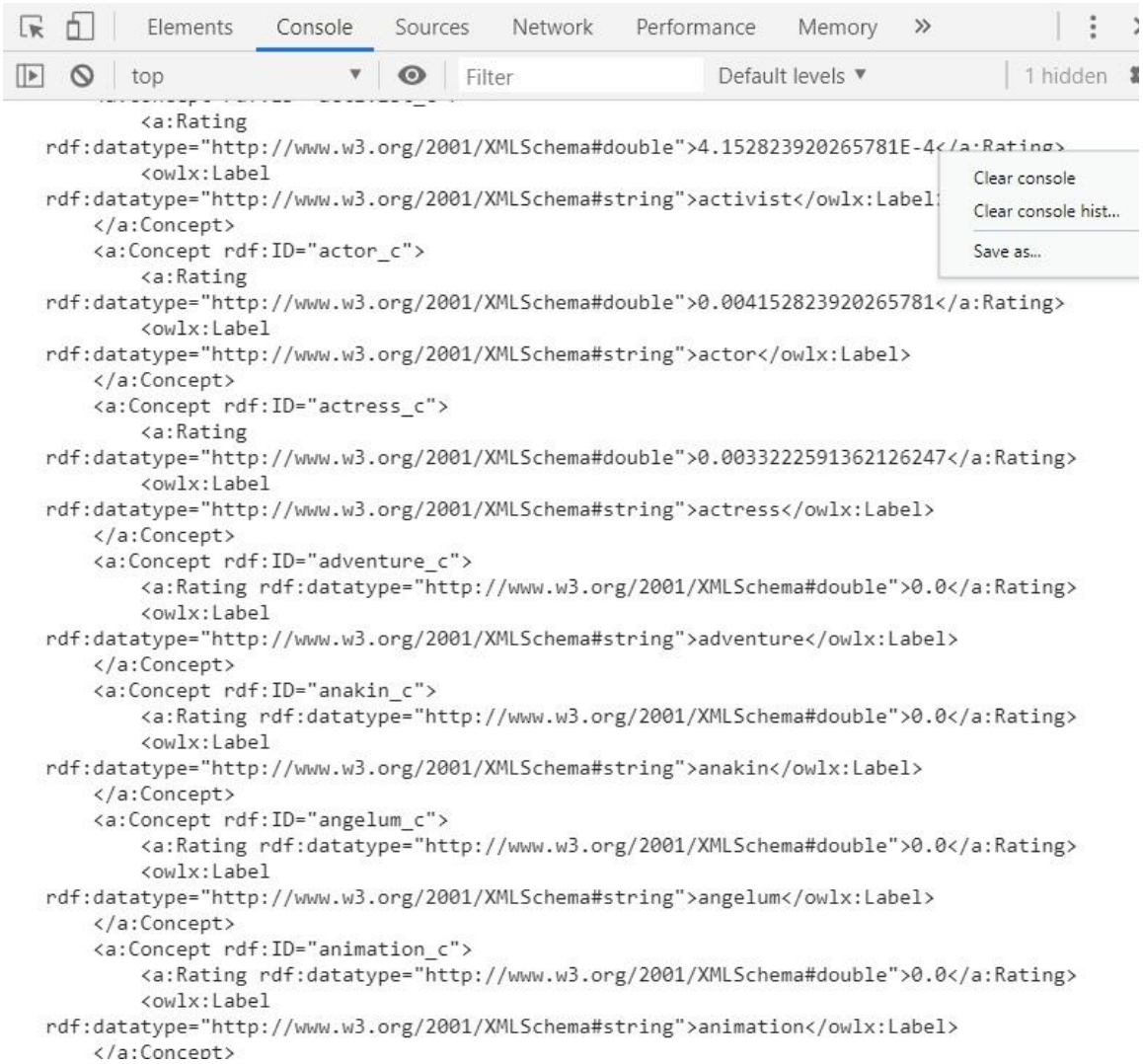
Εικόνα 7-1: Αποτελέσματα της διαδικασίας μέτρησης θεμελιακών στοιχείων

7.3 Η διαδικασία διαγραφής θεμελιακών στοιχείων

Όπως ειπώθηκε στο 5^{ο} κεφάλαιο τα αποτελέσματα των θεμελιακών στοιχείων μοντελοποίησης του Πιθανοτικού Οντολογικού Μοντέλου στο *Text2Onto* μπορούν να φιλτραριστούν, δίνοντας ανατροφοδότηση σε καθένα από αυτά για το αν είναι σκόπιμο να συμπεριληφθούν στην εξαγόμενη οντολογία. Αυτό γίνεται με δεξί κλικ σε κάθε στοιχείο της οντολογίας ξεχωριστά και στη συνέχεια ορίζοντας την κατάλληλη ανατροφοδότηση (*True, False, Don't know*). Δεν υπάρχει όμως η δυνατότητα μαζικής διαγραφής των στοιχείων της οντολογίας όταν για παράδειγμα η τιμή του πεδίου πιθανότητάς τους είναι κάτω από ένα κατώφλι πιθανότητας.

Αυτή η δυνατότητα δίνεται με τη διαδικασία διαγραφής θεμελιακών στοιχείων με την οποία πραγματοποιείται μαζική διαγραφή των στοιχείων της οντολογίας από ένα *owl* αρχείο όταν η τιμή του πεδίου πιθανότητάς τους είναι κάτω από ένα κατώφλι πιθανότητας το οποίο ορίζεται για κάθε θεμελιακό στοιχείο αντίστοιχα. Ο κώδικας περιέχει τρεις συναρτήσεις: τη συνάρτηση *loadOwlDoc* η οποία φορτώνει το *owl* αρχείο και καλεί μια νέα συνάρτηση για την επεξεργασία του. Η συνάρτηση αυτή είναι η *processOwl* η οποία επεξεργάζεται το *DOM* του *owl* αρχείου και στη συνέχεια για κάθε θεμελιακό στοιχείο καλεί μια συνάρτηση για τη διαγραφή στοιχείων οντολογίας με πεδία πιθανότητας κάτω από ένα επιλεγμένο κατώφλι. Η συνάρτηση που καλείται είναι η *deleteNodes* η οποία και διαγράφει τα στοιχεία οντολογίας κάθε θεμελιακού στοιχείου κάτω απ' το επιλεγμένο κατώφλι πιθανότητας.

Η διαδικασία χρησιμοποιεί το πρόσθετο *vkBeautify.js* για τη σωστή εξαγωγή του ενημερωμένου *owl* αρχείου μετά τις διαγραφές. Το πρόσθετο είναι διαθέσιμο στον παρακάτω σύνδεσμο: *<https://github.com/vkiryukhin/vkBeautify>*. Το ενημερωμένο *owl* αρχείο εξάγεται από την κονσόλα του προγράμματος περιήγησης όπως φαίνεται ενδεικτικά στην εικόνα 7-2 που ακολουθεί.



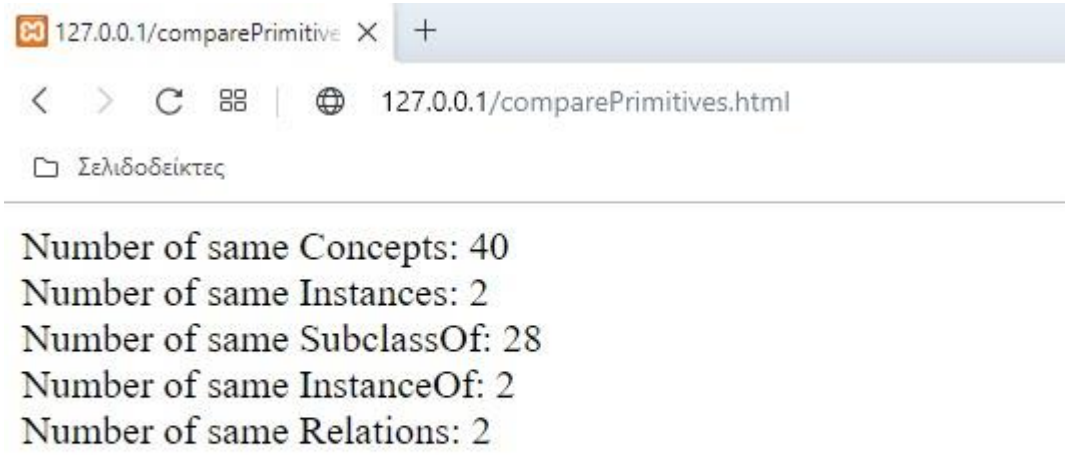
Εικόνα 7-2: Εξαγόμενη οντολογία της διαδικασίας διαγραφής θεμελιακών στοιχείων

7.4 Η διαδικασία σύγκρισης θεμελιακών στοιχείων

Η σύγκριση των εξαγόμενων οντολογιών από δύο διαφορετικά πειράματα ως προς τα κοινά στοιχεία της οντολογίας για κάθε θεμελιακό στοιχείο ήταν η επόμενη κατεύθυνση ως προς την ανάλυση των αποτελεσμάτων των πειραμάτων.

Για την επίτευξη αυτού του σκοπού αναπτύχθηκε η διαδικασία σύγκρισης θεμελιακών στοιχείων με την οποία συγκρίνονται τα *owl* αρχεία που έχουν εξαχθεί από δύο διαφορετικά πειράματα ως προς τον αριθμό των κοινών στοιχείων της οντολογίας για κάθε θεμελιακό στοιχείο αντίστοιχα. Ο κώδικας περιέχει έξι συναρτήσεις. Αρχικά η συνάρτηση *loadOwlDoc1* φορτώνει το πρώτο *owl* αρχείο και καλεί μια συνάρτηση για την επεξεργασία του. Η συνάρτηση αυτή είναι η *processOwl1* η οποία με τη σειρά της καλεί δύο συναρτήσεις, μία για την επεξεργασία του *DOM* του πρώτου *owl* αρχείου και

μία δεύτερη για να φορτωθεί το δεύτερο *owl* αρχείο. Η πρώτη συνάρτηση είναι η *processOwlDom* η οποία επεξεργάζεται το *DOM* και των δύο *owl* αρχείων. Η δεύτερη συνάρτηση είναι η *loadOwlDoc2* που φορτώνει το δεύτερο *owl* αρχείο και καλεί μια άλλη συνάρτηση για την επεξεργασία του. Η συνάρτηση αυτή είναι η *processOwl2* που καλεί τη συνάρτηση *processOwlDom* για την επεξεργασία του *DOM* του δεύτερου *owl* αρχείου και στη συνέχεια καλεί μια άλλη συνάρτηση για τη σύγκριση των στοιχείων οντολογίας των δύο *owl* αρχείων. Η συνάρτηση αυτή είναι η *comparePrim* που ανά θεμελιακό στοιχείο συγκρίνει τα στοιχεία οντολογίας των δύο *owl* αρχείων για την εύρεση κοινών στοιχείων. Στην εικόνα 7-3 που ακολουθεί φαίνονται ενδεικτικά τα αποτελέσματα όπως εμφανίζονται στο πρόγραμμα περιήγησης.



Εικόνα 7-3: Αποτελέσματα της διαδικασίας σύγκρισης θεμελιακών στοιχείων

7.5 Η διαδικασία πιθανοτικής ομαδοποίησης θεμελιακών στοιχείων

Ο επόμενος στόχος της ανάλυσης των αποτελεσμάτων των πειραμάτων ήταν να ομαδοποιηθούν ποσοτικά τα στοιχεία της εξαγόμενης οντολογίας ενός πειράματος ανάλογα με την τιμή του πεδίου πιθανότητάς τους σύμφωνα με κάποια προκαθορισμένα επίπεδα πιθανότητας για κάθε θεμελιακό στοιχείο αντίστοιχα.

Για το σκοπό αυτό αναπτύχθηκε η διαδικασία πιθανοτικής ομαδοποίησης θεμελιακών στοιχείων. Ο κώδικας περιέχει τρεις συναρτήσεις: τη συνάρτηση *loadOwlDoc* η οποία φορτώνει το *owl* αρχείο και καλεί μια νέα συνάρτηση για την επεξεργασία του. Η συνάρτηση αυτή είναι η *processOwl* η οποία επεξεργάζεται το *DOM* του *owl* αρχείου και στη συνέχεια για κάθε θεμελιακό στοιχείο καλεί μια συνάρτηση για

την ομαδοποίηση των στοιχείων οντολογίας τους σύμφωνα με κάποια επιλεγμένα επίπεδα πιθανότητας. Η συνάρτηση που καλείται είναι η *probResults* η οποία ομαδοποιεί τα στοιχεία οντολογίας κάθε θεμελιακού στοιχείου σύμφωνα με τα επιλεγμένα επίπεδα πιθανότητας και εξάγει το άθροισμα αυτών των στοιχείων για κάθε επίπεδο. Ενδεικτικά αποτελέσματα, όπως αυτά εμφανίζονται στο πρόγραμμα περιήγησης, φαίνονται στην εικόνα 7-4 που ακολουθεί.

Εικόνα 7-4: Αποτελέσματα της διαδικασίας πιθανοτικής ομαδοποίησης θεμελιακών στοιχείων

7.6 Η διαδικασία μετατροπής θεμελιακών στοιχείων

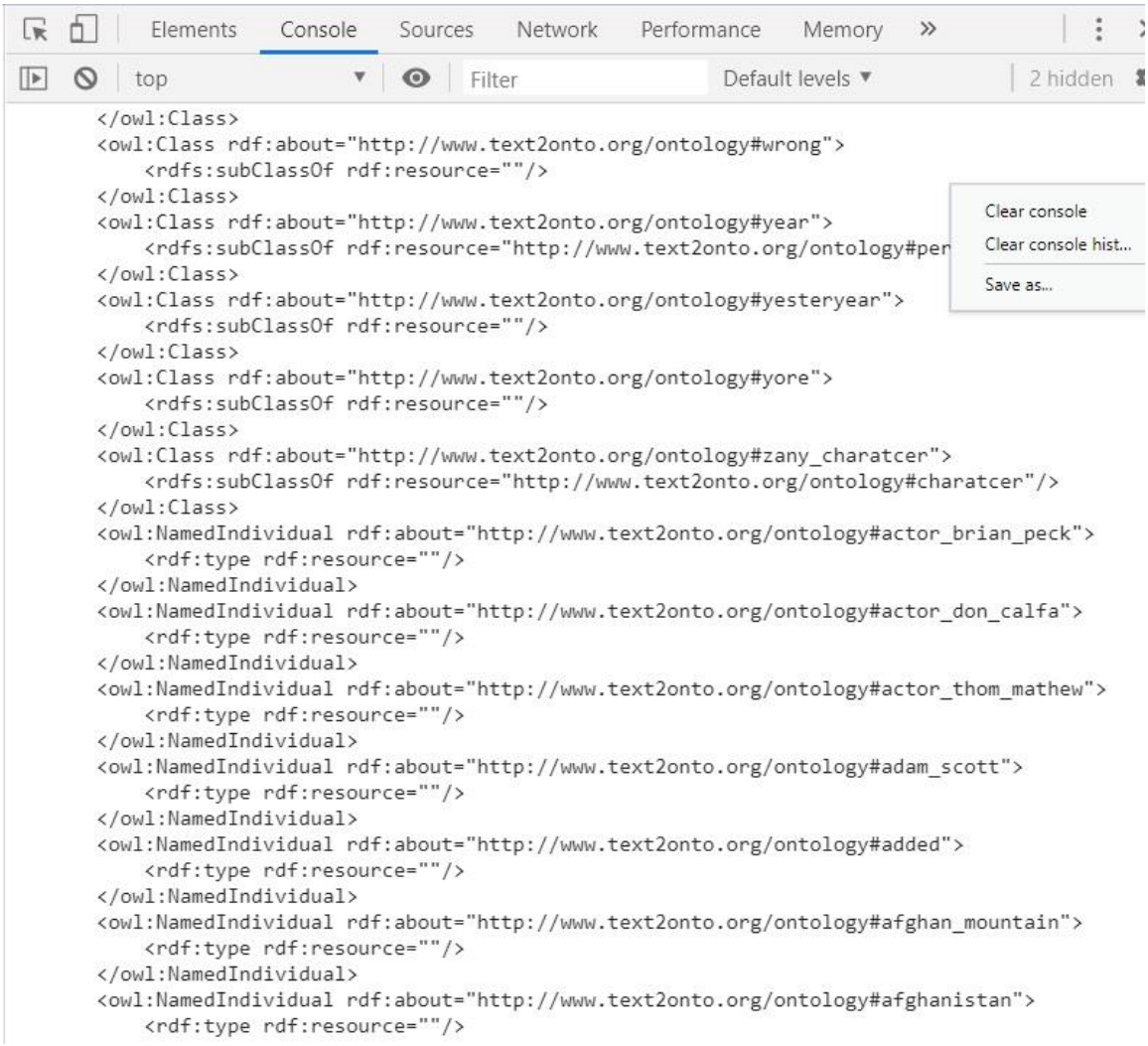
Ένας από τους αρχικούς στόχους της μελέτης ήταν να χρησιμοποιηθεί η εξαγόμενη οντολογία από τα πειράματα με την εφαρμογή *Text2Onto* ως είσοδος στο περιβάλλον οντολογιών *Protégé* για να υπάρχει η δυνατότητα περαιτέρω επεξεργασίας της.

Αυτό που παρατηρήθηκε με τη χρησιμοποίηση των εξαγόμενων από τα πειράματα *owl* αρχείων ως είσοδο στο *Protégé*, ήταν ότι η οντολογία δεν εμφανιζόταν στο περιβάλλον του με τον τρόπο που θα έπρεπε να εμφανίζεται. Πιο συγκεκριμένα οι σχέσεις μεταξύ ατόμων των κλάσεων της οντολογίας δεν εμφανίζονται ως ιδιότητες αντικειμένου στην καρτέλα "*Object Properties*" αλλά ως κλάσεις στην καρτέλα "*Classes*". Επιπλέον τούτου στην ιεραρχία των κλάσεων δεν εμφανιζόταν σωστή ιεράρχηση των κλάσεων-υποκλάσεων ως προς το θεμελιακό στοιχείο *SubclassOf* του αρχείου εισόδου. Επίσης ως προς το θεμελιακό στοιχείο *InstanceOf* δεν εμφανιζόταν στις καρτέλες σωστά η αντιστοίχιση του ατόμου με την αντίστοιχη κλάση στην οποία ανήκει.

Για τη διόρθωση των παραπάνω κρίθηκε σκόπιμο να αναπτυχθεί μια διαδικασία που να μετατρέπει τα *owl* αρχεία που εξάγονται από τα πειράματα με το *Text2Onto* σε *owl* αρχεία που να χρησιμοποιούνται αυτά ως είσοδος στο *Protégé* ώστε να εμφανίζεται στο περιβάλλον του η οντολογία στη σωστή της μορφή.

Η διαδικασία που αναπτύχθηκε είναι αυτή της μετατροπής θεμελιακών στοιχείων. Ο κώδικας περιέχει τέσσερις συναρτήσεις: τη συνάρτηση *loadText2OntoOwl* η οποία φορτώνει το *owl* αρχείο που εξάγεται από το *Text2Onto* και καλεί μια νέα συνάρτηση για την επεξεργασία του. Η συνάρτηση αυτή είναι η *processText2OntoOwl* η οποία επεξεργάζεται το *DOM* του *owl* αρχείου που εξάγεται από το *Text2Onto* και στη συνέχεια καλεί μια άλλη συνάρτηση για να φορτώσει ένα πρότυπο *owl* αρχείο επάνω στο οποίο θα δομηθεί το νέο *owl* αρχείο που θα είναι κατάλληλο για είσοδο στο *Protégé*. Η συνάρτηση αυτή είναι η *loadTemplateOwl* η οποία με τη σειρά της φορτώνει το πρότυπο *owl* αρχείο και καλεί μια νέα συνάρτηση για την επεξεργασία του. Η συνάρτηση που καλείται είναι η *processTemplateOwl* η οποία και τροφοδοτεί το πρότυπο *owl* αρχείο με τα δεδομένα του *owl* αρχείου που εξάγεται από το *Text2Onto*.

Η διαδικασία χρησιμοποιεί το πρόσθετο *vkBeautify.js* για τη σωστή εξαγωγή του νέου *owl* αρχείου μετά την μετατροπή. Το νέο *owl* αρχείο εξάγεται από την κονσόλα του προγράμματος περιήγησης όπως φαίνεται ενδεικτικά στην εικόνα 7-5 που ακολουθεί.



Εικόνα 7-5: Εξαγόμενη οντολογία της διαδικασίας μετατροπής θεμελιακών στοιχείων

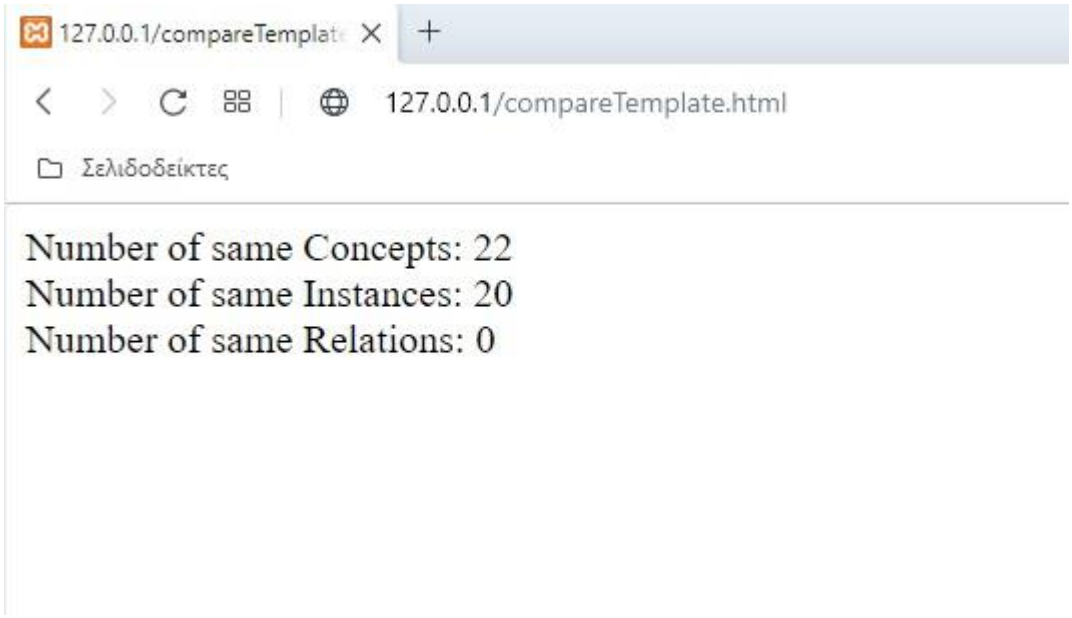
7.7 Η διαδικασία σύγκρισης με πρότυπο οντολογίας

Ένας επιπλέον αρχικός στόχος της μελέτης ήταν η σύγκριση της εξαγόμενης οντολογίας από τα πειράματα με την εφαρμογή *Text2Onto* με ένα πρότυπο οντολογίας σχετικά με κινηματογραφικές ταινίες ως προς τον αριθμό των κοινών στοιχείων των οντολογιών για κάθε θεμελιακό στοιχείο αντίστοιχα.

Για το σκοπό αυτό αναπτύχθηκε μια διαδικασία η οποία συγκρίνει το αρχείο *owl* που παίρνουμε ως έξοδο από τα πειράματα αφού πρώτα μετατραπεί σε μορφή κατάλληλη (μέσω της διαδικασίας μετατροπής θεμελιακών στοιχείων) με ένα πρότυπο αρχείο οντολογίας *owl* σχετικά με κινηματογραφικές ταινίες (Amancio Bouza (2010)). Η σύγκριση αφορά την εύρεση κοινών στοιχείων ανάμεσα στα δύο αρχεία οντολογίας όσον αφορά τα θεμελιακά στοιχεία των κλάσεων, των ατόμων και των σχέσεων. Το παραπάνω

πρότυπο οντολογίας φαίνεται αναλυτικά μέσω του περιβάλλοντος οντολογίας Protégé στο 4ο κεφάλαιο.

Η διαδικασία που αναπτύχθηκε είναι αυτή της σύγκρισης με πρότυπο οντολογίας. Ο κώδικας περιέχει έξι συναρτήσεις. Αρχικά η συνάρτηση *loadOwlDoc1* φορτώνει το *owl* αρχείο που εξάγεται από το *Text2Onto* και καλεί μια συνάρτηση για την επεξεργασία του. Η συνάρτηση αυτή είναι η *processOwl1* η οποία με τη σειρά της καλεί δύο συναρτήσεις, μία για την επεξεργασία του *DOM* του *owl* αρχείου που εξάγεται από το *Text2Onto* και μια δεύτερη για να φορτωθεί το πρότυπο αρχείο οντολογίας *owl*. Η πρώτη συνάρτηση είναι η *processOwlDom* η οποία επεξεργάζεται το *DOM* και των δύο *owl* αρχείων. Η δεύτερη συνάρτηση είναι η *loadOwlDoc2* που φορτώνει το πρότυπο αρχείο οντολογίας *owl* και καλεί μια άλλη συνάρτηση για την επεξεργασία του. Η συνάρτηση αυτή είναι η *processOwl2* που καλεί τη συνάρτηση *processOwlDom* για την επεξεργασία του *DOM* του πρότυπου αρχείου οντολογίας *owl* και στη συνέχεια καλεί μια άλλη συνάρτηση για τη σύγκριση των στοιχείων οντολογίας των δύο *owl* αρχείων. Η συνάρτηση αυτή είναι η *comparePrim* που ανά θεμελιακό στοιχείο συγκρίνει τα στοιχεία οντολογίας των δύο *owl* αρχείων για την εύρεση κοινών στοιχείων. Στην εικόνα 7-6 που ακολουθεί φαίνονται ενδεικτικά τα αποτελέσματα όπως εμφανίζονται στο πρόγραμμα περιήγησης.



Εικόνα 7-6: Αποτελέσματα της διαδικασίας σύγκρισης με πρότυπο οντολογίας

8 Γραφήματα – Σχολιασμός – Συμπεράσματα

8.1 Μια συνοπτική θεώρηση

Οι διαδικασίες που αναπτύχθηκαν και περιγράφηκαν στο προηγούμενο κεφάλαιο έδωσαν τη δυνατότητα για αρκετά μεγάλη ποικιλία ως προς την ποσοτική και ποιοτική ανάλυση των αποτελεσμάτων των πειραμάτων. Ως επακόλουθο με τη χρήση των παραπάνω διαδικασιών πραγματοποιήθηκε στη συνέχεια μια πληθώρα αναλύσεων των αποτελεσμάτων ως προς κάποιες συγκεκριμένες κατευθύνσεις και με διάφορους συνδυασμούς των αποτελεσμάτων. Για τις αναλύσεις αυτές έγινε εξαγωγή των αποτελεσμάτων σε λογιστικά φύλλα εργασίας και δημιουργήθηκαν για κάθε περίπτωση τα αντίστοιχα γραφήματα.

Μια βασική δομή των παραπάνω κατευθύνσεων δίνεται συνοπτικά παρακάτω:

- Μέτρηση του αριθμού των στοιχείων της οντολογίας για κάθε θεμελιακό στοιχείο των *owl* αρχείων που παίρνουμε ως έξοδο από τα πειράματα με το *Text2Onto*.
- Σύγκριση των *owl* αρχείων που έχουν εξαχθεί από δύο διαφορετικά πειράματα ως προς τον αριθμό των κοινών στοιχείων της οντολογίας για κάθε θεμελιακό στοιχείο αντίστοιχα με εύρεση ενός ποσοστιαίου λόγου των κοινών στοιχείων.
- Ποσοτική ομαδοποίηση των στοιχείων της οντολογίας των *owl* αρχείων ανάλογα με την τιμή του πεδίου πιθανότητάς τους σύμφωνα με κάποια προκαθορισμένα επίπεδα πιθανότητας, για κάθε θεμελιακό στοιχείο αντίστοιχα, με εύρεση ενός ποσοστιαίου λόγου για κάθε ομαδοποίηση στοιχείων.
- Σύγκριση των *owl* αρχείων που παίρνουμε ως έξοδο από τα πειράματα με το *Text2Onto* με ένα πρότυπο *owl* αρχείο μιας οντολογίας σχετικά με κινηματογραφικές ταινίες ως προς τον αριθμό των κοινών στοιχείων των οντολογιών για τα βασικά θεμελιακά στοιχεία (*Concept*, *Instance*, *Relation*).
- Χρησιμοποίηση των εξαγόμενων *owl* αρχείων ως είσοδο στο περιβάλλον οντολογιών *Protégé* για δυνατότητα περαιτέρω επεξεργασίας της εξαγόμενης οντολογίας και οπτικοποίησής της.

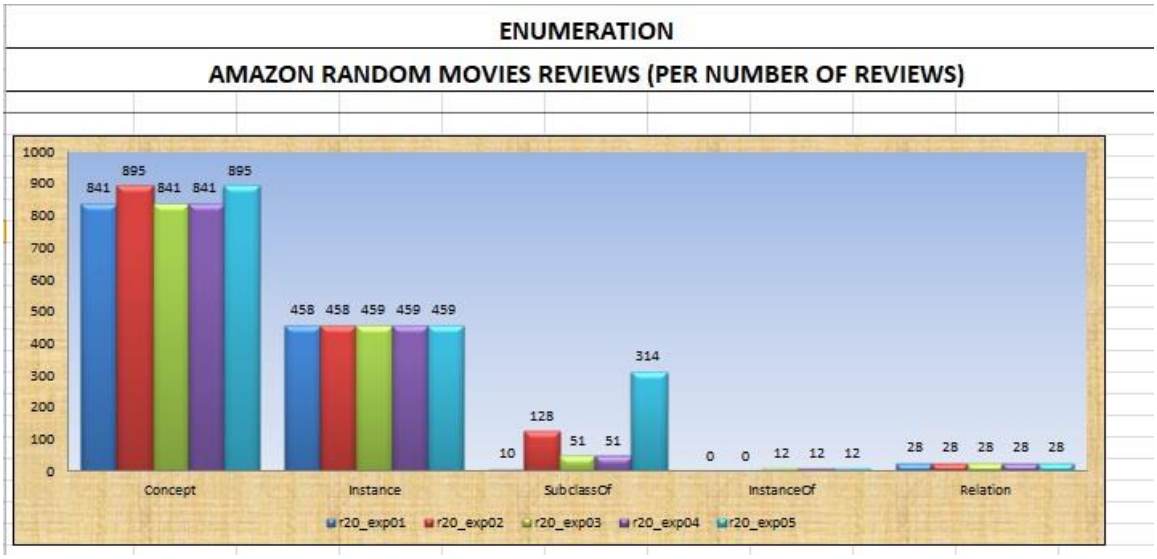
Για τις παραπάνω κατευθύνσεις η ανάλυση πραγματοποιήθηκε χρησιμοποιώντας διάφορους συνδυασμούς αποτελεσμάτων των πειραμάτων. Με βάση αυτούς τους συνδυασμούς δημιουργήθηκαν και τα αντίστοιχα γραφήματα. Μια συνοπτική δομή αυτών των συνδυασμών δίνεται παρακάτω:

- Τέσσερις διαφορετικές κατηγορίες δεδομένων εισόδου (*amazon random movies reviews, amazon same movies reviews, imdb movies reviews, imdb movies data*), όπως αυτές περιγράφηκαν στο κεφάλαιο 6.
- Πέντε διαφορετικά είδη πειραμάτων ως προς τον συνδυασμό των αλγορίθμων. Τα είδη αυτά περιγράφηκαν στο κεφάλαιο 6.
- Ποσοτική μεταβολή των δεδομένων εισόδου, της τάξης των 20,100,1000. Στο σημείο αυτό να τονιστεί, όπως αναφέρθηκε και στο κεφάλαιο 6, πως η ποσοτική μεταβολή των δεδομένων εισόδου αφορά τις κατηγορίες δεδομένων εισόδου *amazon_random* και *imdb_reviews*. Τα δεδομένα εισόδου στις κατηγορίες *amazon_same* και *imdb_movies_data* είναι της τάξης των 1000.
- Γραφήματα ανά αριθμό κριτικών ή εγγραφών δεδομένων και γραφήματα ανά πείραμα με ποσοτική μεταβολή των δεδομένων.
- Αντίστοιχα γραφήματα με μαζική διαγραφή από τα *owl* αρχεία, των στοιχείων της οντολογίας όπου η τιμή του πεδίου πιθανότητάς τους είναι κάτω από ένα κατώφλι πιθανότητας το οποίο ορίζεται για κάθε θεμελιακό στοιχείο αντίστοιχα.
- Συγκρίσεις μεταξύ διαφορετικών κατηγοριών δεδομένων εισόδου με βάση ένα συγκεκριμένο είδος πειράματος συνδυασμού των αλγορίθμων.

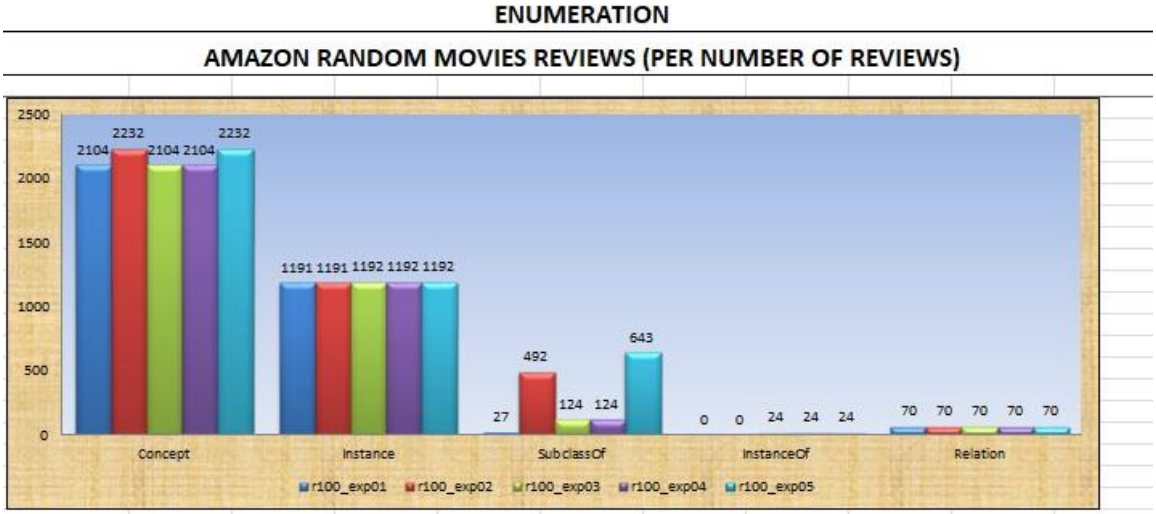
Λόγω της πληθώρας των περιπτώσεων στη συνέχεια αυτού του κεφαλαίου αναλύονται και σχολιάζονται περαιτέρω οι πιο ενδιαφέρουσες από τις αναλύσεις αποτελεσμάτων των πειραμάτων. Στο Παράρτημα Α στο τέλος της εργασίας παρατίθενται τα αντίστοιχα γραφήματα για όλες τις αναλύσεις των αποτελεσμάτων των πειραμάτων που πραγματοποιήθηκαν.

8.2 Μέτρηση αριθμού στοιχείων οντολογίας

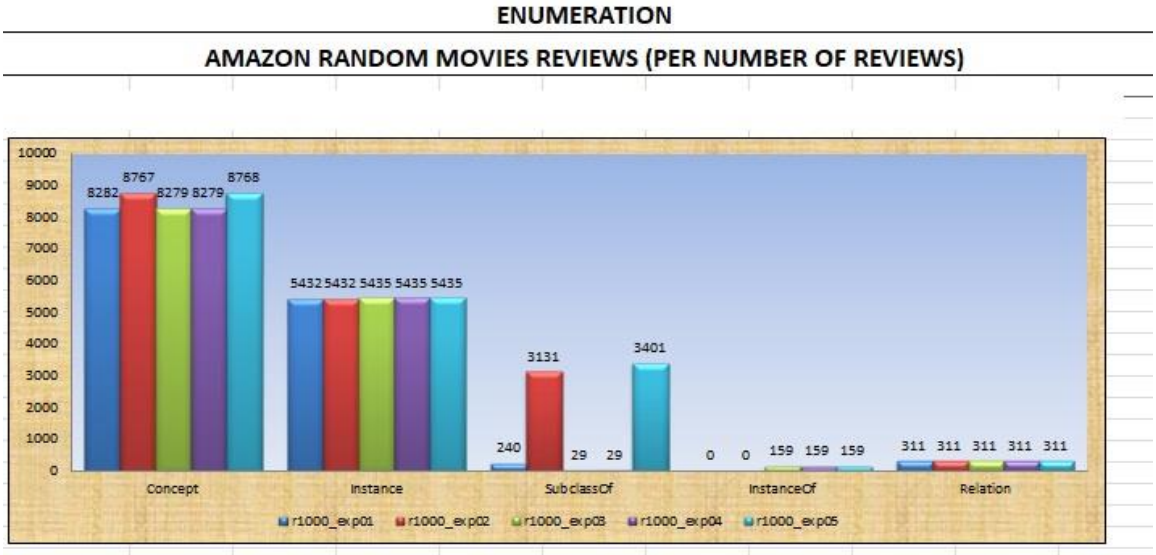
Στη συγκεκριμένη ανάλυση αποτελεσμάτων των πειραμάτων έγινε μέτρηση του αριθμού των στοιχείων της οντολογίας για κάθε θεμελιακό στοιχείο από τα *owl* αρχεία και των τεσσάρων κατηγοριών δεδομένων εισόδου. Στην ενότητα αυτή αναλύονται περαιτέρω τα γραφήματα που εξάχθηκαν και αφορούν την κατηγορία δεδομένων εισόδου *amazon_random*. Τα πρώτα γραφήματα από τη συγκεκριμένη ανάλυση αφορούν όλα τα θεμελιακά στοιχεία και των πέντε ειδών πειραμάτων. Υπάρχουν τρία γραφήματα ανά αριθμό κριτικών αντίστοιχα. Τα γραφήματα αυτά φαίνονται παρακάτω στις εικόνες 8-1 έως 8-3.



Εικόνα 8-1: Μέτρηση αριθμού στοιχείων οντολογίας – αριθμός κριτικών 20



Εικόνα 8-2: Μέτρηση αριθμού στοιχείων οντολογίας – αριθμός κριτικών 100



Εικόνα 8-3: Μέτρηση αριθμού στοιχείων οντολογίας – αριθμός κριτικών 1000

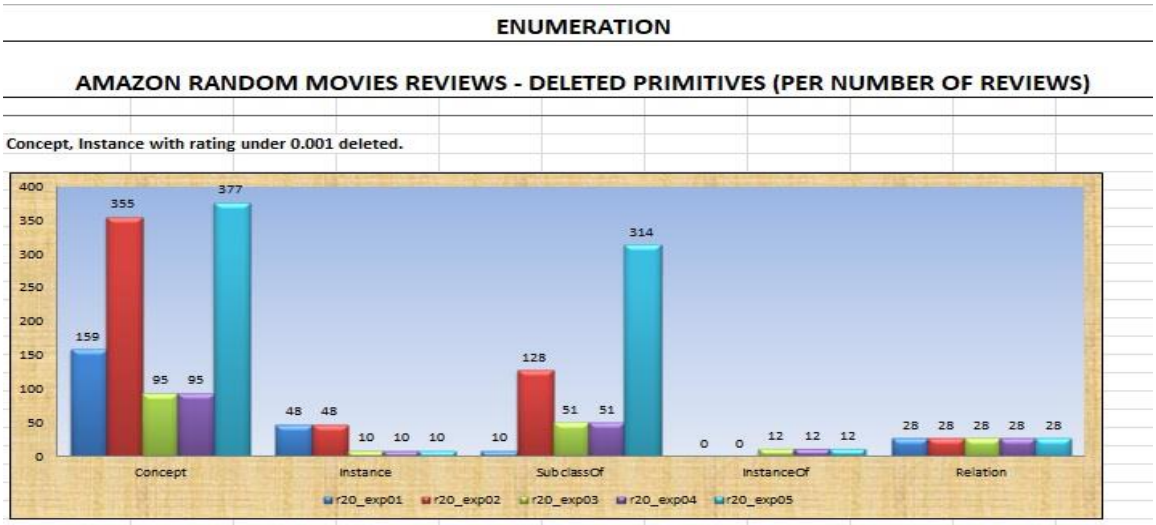
Από τα γραφήματα αυτά μπορούν να εξαχθούν κάποια πρώτα συμπεράσματα. Όσον αφορά τα θεμελιακά στοιχεία *Concept* και *Instance* παρατηρούμε σε γενικές γραμμές μικρές διαφοροποιήσεις σε όλα τα γραφήματα για καθένα από τα πέντε είδη πειραμάτων. Ως προς το θεμελιακό στοιχείο *SubclassOf* παρατηρούμε ότι στο δεύτερο και στο πέμπτο είδος πειραμάτων έχουμε αυξημένο αριθμό αποτελεσμάτων σε σχέση με τα υπόλοιπα. Για το θεμελιακό στοιχείο *InstanceOf* βλέπουμε ότι τα δύο πρώτα είδη πειραμάτων δεν μας δίνουν καθόλου αποτελέσματα. Τέλος για το θεμελιακό στοιχείο *Relation* σε κάθε γράφημα τα αποτελέσματα είναι τα ίδια καθώς υπάρχει μόνο ένας διαθέσιμος αλγόριθμος που χρησιμοποιείται και στα πέντε είδη πειραμάτων.

Όπως αναφέρθηκε στο κεφάλαιο 6 και θα φανεί στη συνέχεια του κεφαλαίου, από τα γραφήματα που αφορούν τις τιμές πιθανότητας των στοιχείων οντολογίας, για τα θεμελιακά στοιχεία *Concept* και *Instance*, παρατηρήθηκε σε όλες τις περιπτώσεις των πειραμάτων ότι οι τιμές πιθανότητας των στοιχείων οντολογίας τους είναι αρκετά χαμηλές σε σχέση και με τα υπόλοιπα θεμελιακά στοιχεία. Επίσης παρατηρήθηκε για τα θεμελιακά στοιχειά *Concept* και *Instance* καθώς και για το δεύτερο και το πέμπτο είδος πειράματος του θεμελιακού στοιχείου *SubclassOf*, πως ο αριθμός των οντολογικών στοιχείων είναι πολύ μεγάλος σε όλα τα πειράματα και οι αυξητικές τους τάσεις είναι επίσης πολύ μεγάλες καθώς αυξάνεται ο αριθμός των κριτικών από 20 σε 100 και από 100 σε 1000. Κάτι τέτοιο δεν είναι ιδιαίτερα χρήσιμο στο χτίσιμο μιας σωστής οντολογίας γιατί το τοπίο γίνεται ιδιαίτερα χαοτικό και λόγω των μικρών τιμών πιθανοτήτων των στοιχείων τα περισσότερα από αυτά είναι μη αξιοποιήσιμα.

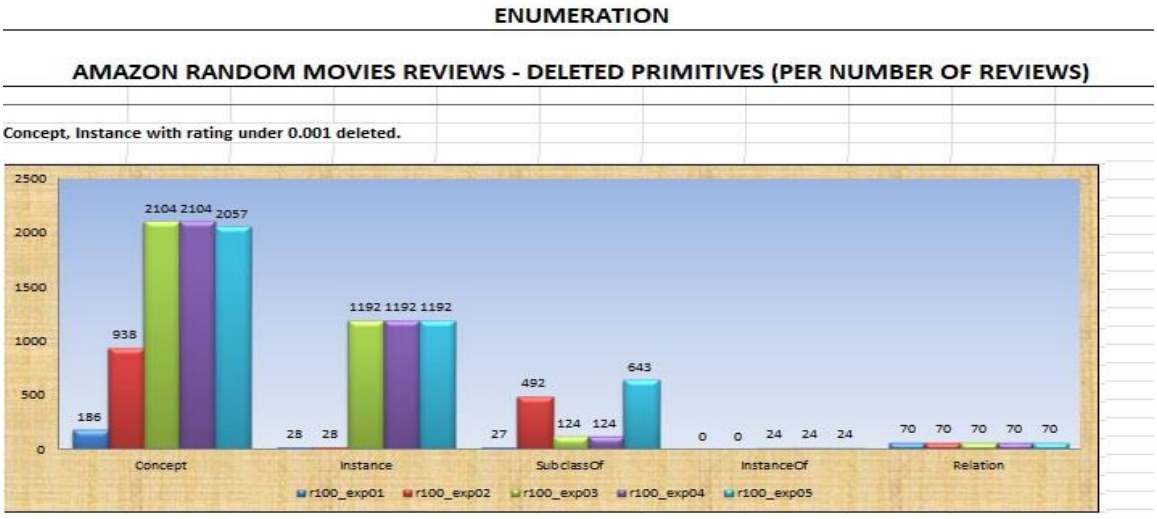
Με βάση τις πιο πάνω παρατηρήσεις κρίθηκε σκόπιμο ως επόμενο βήμα για τη συγκεκριμένη ανάλυση να εξαχθούν αντίστοιχα αποτελέσματα, αφού έχει προηγηθεί μαζική διαγραφή από τα *owl* αρχεία, των στοιχείων της οντολογίας για τα θεμελιακά στοιχεία *Concept* και *Instance* όπου η τιμή του πεδίου πιθανότητάς τους είναι κάτω από ένα κατώφλι πιθανότητας. Στην προκειμένη περίπτωση το κατώφλι πιθανότητας ορίστηκε σε 0,001. Στη συνέχεια πραγματοποιήθηκε και πάλι μέτρηση του αριθμού των στοιχείων της οντολογίας για κάθε θεμελιακό στοιχείο και για τις τέσσερις κατηγορίες δεδομένων εισόδου. Στις εικόνες 8-4 έως 8-6 που ακολουθούν, φαίνονται τα αντίστοιχα γραφήματα με αυτά των εικόνων 8-1 έως 8-3 μετά τις διαγραφές.

Οι διαφοροποιήσεις σε σχέση με τα προηγούμενα γραφήματα, όπως είναι φυσικό επακόλουθο εμφανίζονται όσον αφορά τα θεμελιακά στοιχεία *Concept* και *Instanc*e όπου και πραγματοποιήθηκαν οι διαγραφές. Παρατηρούμε και για τα δύο θεμελιακά στοιχεία

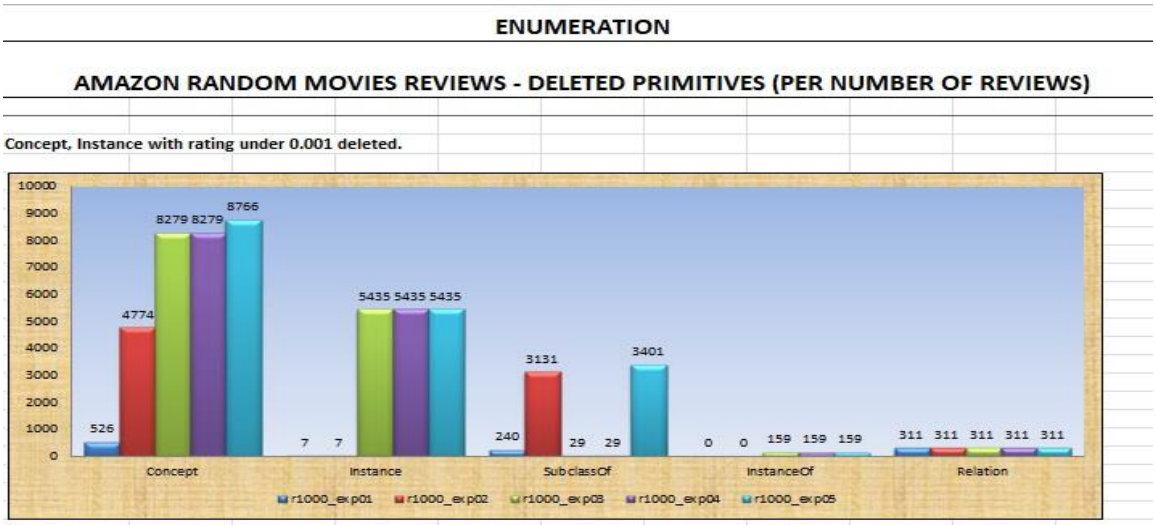
ότι, αντίθετα με πριν, υπάρχει διαφοροποίηση του αριθμού των στοιχείων ανάλογα με το είδος πειράματος και στα τρία γραφήματα. Στο πρώτο γράφημα παρατηρούμε ότι υπάρχει αισθητή πτώση του αριθμού των στοιχείων και για τα πέντε είδη πειραμάτων, κάτι που δείχνει και την πολύ χαμηλή τιμή της πιθανότητας των στοιχείων για τα πειράματα με είκοσι κριτικές. Αντίθετα στα επόμενα δύο γραφήματα, όπου ο αριθμός των κριτικών αυξάνεται, παρατηρούμε ότι ο αριθμός των στοιχείων διατηρείται στα ίδια επίπεδα με προηγουμένως μόνο για τα τρία τελευταία είδη πειραμάτων. Ειδικότερα για το θεμελιακό στοιχείο *Instance* παρατηρούμε επίσης ότι για τα δύο πρώτα είδη πειραμάτων ο αριθμός των στοιχείων μειώνεται αντί να αυξάνεται όσο αυξάνονται οι κριτικές ανάμεσα στα τρία γραφήματα.



Εικόνα 8-4: Μέτρηση αριθμού στοιχείων οντολογίας με διαγραφή θεμελιακών στοιχείων – αριθμός κριτικών 20



Εικόνα 8-5: Μέτρηση αριθμού στοιχείων οντολογίας με διαγραφή θεμελιακών στοιχείων – αριθμός κριτικών 100



Εικόνα 8-6: Μέτρηση αριθμού στοιχείων οντολογίας με διαγραφή θεμελιακών στοιχείων – αριθμός κριτικών 1000

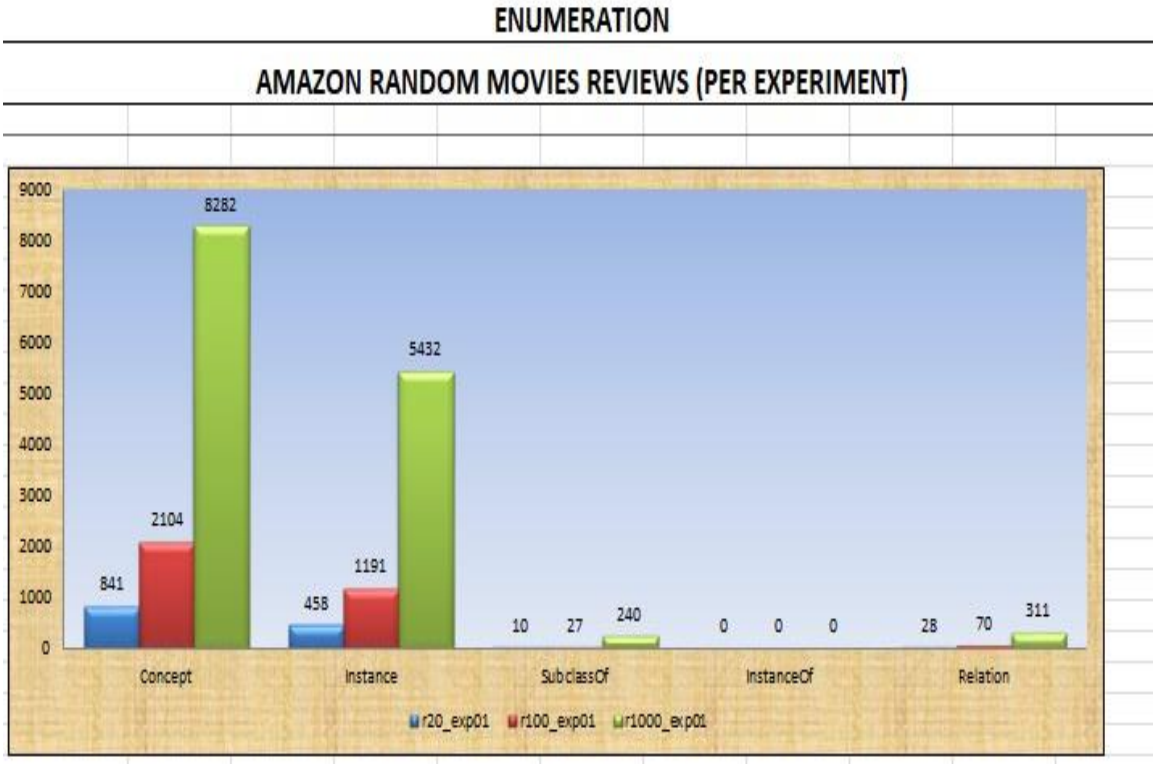
Συνοψίζοντας όσον αφορά τις μετρήσεις ως προς τον αριθμό των κριτικών και για τα πέντε είδη πειραμάτων, μετά τις παρατηρήσεις των παραπάνω γραφημάτων, ένα βασικό συμπέρασμα που μπορούμε να βγάλουμε είναι όσον αφορά την επιλογή των αλγορίθμων.

Ως προς το θεμελιακό στοιχείο *Concept*, οι αλγόριθμοι *RTFConceptExtraction* και *TFIDFConceptExtraction* που χρησιμοποιούνται στο τρίτο και στο τέταρτο είδος πειραμάτων βλέπουμε ότι δίνουν καλύτερα αποτελέσματα καθώς ο αριθμός των στοιχείων διατηρείται στα ίδια επίπεδα μετά τη διαγραφή των στοιχείων με μικρότερες πιθανότητες. Ειδικότερα είναι γνωστό πως ο αλγόριθμος *TFIDFConceptExtraction* εξάγει τις πιο σημαντικές έννοιες όσον αφορά έναν δεδομένο τομέα ενδιαφέροντος (Johanna Volker, Denny Vrandecic, York Sure (2005, σ.14)). Ως προς το θεμελιακό στοιχείο *Instance* ο αλγόριθμος *TFIDFInstanceExtraction* που χρησιμοποιείται στο τρίτο και στο τέταρτο είδος πειραμάτων, βλέπουμε ότι και αυτός δίνει καλύτερα αποτελέσματα καθώς και σε αυτήν την περίπτωση ο αριθμός των στοιχείων διατηρείται στα ίδια επίπεδα μετά τη διαγραφή των στοιχείων με μικρότερες πιθανότητες. Για τους παραπάνω αλγορίθμους που διατηρούν τα αριθμητικά τους μεγέθη και μετά τις διαγραφές, περαιτέρω μείωση στο κατώφλι πιθανότητας πολύ πιθανόν να μειώσει τα αριθμητικά αποτελέσματα. Διάφοροι πειραματισμοί με την επιλογή του κατωφλίου πιθανότητας και την παρατήρηση των αριθμητικών αποτελεσμάτων που αντίστοιχα παίρνουμε, μπορεί να οδηγήσει στο χτίσιμο μιας πιο σωστής οντολογίας. Όσον αφορά το θεμελιακό στοιχείο *SubclassOf* βλέπουμε πως ο αλγόριθμος *VerticalRelationsConceptClassification* που χρησιμοποιείται στο δεύτερο είδος πειράματος δίνει πολύ αυξημένα αριθμητικά

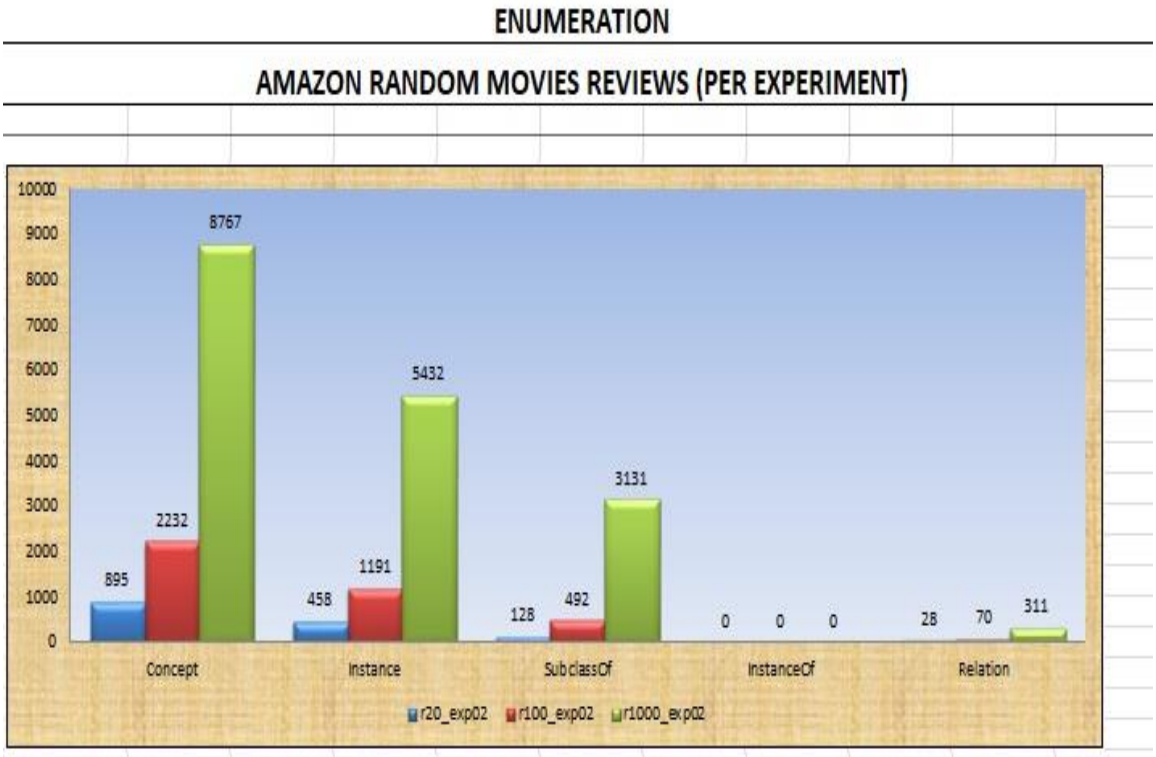
αποτελέσματα σε σχέση με τους υπόλοιπους με τις τιμές πιθανότητας των αποτελεσμάτων να είναι σε υψηλά επίπεδα. Αντίστοιχα όσον αφορά το θεμελιακό στοιχείο *InstanceOf* ο αλγόριθμος *PatternInstanceClassification* του τρίτου και τέταρτου είδους πειραμάτων είναι ο βέλτιστος καθώς τα πρώτα δύο είδη πειραμάτων δεν μας έδωσαν καθόλου αποτελέσματα.

Τα στοιχεία αυτά που αφορούν την επιλογή των αλγορίθμων και που μπορούν να δώσουν τα βέλτιστα δυνατά αποτελέσματα θα επαληθευτούν και παρακάτω στην ενότητα που αφορά την ποσοτική ομαδοποίηση των στοιχείων της οντολογίας ανάλογα με την τιμή του πεδίου πιθανότητάς τους.

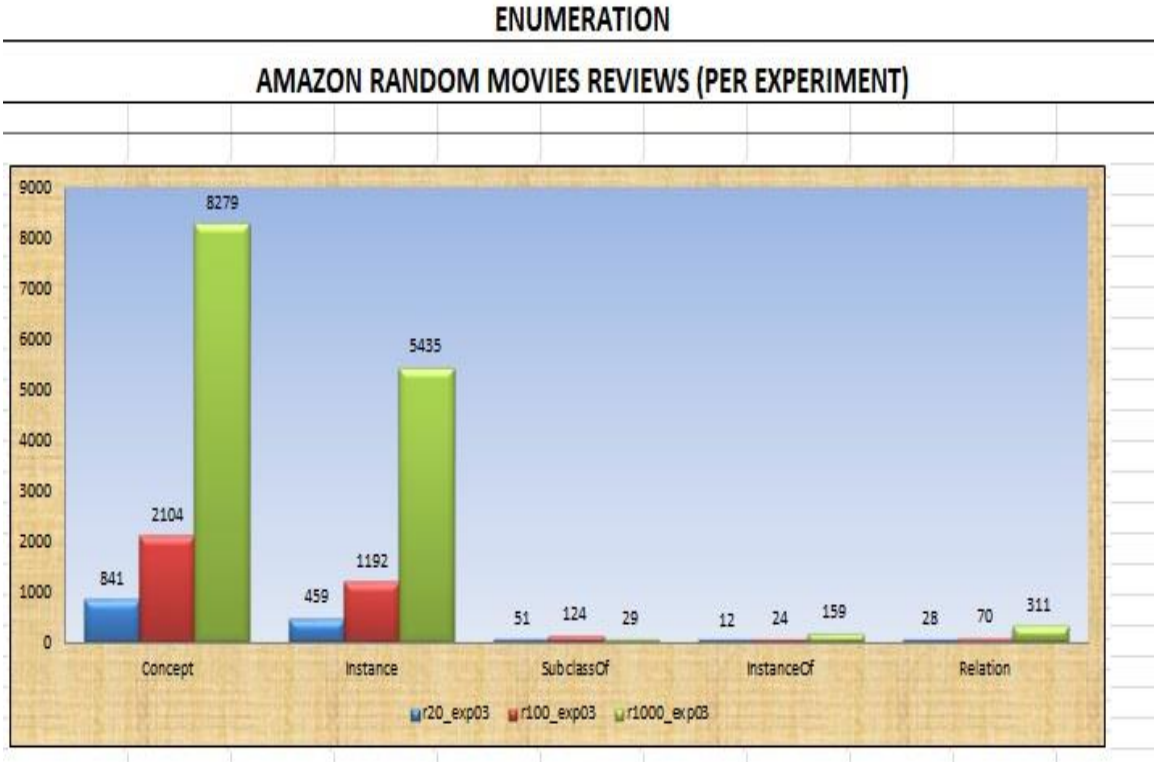
Ένα άλλο είδος γραφημάτων που προέκυψε, ως προς τη μέτρηση του αριθμού στοιχείων της οντολογίας είναι με βάση την ποσοτική μεταβολή των δεδομένων εισόδου ανά είδος πειράματος αντίστοιχα. Τα γραφήματα αυτά αφορούν τις κατηγορίες δεδομένων εισόδου *amazon_random* και *imdb_reviews*. Ενδεικτικά στην ενότητα αυτή αναλύονται τα γραφήματα της κατηγορίας *amazon_random* όπως και προηγουμένως. Υπάρχουν πέντε γραφήματα ανά είδος πειραμάτων αντίστοιχα. Τα γραφήματα αυτά φαίνονται παρακάτω στις εικόνες 8-7 έως 8-11.



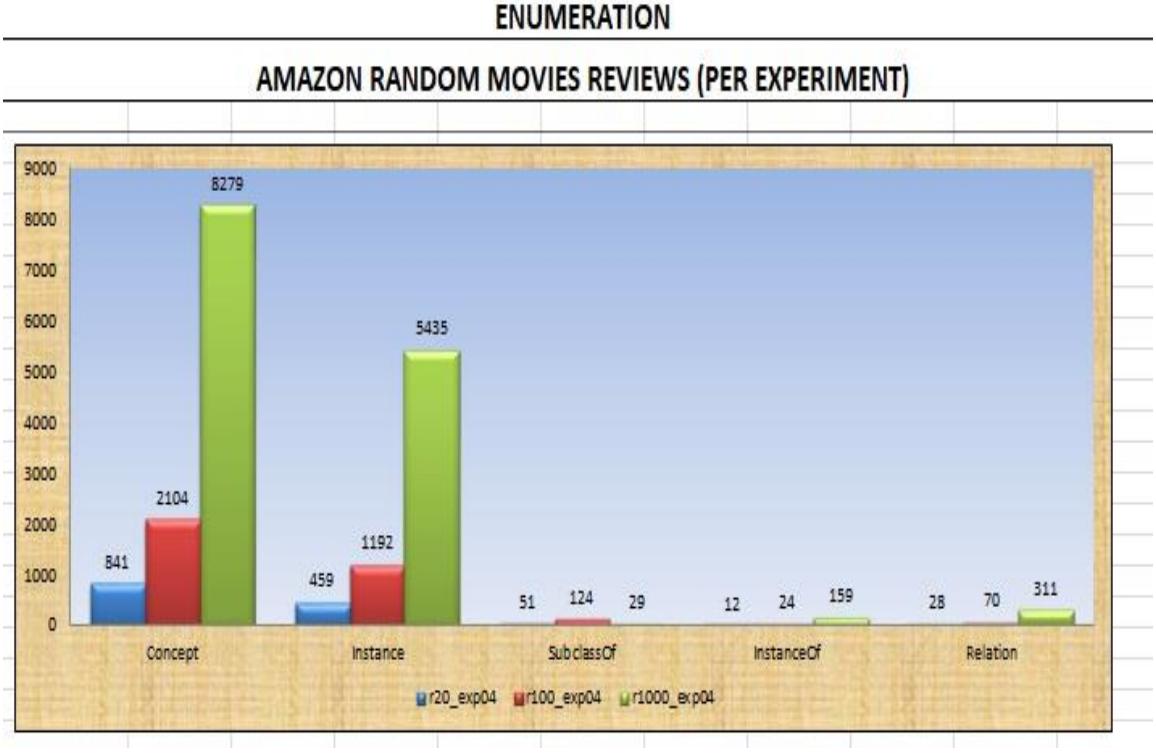
Εικόνα 8-7: Μέτρηση αριθμού στοιχείων οντολογίας – είδος πειράματος 01



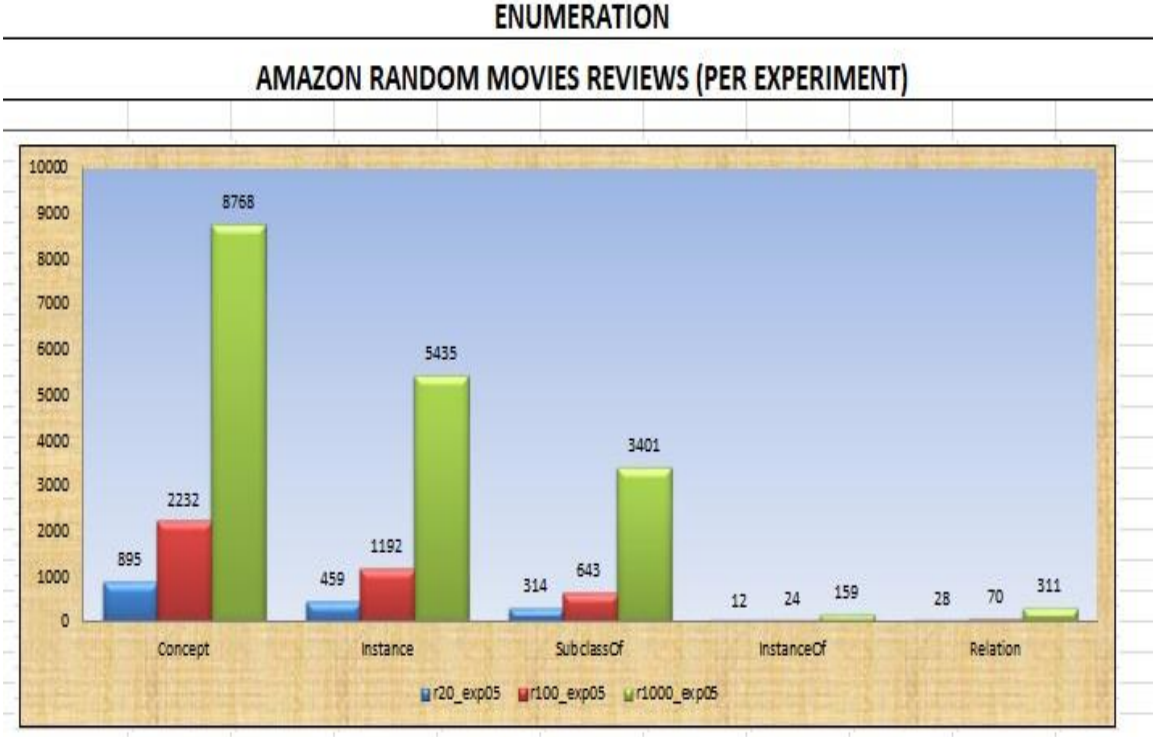
Εικόνα 8-8: Μέτρηση αριθμού στοιχείων οντολογίας – είδος πειράματος 02



Εικόνα 8-9: Μέτρηση αριθμού στοιχείων οντολογίας – είδος πειράματος 03



Εικόνα 8-10: Μέτρηση αριθμού στοιχείων οντολογίας – είδος πειράματος 04



Εικόνα 8-11: Μέτρηση αριθμού στοιχείων οντολογίας – είδος πειράματος 05

Η βασική παρατήρηση που μπορεί να γίνει βλέποντας τα παραπάνω γραφήματα είναι πως κατά βάση για όλα τα θεμελιακά στοιχεία και για τα πέντε είδη πειραμάτων τα γραφήματα έχουν την ίδια συμπεριφορά. Υπάρχει δηλαδή μια αύξηση του αριθμού των

στοιχείων της οντολογίας με την αύξηση του αριθμού των δεδομένων εισόδου. Μια μικρή εξαίρεση στο συγκεκριμένο φαινόμενο αποτελεί το θεμελιακό στοιχείο *SubclassOf* στο τρίτο και στο τέταρτο είδος πειραμάτων. Ακόμα και στο θεμελιακό στοιχείο *Concept* όπου θεωρητικά σε μία οντολογία οι κλάσεις της είναι σχετικά το πιο σταθερό μέγεθος, παρατηρούμε ότι υπάρχει μεγάλη αύξηση του αριθμού των στοιχείων καθώς αυξάνεται ο αριθμός των κριτικών που χρησιμοποιούμε ως είσοδο στα πειράματα.

Όπως αναφέραμε και παραπάνω η αύξηση αυτή σε συνδυασμό με τις μικρές τιμές πιθανότητας των οντολογικών στοιχείων στα θεμελιακά στοιχεία Concept και Instance δεν είναι ένα θετικό στοιχείο στο χτίσιμο μιας σωστής οντολογίας καθώς οδηγεί σε ένα κάπως χαοτικό τοπίο και σε πληθώρα μη αξιοποιήσιμων στοιχείων. Η βελτίωση των αποτελεσμάτων μπορεί να προέλθει από πειραματισμούς σχετικά με την ποσοτική μεταβολή των δεδομένων εισόδου σε συνδυασμό με επιλογές ενός κατωφλίου πιθανότητας για τη διαγραφή μεγάλου αριθμού πιθανώς ανεπιθύμητων στοιχείων.

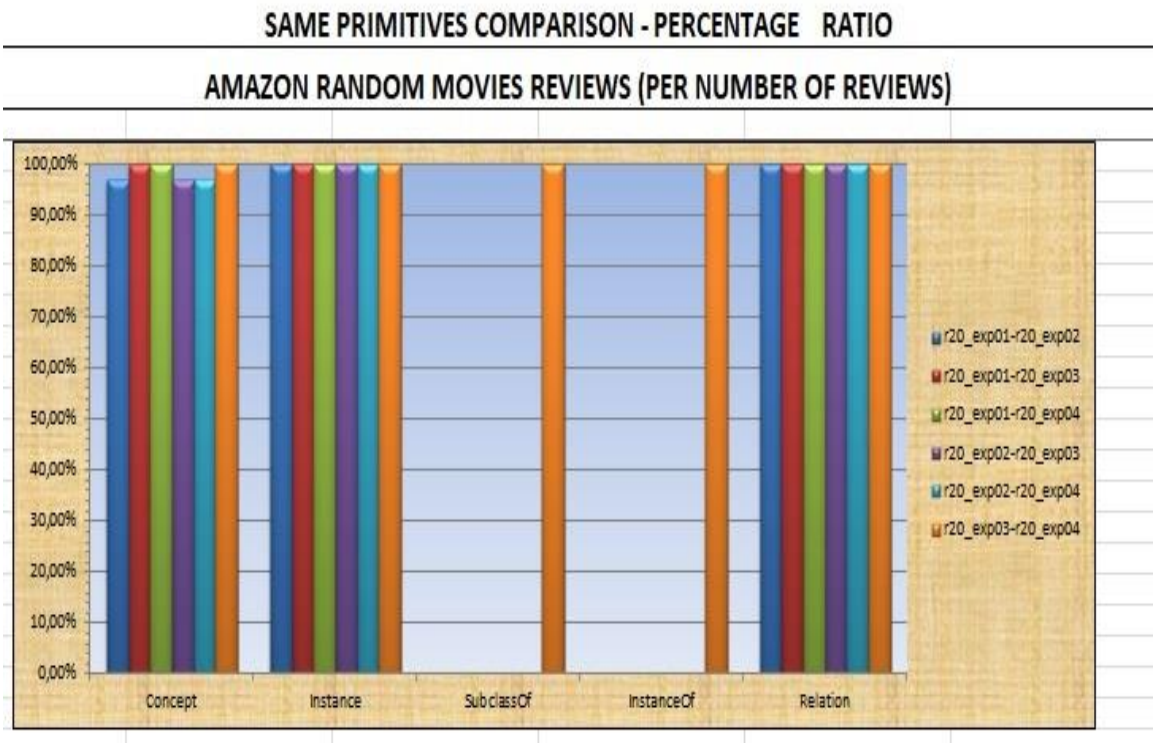
8.3 Σύγκριση ως προς τον αριθμό κοινών στοιχείων οντολογιών

Το επόμενο βήμα στην ανάλυση των αποτελεσμάτων αφορά τη σύγκριση των *owl* αρχείων που εξάγονται από δύο διαφορετικά πειράματα με σκοπό την εύρεση των κοινών στοιχείων της οντολογίας για κάθε θεμελιακό στοιχείο αντίστοιχα. Μέσω της συγκεκριμένης σύγκρισης επιδιώκουμε να δούμε το βαθμό ομοιότητας των οντολογιών που εξάγονται από δύο διαφορετικά πειράματα για το κάθε θεμελιακό στοιχείο μοντελοποίησης. Είναι σαφές πως όσο μεγαλύτερο είναι το ποσοστό των κοινών στοιχείων μεταξύ των δύο πειραμάτων τόσο πιο πετυχημένο είναι και το αποτέλεσμα των εξαγόμενων οντολογιών. Για την εύρεση του ποσοστού των κοινών στοιχείων υπολογίζεται ο συντελεστής ομοιότητας *Dice* που φαίνεται παρακάτω όπου ***Κοινά*** ο αριθμός των κοινών στοιχείων οντολογίας και ***Στοιχεία1***, ***Στοιχεία2*** ο αριθμός των στοιχείων του κάθε πειράματος αντίστοιχα:

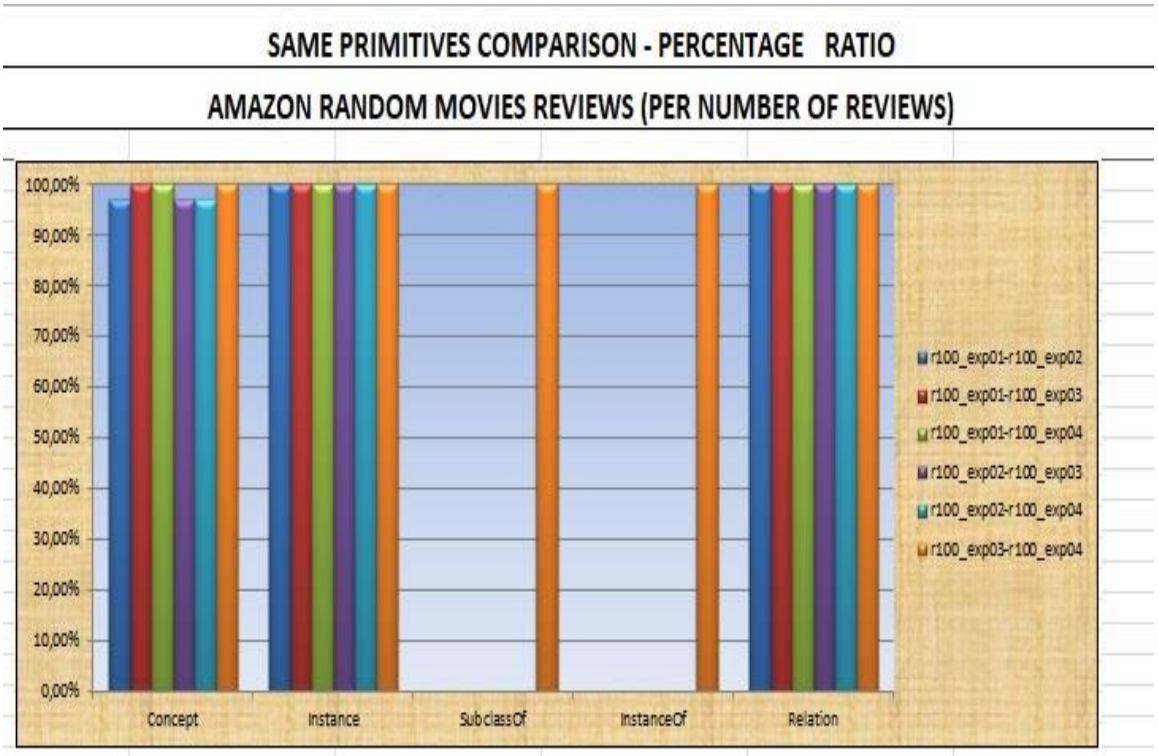
$$\frac{\mathbf{2} * \boldsymbol{Κοιν}\alpha}{\boldsymbol{Στοιχεί}\boldsymbol{\alpha}\mathbf{1} + \boldsymbol{Στοιχεί}\boldsymbol{\alpha}\mathbf{2}}$$

Η σύγκριση πραγματοποιήθηκε και για τις τέσσερις κατηγορίες δεδομένων εισόδου ξεχωριστά. Στην ενότητα αυτή, όπως και προηγουμένως, αναλύονται περαιτέρω τα γραφήματα που εξήχθησαν και αφορούν την κατηγορία δεδομένων εισόδου *amazon_random*. Αρχικά εξάγονται τρία γραφήματα και πάλι ανά αριθμό κριτικών

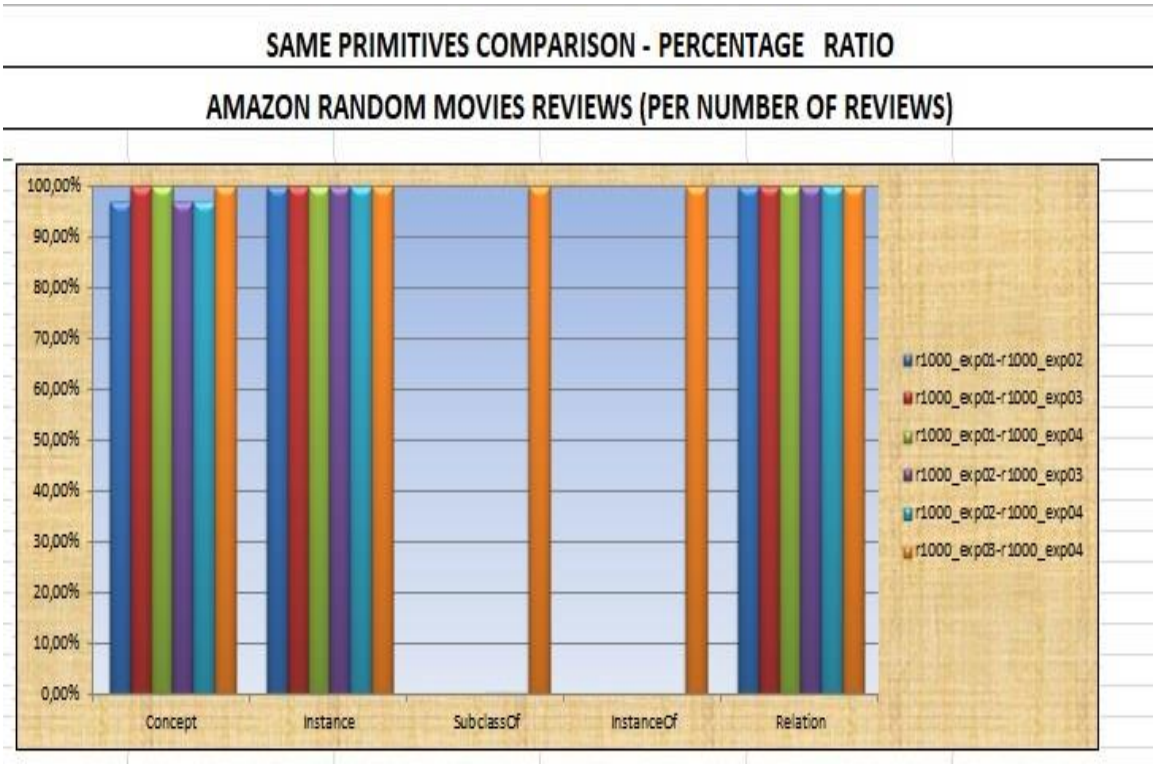
αντίστοιχα. Απεικονίζονται συγκρίσεις ανάμεσα στα τέσσερα πρώτα είδη πειραμάτων ανά δύο για όλα τα θεμελιακά στοιχεία. Τα γραφήματα αυτά φαίνονται παρακάτω στις εικόνες 8-12 έως 8-14.



Εικόνα 8-12: Σύγκριση κοινών στοιχείων οντολογιών – αριθμός κριτικών 20



Εικόνα 8-13: Σύγκριση κοινών στοιχείων οντολογιών – αριθμός κριτικών 100

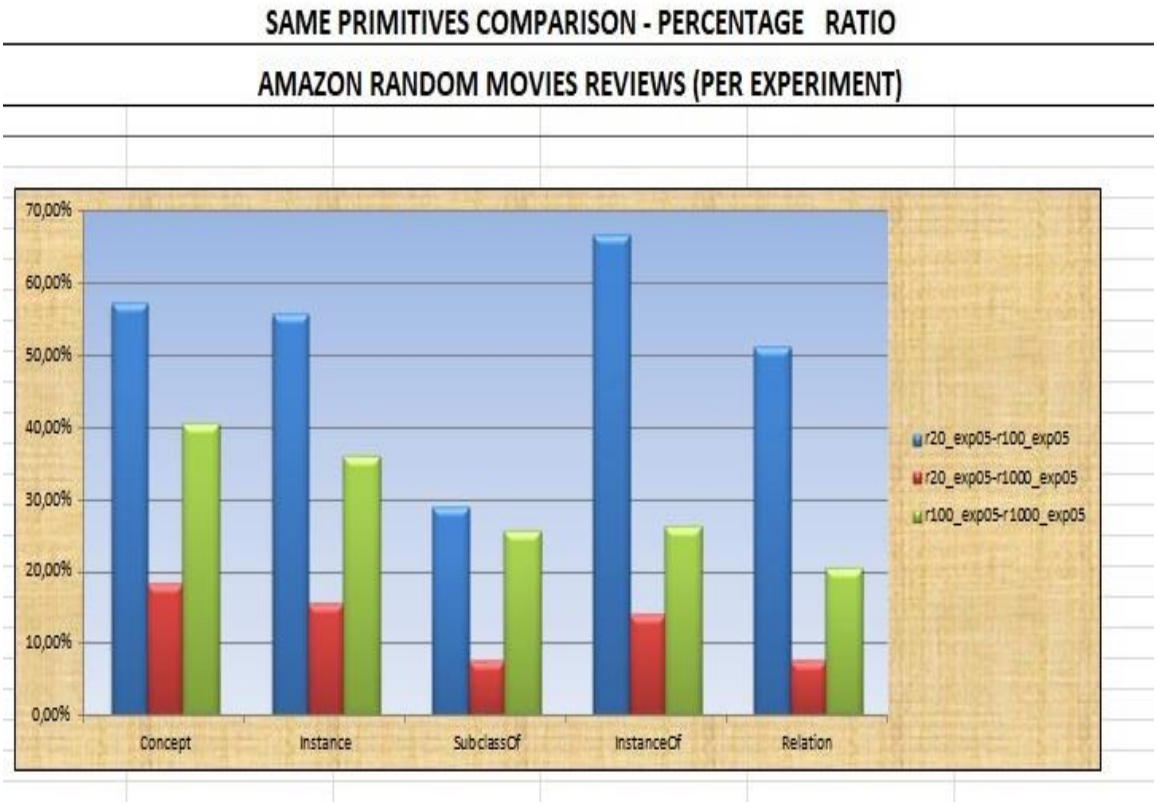


Εικόνα 8-14: Σύγκριση κοινών στοιχείων οντολογιών – αριθμός κριτικών 1000

Παρατηρούμε αρχικά πως και τα τρία γραφήματα παρουσιάζουν σε γενικές γραμμές την ίδια μορφή επομένως η εικόνα των συγκρίσεων δεν επηρεάζεται από την ποσοτική μεταβολή των δεδομένων εισόδου. Όσον αφορά τα θεμελιακά στοιχεία *Concept* και *Instance* βλέπουμε πως και στις έξι διαφορετικές συγκρίσεις που πραγματοποιήθηκαν τα κοινά στοιχεία σχεδόν αγγίζουν το 100%. Παρατηρούμε δηλαδή πως ανεξάρτητα από τον αλγόριθμο που χρησιμοποιείται σε καθένα από τα τέσσερα είδη πειραμάτων, βρίσκονται για όλες τις περιπτώσεις σχεδόν τα ίδια αποτελέσματα. Αντίθετα αυτό δεν ισχύει για τα θεμελιακά στοιχεία *Concept* και *Instance* όπου έχουμε κοινά στοιχεία στο 100% μόνο στις συγκρίσεις μεταξύ των πειραμάτων τρίτου και τέταρτου είδους όπου χρησιμοποιείται και για τα δύο θεμελιακά στοιχεία αντίστοιχα ο ίδιος αλγόριθμος. Για τις υπόλοιπες περιπτώσεις συγκρίσεων τα κοινά στοιχεία είναι σχεδόν μηδενικά. Υπενθυμίζουμε σε αυτό το σημείο ότι στα πρώτα δύο είδη πειραμάτων για το θεμελιακό στοιχείο *Instance* δεν είχαν βρεθεί καθόλου αποτελέσματα. Τέλος στην περίπτωση του θεμελιακού στοιχείου *Relation*, όπως αναμενόταν, τα κοινά στοιχεία σε όλες τις συγκρίσεις είναι στο 100% καθώς σε όλα τα είδη πειραμάτων χρησιμοποιείται ο ίδιος αλγόριθμος.

Ένα άλλο είδος γραφήματος που προέκυψε, ως προς την εύρεση των κοινών στοιχείων της οντολογίας είναι με βάση την ποσοτική μεταβολή των δεδομένων εισόδου

και αφορά το πέμπτο είδος πειράματος. Απεικονίζονται ανά δύο οι συγκρίσεις για το πέμπτο είδος πειράματος με 20, 100 και 1000 κριτικές αντίστοιχα για όλα τα θεμελιακά στοιχεία. Το γράφημα αυτό φαίνεται παρακάτω στην εικόνα 8-15.



Εικόνα 8-15: Σύγκριση κοινών στοιχείων οντολογιών – είδος πειράματος 05

Παρατηρώντας το παραπάνω γράφημα βλέπουμε πως η σύγκριση μεταξύ των πειραμάτων με 20 και 100 κριτικές έδωσε τα μεγαλύτερα ποσοστά κοινών στοιχείων για όλα τα θεμελιακά στοιχεία με ποσοστά που κυμαίνονται από λίγο πάνω του 50% έως και κοντά στο 70%. Η σύγκριση μεταξύ των πειραμάτων με 100 και 1000 κριτικές ακολούθησε με ποσοστά που κυμαίνονται από το 20% έως το 40% για όλα τα θεμελιακά στοιχεία αντίστοιχα. Τέλος τα μικρότερα ποσοστά κοινών στοιχείων έδωσε η σύγκριση μεταξύ των πειραμάτων με 20 και 1000 κριτικές με ποσοστά που κυμαίνονται από περίπου 7% έως λίγο κάτω του 20%.

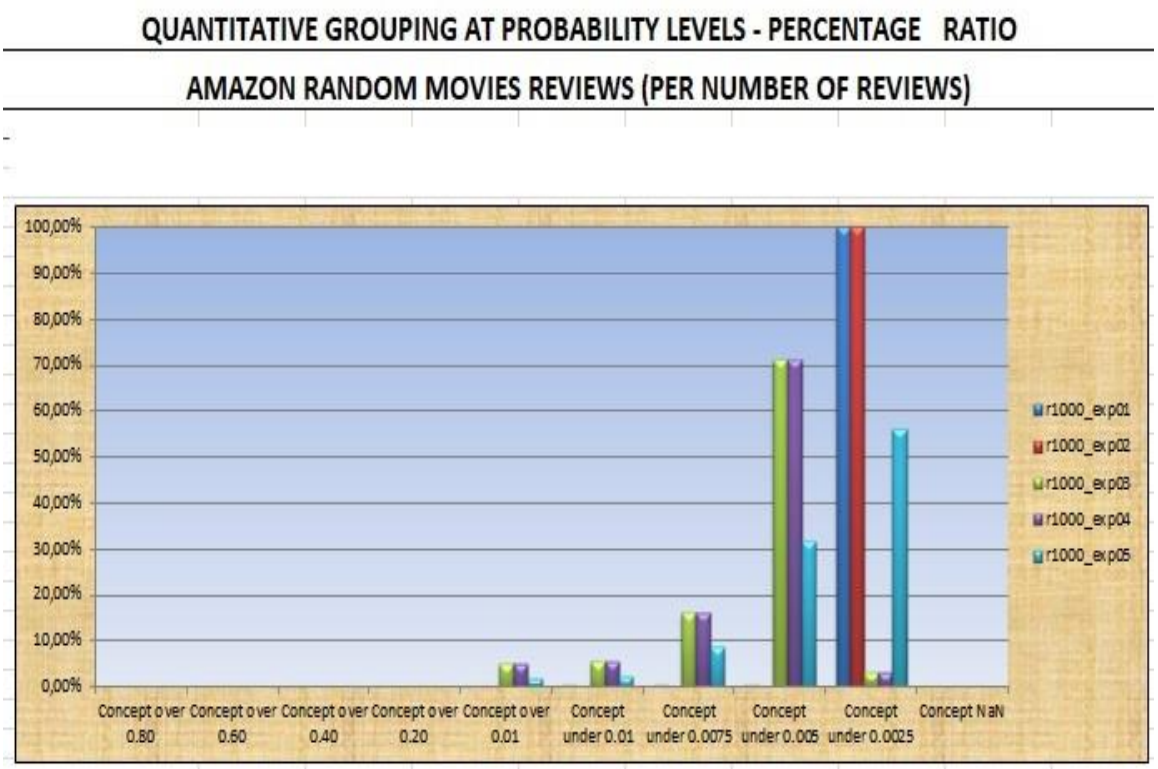
Το συγκεκριμένο γράφημα μπορεί να μας δείξει πως η σύγκριση για την εύρεση κοινών στοιχείων οντολογίας με ποσοτική μεταβολή των δεδομένων μπορεί να συμβάλει σε μεγάλο βαθμό στη βελτίωση των οντολογικών αποτελεσμάτων στην προσπάθεια για το χτίσιμο μιας σωστής οντολογίας. Είναι εύκολο να συμπεράνουμε πως τα κοινά οντολογικά στοιχεία μεταξύ των πειραμάτων με μεταβολή του αριθμού των κριτικών από 20 σε 100 και σε 1000 θα είναι πιθανόν και πιο αξιοποιήσιμα στο χτίσιμο της

οντολογίας. Ακόμα και στην περίπτωση της σύγκρισης μεταξύ των πειραμάτων με 20 και 1000 κριτικές που το ποσοστό των κοινών στοιχείων είναι μικρό αυτό μπορεί να αποδειχθεί ιδιαίτερα χρήσιμο αφού προέρχεται από σύγκριση δύο πειραμάτων με πολύ μεγάλη διαφορά μεταξύ του αριθμού των κριτικών κάτι που δείχνει πως αυτά τα κοινά στοιχεία που βρέθηκαν παραμένουν σταθερά είτε ο αριθμός των κριτικών είναι 20 είτε 1000. Περαιτέρω πειραματισμός συγκρίσεων ως προς την εύρεση των κοινών στοιχείων της οντολογίας πειραμάτων με διαφορετικό αριθμό δεδομένων εισόδου σε συνδυασμό με διαγραφή θεμελιακών στοιχείων κάτω από ένα κατώφλι πιθανότητας μπορεί να βελτιώσει περαιτέρω τα οντολογικά αποτελέσματα.

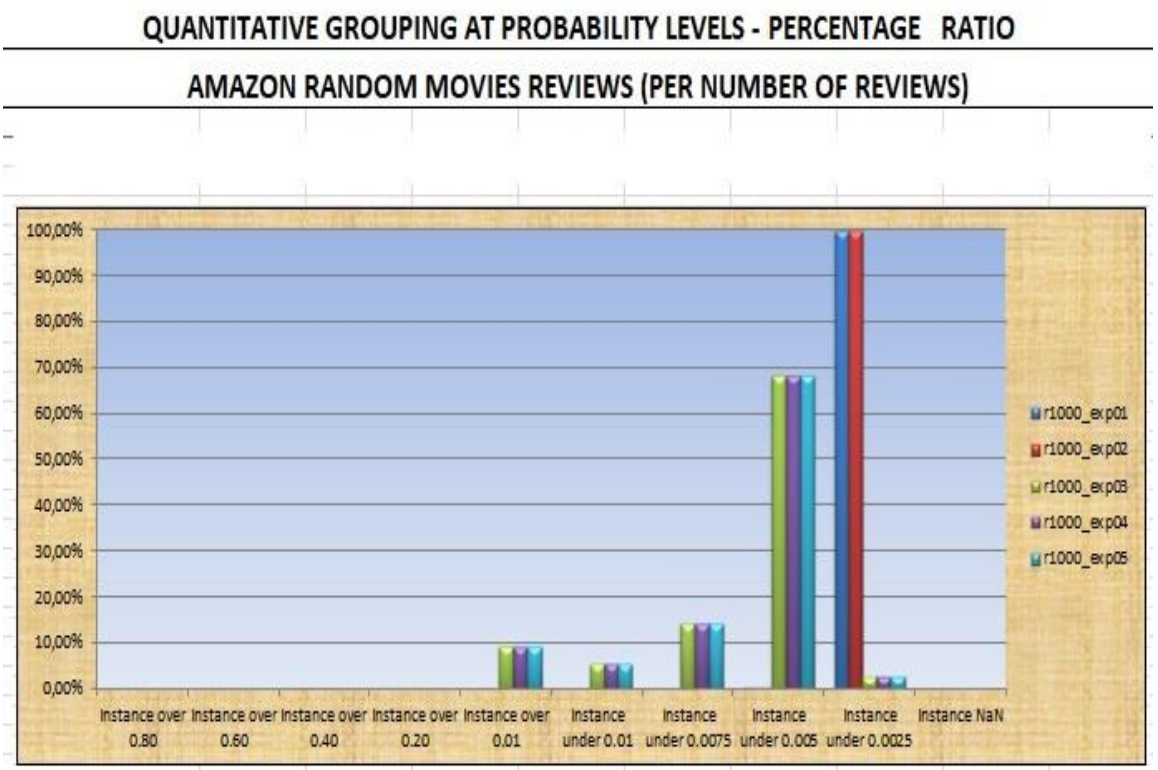
8.4 Ποσοτική ομαδοποίηση στοιχείων οντολογίας με βάση την τιμή πιθανότητας

Ως επόμενο στάδιο στην ανάλυση των αποτελεσμάτων των πειραμάτων κρίθηκε σκόπιμο να φανεί η σχέση που διέπει τα αριθμητικά αποτελέσματα των μετρήσεων των στοιχείων των οντολογιών και τιμών πιθανοτήτων τους. Κατόπιν αυτού πραγματοποιήθηκε μια ποσοτική ομαδοποίηση των στοιχείων της οντολογίας που πήραμε ως αποτέλεσμα σε κάθε πείραμα ανάλογα με την τιμή του πεδίου πιθανότητάς τους σύμφωνα με κάποια προκαθορισμένα επίπεδα πιθανότητας και για κάθε θεμελιακό στοιχείο αντίστοιχα. Τα αποτελέσματα της ομαδοποίησης εκτός από απόλυτο αριθμό αποτυπώθηκαν και σε μορφή ενός ποσοστιαίου λόγου για κάθε ομαδοποίηση στοιχείων.

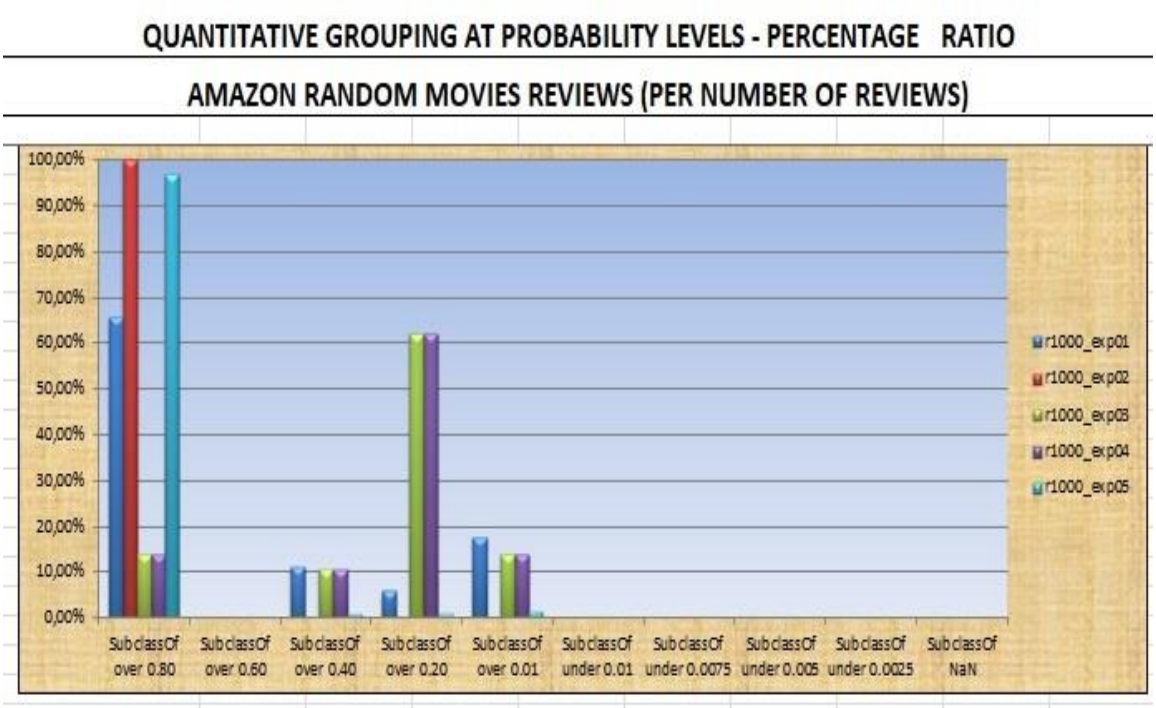
Η ομαδοποίηση πραγματοποιήθηκε και για τις τέσσερις κατηγορίες δεδομένων εισόδου ξεχωριστά και για τις αντίστοιχες ποσοτικές μεταβολές των δεδομένων. Για κάθε κατηγορία εξήχθησαν γραφήματα ανά θεμελιακό στοιχείο αντίστοιχα. Στην ενότητα αυτή εξετάζονται ενδεικτικά ορισμένα από τα γραφήματα που εξήχθησαν και αφορούν την κατηγορία δεδομένων εισόδου *amazon_random*. Τα πρώτα γραφήματα που εξετάζονται είναι για δεδομένα εισόδου της τάξης των 1000 κριτικών για το καθένα από τα πέντε είδη πειραμάτων. Τα γραφήματα αυτά φαίνονται παρακάτω στις εικόνες 8-16 έως 8-20.



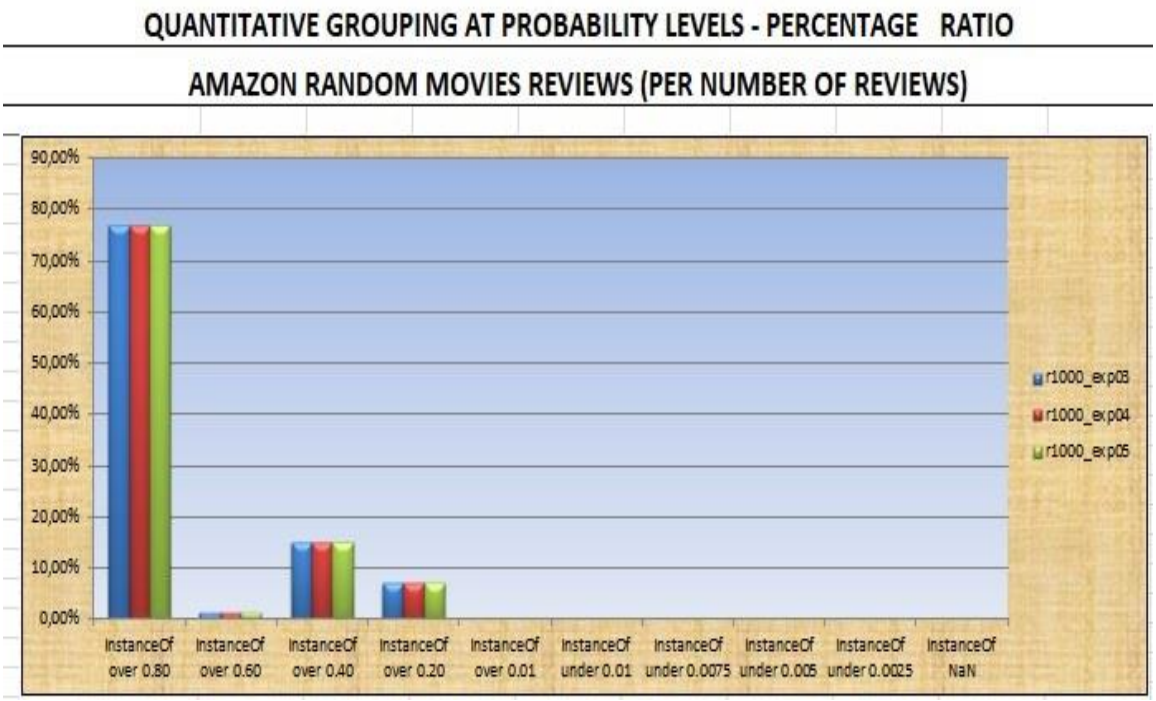
Εικόνα 8-16: Ποσοτική ομαδοποίηση στοιχείων οντολογιών με βάση την τιμή πιθανότητας – Θεμελιακό στοιχείο Concept – αριθμός κριτικών 1000



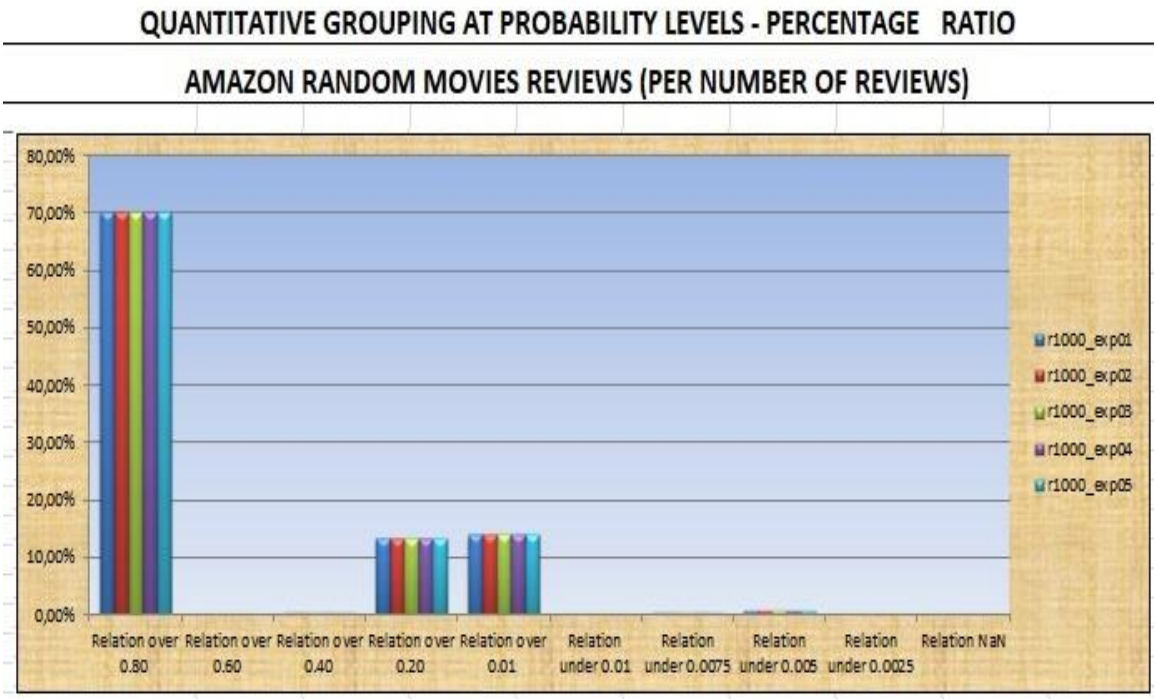
Εικόνα 8-17: Ποσοτική ομαδοποίηση στοιχείων οντολογιών με βάση την τιμή πιθανότητας – Θεμελιακό στοιχείο Instance – αριθμός κριτικών 1000



Εικόνα 8-18: Ποσοτική ομαδοποίηση στοιχείων οντολογιών με βάση την τιμή πιθανότητας – Θεμελιακό στοιχείο SubclassOf – αριθμός κριτικών 1000



Εικόνα 8-19: Ποσοτική ομαδοποίηση στοιχείων οντολογιών με βάση την τιμή πιθανότητας – Θεμελιακό στοιχείο InstanceOf – αριθμός κριτικών 1000



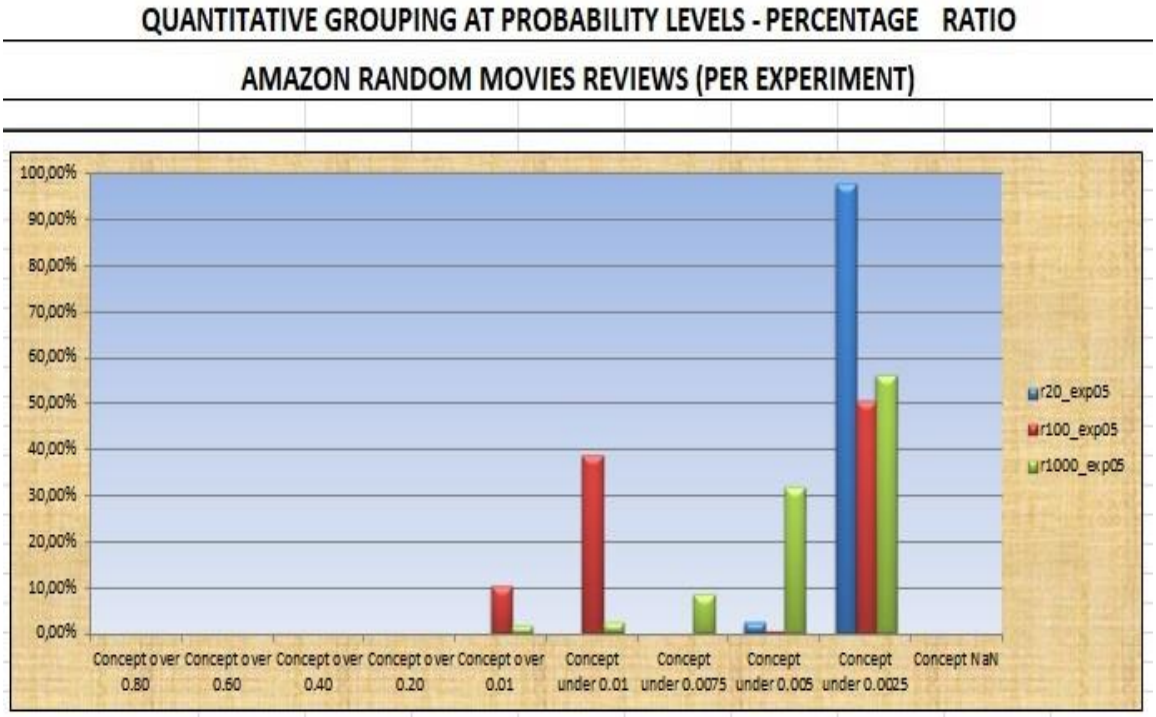
Εικόνα 8-20: Ποσοτική ομαδοποίηση στοιχείων οντολογιών με βάση την τιμή πιθανότητας – Θεμελιακό στοιχείο Relation – αριθμός κριτικών 1000

Η πρώτη βασική παρατήρηση που μπορεί να γίνει βλέποντας τα γραφήματα είναι πως, όσον αφορά τα θεμελιακά στοιχεία *Concept* και *Instance*, οι τιμές πιθανότητας των στοιχείων τους, όπως έχει αναφερθεί και νωρίτερα, κυμαίνονται γενικά σε χαμηλά επίπεδα. Παρατηρούμε επίσης και για τα δύο θεμελιακά στοιχεία πως για τα τρία τελευταία είδη πειραμάτων, ένα ποσοστό των αποτελεσμάτων τους εμφανίζεται και στις υψηλότερες στάθμες πιθανοτήτων σε σχέση με τα δύο πρώτα είδη όπου το σύνολο των αποτελεσμάτων έχει τιμή πιθανότητας κάτω από 0,0025. Ως προς το θεμελιακό στοιχείο *SubclassOf* αν κοιτάξουμε το αντίστοιχο γράφημα παρατηρούμε ότι το δεύτερο είδος πειραμάτων εμφανίζει τα καλύτερα αποτελέσματα καθώς το σύνολο των αποτελεσμάτων του έχει τιμή πιθανότητας άνω του 0,80. Αντίθετα τα χειρότερα πιθανοτικά αποτελέσματα εμφανίζουν το τρίτο και το τέταρτο είδος πειραμάτων. Για το θεμελιακό στοιχείο *InstanceOf* είχαμε, όπως έχει ειπωθεί, αποτελέσματα μόνο στα τελευταία είδη πειραμάτων με ένα ικανοποιητικό ποσοστό των αποτελεσμάτων άνω του 70% να έχει τιμή πιθανότητας άνω του 0,80. Το ίδιο ισχύει και για το θεμελιακό στοιχείο Relation όπου για όλα τα είδη πειραμάτων ένα ποσοστό των αποτελεσμάτων που αγγίζει το 70% έχει τιμή πιθανότητας άνω του 0,80.

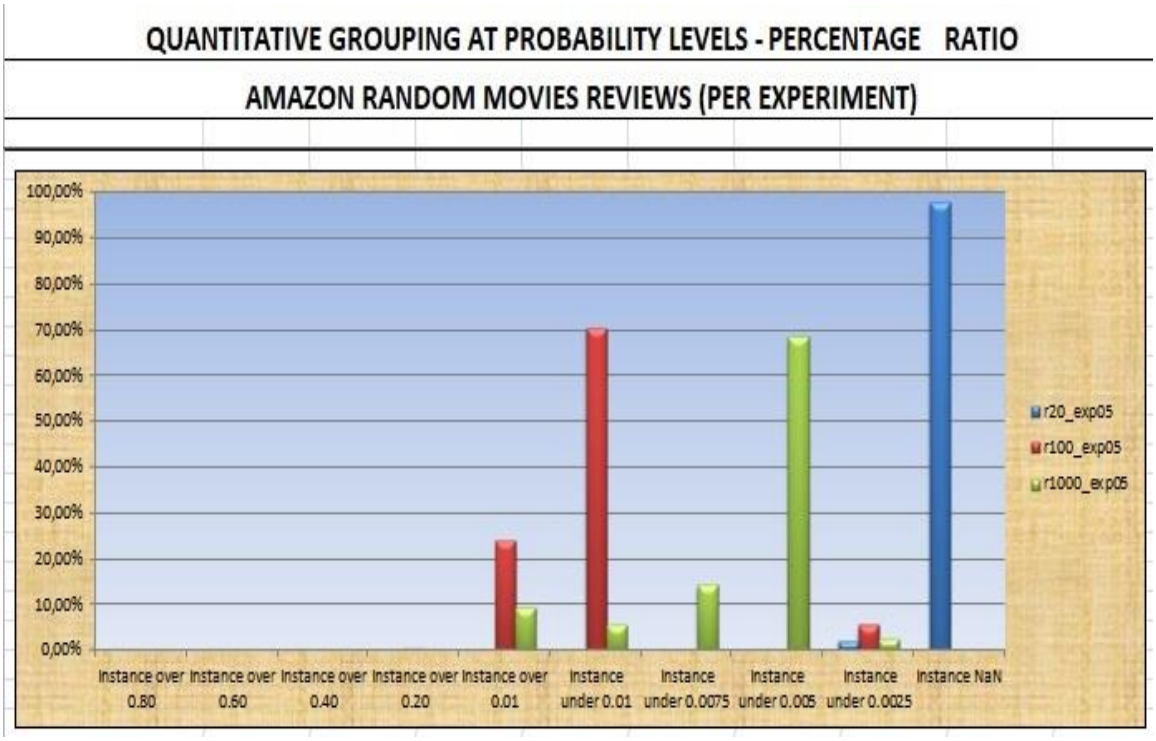
Οι παραπάνω παρατηρήσεις έρχονται να επιβεβαιώσουν τα συμπεράσματα που είχαν εξαχθεί στην ενότητα 8.2 σχετικά με τους αλγορίθμους που δίνουν τα βέλτιστα

αποτελέσματα για κάθε θεμελιακό στοιχείο. Όσον αφορά το θεμελιακό στοιχείο *Concept* ο αλγόριθμος *TFIDFConceptExtraction* που χρησιμοποιείται στο τέταρτο είδος πειραμάτων κρίνεται και πάλι ο πλέον κατάλληλος. Ως προς το θεμελιακό στοιχείο *Instance* ο αλγόριθμος *TFIDFInstanceExtraction* που χρησιμοποιείται στο τρίτο και στο τέταρτο είδος πειραμάτων βλέπουμε ότι και εδώ εμφανίζει καλύτερα πιθανοτικά αποτελέσματα. Όσον αφορά το θεμελιακό στοιχείο *SubclassOf* βλέπουμε πως ο αλγόριθμος *VerticalRelationsConceptClassification* που χρησιμοποιείται στο δεύτερο είδος πειραμάτων δίνει και πάλι τα καλύτερα αποτελέσματα. Αντίστοιχα όσον αφορά το θεμελιακό στοιχείο *InstanceOf* ο αλγόριθμος *PatternInstanceClassification* του τρίτου και τέταρτου είδους πειραμάτων είναι ο βέλτιστος καθώς τα πρώτα δύο είδη πειραμάτων δεν μας έδωσαν καθόλου αποτελέσματα.

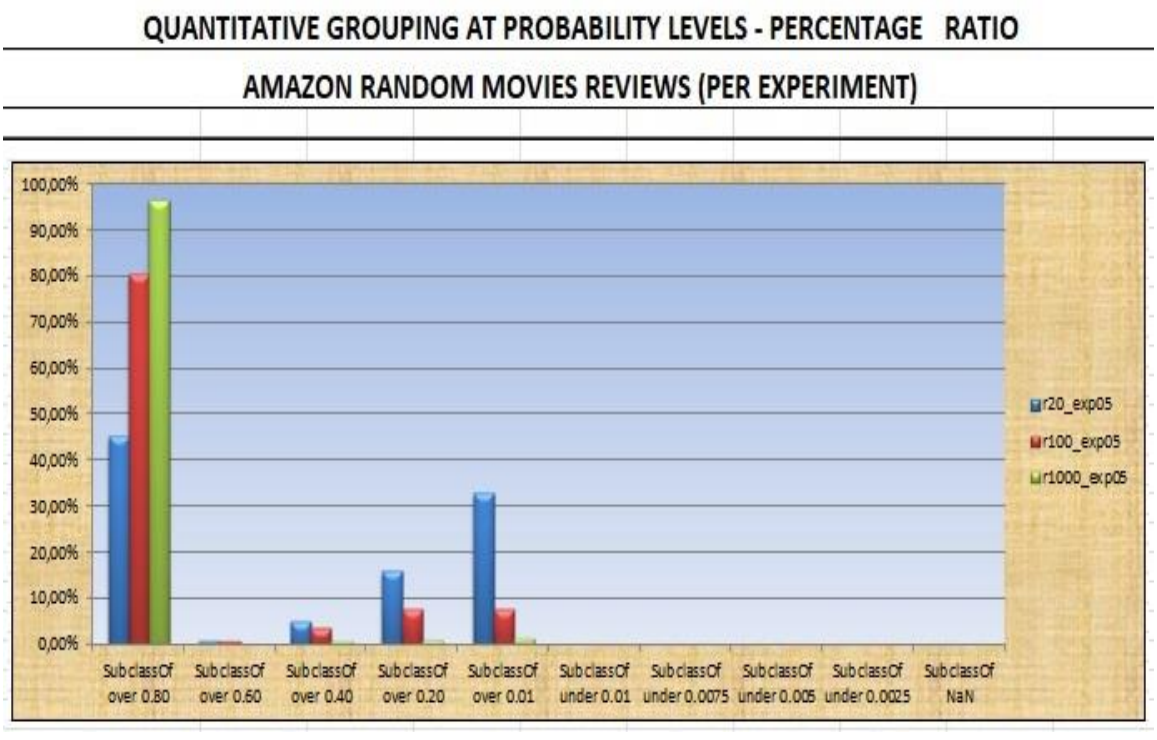
Ένα άλλο είδος γραφημάτων που κρίθηκε σκόπιμο να μελετηθεί στη συγκεκριμένη ανάλυση είναι αυτό που δείχνει για ένα συγκεκριμένο είδος πειραμάτων, στην προκειμένη περίπτωση το πέμπτο, πώς ομαδοποιούνται τα αποτελέσματα των πειραμάτων ανάλογα με την ποσοτική μεταβολή των δεδομένων εισόδου. Δημιουργήθηκαν και πάλι γραφήματα ανά θεμελιακό στοιχείο για τα δεδομένα εισόδου της τάξης των 20, 100 και 1000 κριτικών. Τα γραφήματα αυτά φαίνονται πιο κάτω στις εικόνες 8-21 έως 8-25.



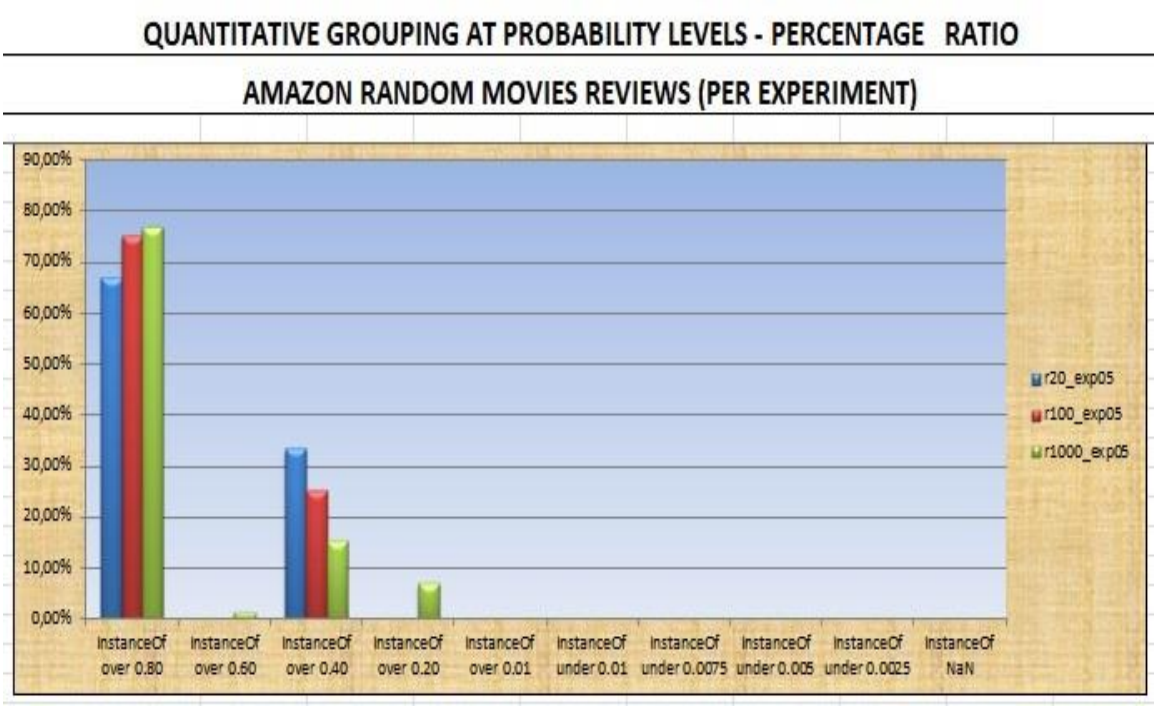
Εικόνα 8-21: Ποσοτική ομαδοποίηση στοιχείων οντολογιών με βάση την τιμή πιθανότητας – Θεμελιακό στοιχείο Concept – είδος πειράματος 05



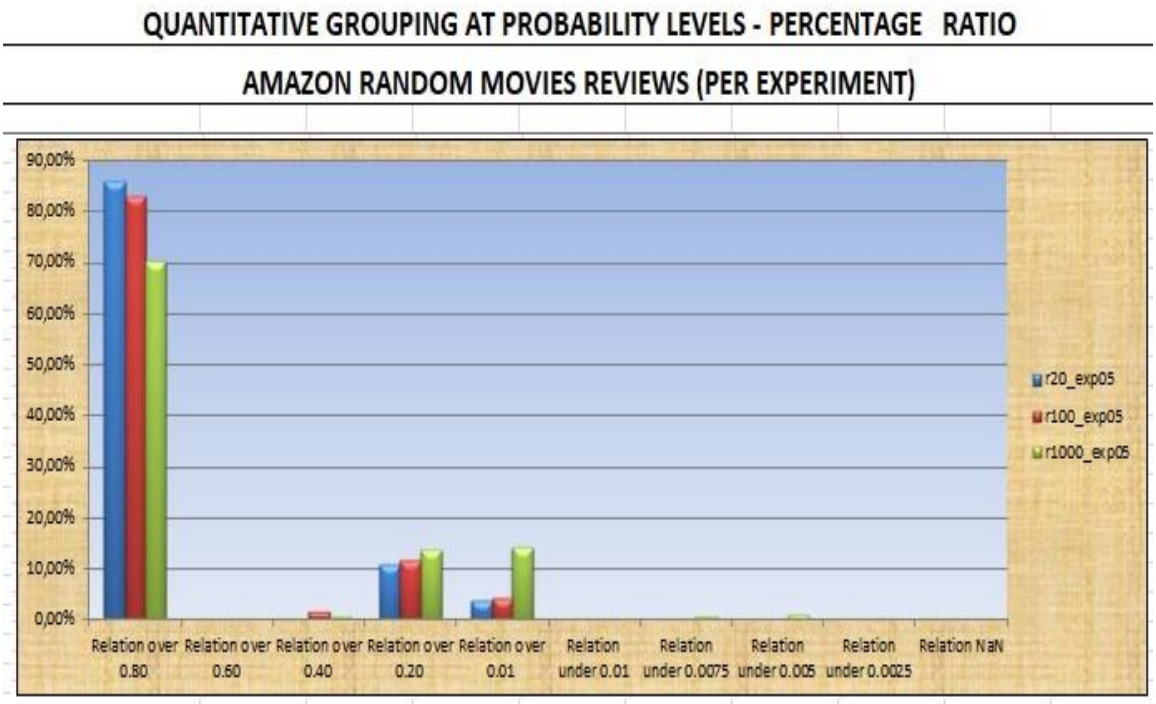
Εικόνα 8-22: Ποσοτική ομαδοποίηση στοιχείων οντολογιών με βάση την τιμή πιθανότητας – Θεμελιακό στοιχείο Instance – είδος πειράματος 05



Εικόνα 8-23: Ποσοτική ομαδοποίηση στοιχείων οντολογιών με βάση την τιμή πιθανότητας – Θεμελιακό στοιχείο SubclassOf – είδος πειράματος 05



Εικόνα 8-24: Ποσοτική ομαδοποίηση στοιχείων οντολογιών με βάση την τιμή πιθανότητας – Θεμελιακό στοιχείο InstanceOf – είδος πειράματος 05



Εικόνα 8-25: Ποσοτική ομαδοποίηση στοιχείων οντολογιών με βάση την τιμή πιθανότητας – Θεμελιακό στοιχείο Relation – είδος πειράματος 05

Παρατηρώντας τα δύο πρώτα γραφήματα για τα θεμελιακά στοιχεία *Concept* και *Instance* βλέπουμε πως τα καλύτερα αποτελέσματα από πλευράς πιθανοτήτων τα παίρνουμε ποσοστιαία στην περίπτωση των 100 κριτικών ως δεδομένων εισόδου. Στην

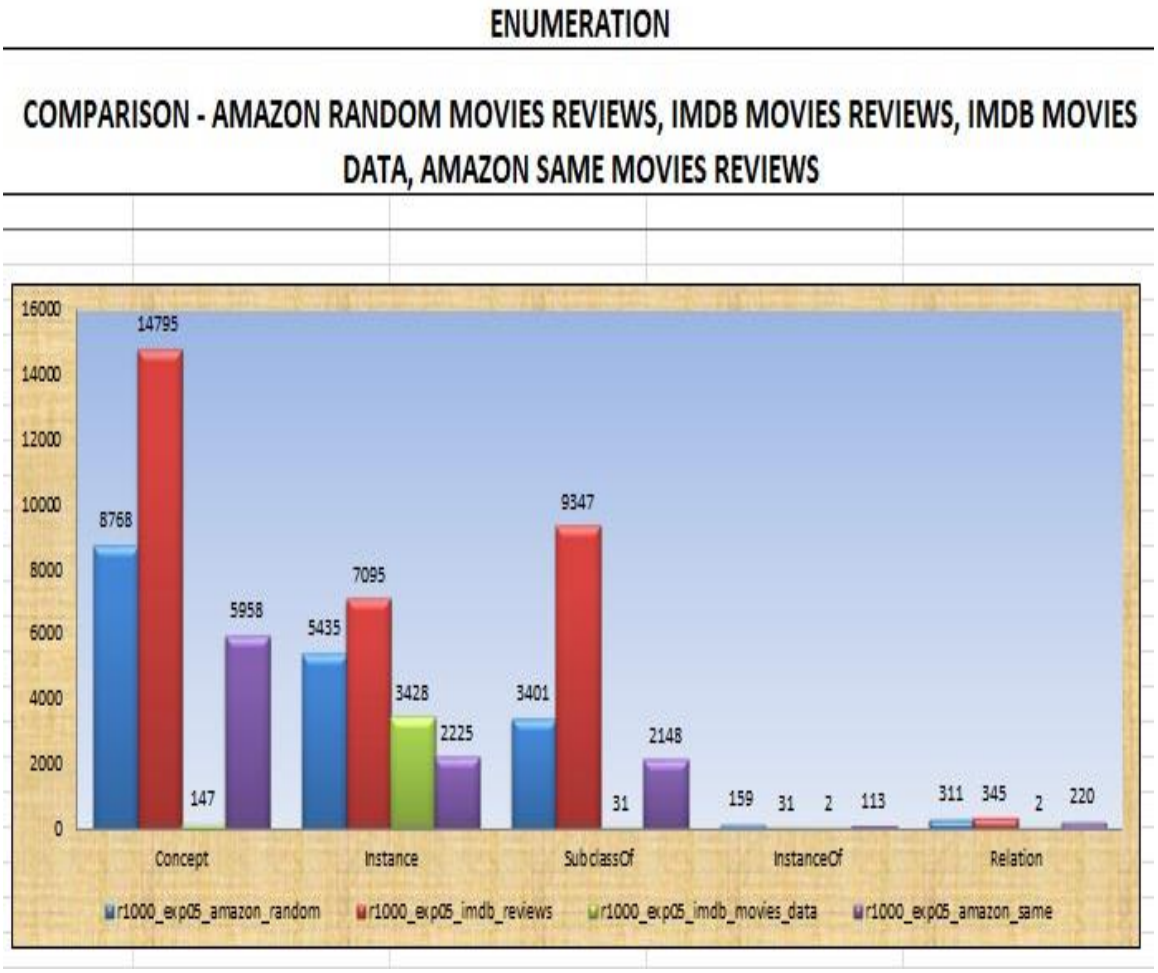
περίπτωση των 20 κριτικών βλέπουμε ότι όλα τα στοιχεία οντολογίας και στις δύο περιπτώσεις έχουν πολύ χαμηλή πιθανότητα κάτω του 0,0025. Ενώ τα πιθανοτικά αποτελέσματα παρουσιάζουν βελτίωση στην περίπτωση των 100 κριτικών, στην περίπτωση των 1000 κριτικών υπάρχει συγκέντρωση των αποτελεσμάτων σε χαμηλότερο επίπεδα στάθμης σε σχέση με τις 100 κριτικές. Ενώ δηλαδή αυξάνεται ο αριθμός των αποτελεσμάτων που παίρνουμε με είσοδο τις 1000 κριτικές, αυτά εμφανίζουν ποσοστιαία μικρότερες τιμές πιθανότητας από την περίπτωση του πειράματος με τις 100. Το γράφημα για το θεμελιακό στοιχείο *SubclassOf* μας δείχνει ότι η βελτίωση των πιθανοτικών αποτελεσμάτων συμβαδίζει αναλογικά με την ποσοτική αύξηση των δεδομένων εισόδου με τα στοιχεία στην περίπτωση των 1000 κριτικών, με τιμές πιθανότητας άνω του 0,80 να αγγίζουν το 100%. Σχετικά ανάλογη είναι η εικόνα του γραφήματος για το θεμελιακό στοιχείο *InstanceOf* παρουσιάζοντας όμως μεγαλύτερη συγκέντρωση και για τις τρεις περιπτώσεις, στις τιμές πιθανότητας άνω του 0,80. Αντίστοιχα υψηλή είναι η συγκέντρωση στις τιμές πιθανότητας άνω του 0,80 και στην περίπτωση του θεμελιακού στοιχείου *Relation*, με την εικόνα όμως σε αυτή την περίπτωση να είναι αντιστρόφως ανάλογη με την ποσοτική αύξηση των δεδομένων εισόδου και τα καλύτερα πιθανοτικά αποτελέσματα να τα παίρνουμε στην περίπτωση των 20 κριτικών και τα σχετικά χειρότερα στην περίπτωση των 1000.

8.5 Σύγκριση μετρήσεων αριθμού στοιχείων οντολογιών μεταξύ διαφορετικών κατηγοριών δεδομένων εισόδου

Μετά την ολοκλήρωση των αναλύσεων αποτελεσμάτων, με γραφήματα που αφορούν μια συγκεκριμένη κατηγορία δεδομένων εισόδου, το επόμενο στάδιο αφορά τις συγκρίσεις μεταξύ διαφορετικών κατηγοριών δεδομένων εισόδου. Οι αναλύσεις αυτού του είδους περιγράφονται σε αυτήν και τις επόμενες ενότητες.

Η πρώτη ανάλυση συγκρίσεων μεταξύ διαφορετικών κατηγοριών εισόδου αφορά τη μέτρηση του αριθμού των στοιχείων οντολογίας. Ενδεικτικά παρουσιάζεται στη συγκεκριμένη ενότητα το γράφημα που αφορά τη σύγκριση και για τις τέσσερις κατηγορίες εισόδου. Για την εξαγωγή του γραφήματος της συγκεκριμένης ανάλυσης επιλέχθηκαν δεδομένα εισόδου της τάξεως των 1000 για το πέμπτο είδος πειράματος. Γίνεται απεικόνιση της σύγκρισης του αριθμού των στοιχείων οντολογίας μεταξύ των τεσσάρων κατηγοριών εισόδου για κάθε θεμελιακό στοιχείο αντίστοιχα. Το γράφημα αυτό φαίνεται παρακάτω στην εικόνα 8-26. Αντίστοιχο γράφημα σύγκρισης μόνο των

κατηγοριών δεδομένων εισόδου *amazon_random* και *imdb_reviews* με μεταβολή της ποσότητας των δεδομένων εισόδου παρατίθεται στο παράρτημα.



Εικόνα 8-26: Σύγκριση μετρήσεων αριθμού στοιχείων οντολογιών μεταξύ όλων των κατηγοριών δεδομένων εισόδου – είδος πειράματος 05

Ξεκινώντας τις παρατηρήσεις από το θεμελιακό στοιχείο *Concept* βλέπουμε ότι το μεγαλύτερο αριθμό στοιχείων μας τον έδωσε η κατηγορία *imdb_reviews*. Αντίθετα η κατηγορία *imdb_movies_data* που ήταν και η μοναδική κατηγορία δεδομένων που δεν αφορούσε κριτικές ταινιών αλλά στοιχεία από ταινίες, μας έδωσε σε σχέση με τις υπόλοιπες μετρήσεις, ένα πολύ μικρό αριθμό στοιχείων. Οι κατηγορίες *amazon_random* και *amazon_same* με κριτικές από την *Amazon* μας έδωσαν ενδιάμεσες μετρήσεις σε σχέση με τις δύο προηγούμενες με την κατηγορία δεδομένων με κριτικές από τυχαίες ταινίες να έχει δώσει αρκετά μεγαλύτερο αριθμό στοιχείων. Συνεχίζοντας με το θεμελιακό στοιχείο *Instance* βλέπουμε ότι και πάλι η κατηγορία *imdb_reviews* έδωσε το μεγαλύτερο αριθμό στοιχείων, με την κατηγορία *amazon_random* και πάλι να ακολουθεί. Αντίθετα με πριν βλέπουμε ότι σε αυτό το θεμελιακό στοιχείο η κατηγορία

imdb_movies_data έδωσε έναν αρκετά μεγαλύτερο αριθμό στοιχείων πολύ πιο συγκρίσιμο με τις υπόλοιπες κατηγορίες. Το μικρότερο αριθμό στοιχείων τον έδωσε σε αυτό το θεμελιακό στοιχείο η κατηγορία *amazon_same*. Όσον αφορά το θεμελιακό στοιχείο *SubclassOf* αυτό παρουσιάζει μετρήσεις ακριβώς ανάλογες με αυτές του θεμελιακού στοιχείου *Concept* αλλά με σαφώς μειωμένους αριθμούς για όλες τις κατηγορίες. Η κατηγορία *imdb_reviews* παρουσιάζει και πάλι την υψηλότερη μέτρηση, η κατηγορία *imdb_movies_data* παρουσιάζει και πάλι έναν ελάχιστο αριθμό στοιχείων και οι υπόλοιπες δύο βρίσκονται κάπου ενδιάμεσα. Παρατηρώντας στη συνέχεια το θεμελιακό στοιχείο *InstanceOf* βλέπουμε ότι οι μετρήσεις δίνουν έναν πολύ μικρότερο αριθμό στοιχείων για όλες τις κατηγορίες σε σχέση με τα προηγούμενα θεμελιακά στοιχεία. Σε αυτή την περίπτωση τις καλύτερες μετρήσεις έδωσαν κατά σειρά οι κατηγορίες *amazon_random* και *amazon_same* με κριτικές από την *Amazon*, με την κατηγορία *imdb_reviews* να δίνει έναν πολύ μικρότερο αριθμό στοιχείων και την κατηγορία *imdb_movies_data*, σχεδόν μηδενικό. Αντίστοιχα μικρά είναι τα νούμερα και για το θεμελιακό στοιχείο *Relation*, με τις τρεις κατηγορίες κριτικών imdb_reviews, amazon_reviews, amazon_same να δίνουν κατά σειρά τις καλύτερες μετρήσεις, περίπου της ίδια τάξης και την κατηγορία *imdb_movies_data* να δίνει και πάλι σχεδόν μηδενικά αποτελέσματα.

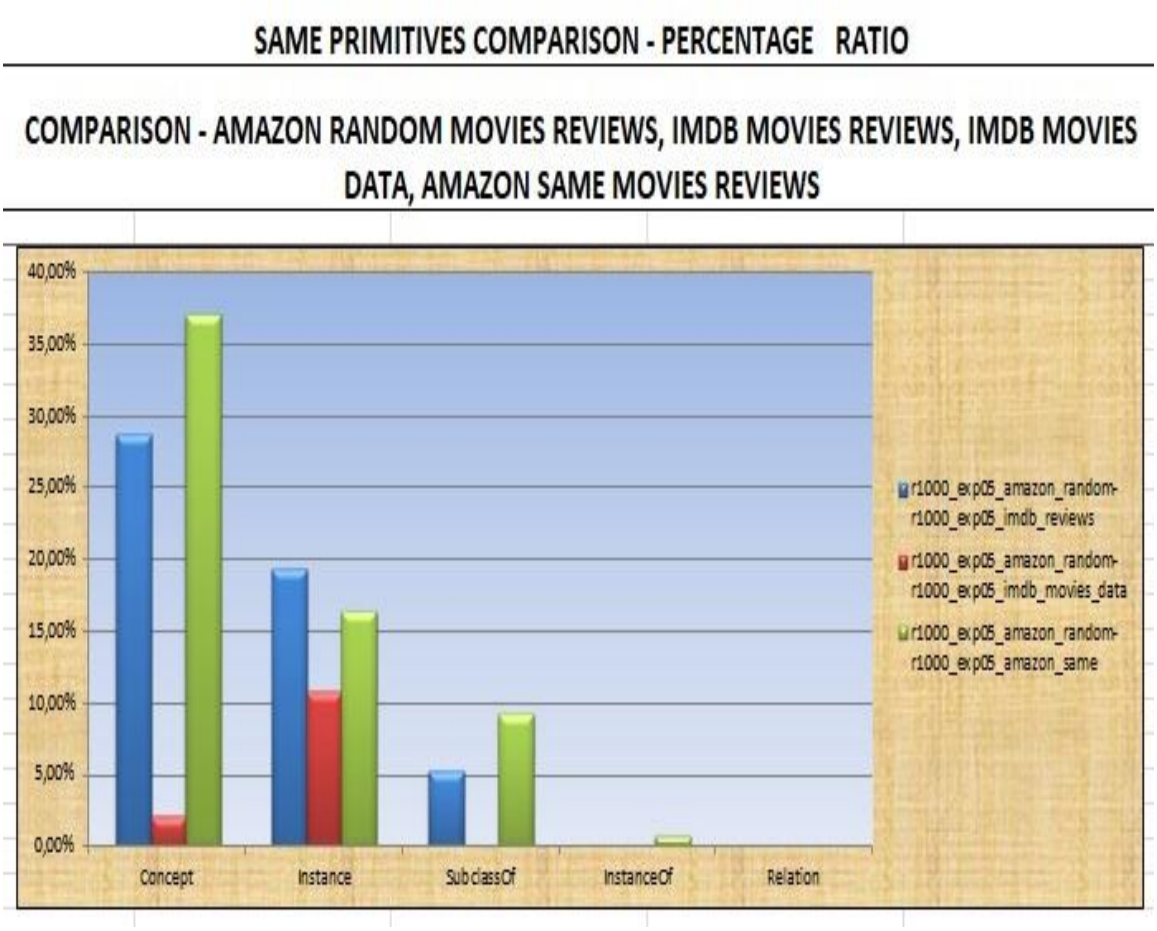
Ένα πρώτο βασικό συμπέρασμα που μπορεί να εξαχθεί από τις παρατηρήσεις σχετικά με το γράφημα είναι πως η κατηγορία δεδομένων εισόδου *imdb_movies_data* σε αντίθεση με τις κατηγορίες που αφορούν κριτικές δεν δείχνει ως μια πηγή δεδομένων εισόδου που μπορεί να αξιοποιηθεί επαρκώς στο χτίσιμο μιας σωστής οντολογίας. Στα θεμελιακά στοιχεία *SubclassOf, InstanceOf* και *Relation* ο αριθμός των οντολογικών στοιχείων είναι πάρα πολύ μικρός ενώ και στα θεμελιακά στοιχεία *Concept* και *Instance* που ο αριθμός των στοιχείων αυξάνεται σε σχέση με τα υπόλοιπα οι τιμές πιθανότητας αυτών των στοιχείων είναι πολύ μικρές κάτι που καθιστά και αυτά τα στοιχεία ελάχιστα αξιοποιήσιμα.

Όσον αφορά τις κατηγορίες που αφορούν κριτικές για τα θεμελιακά στοιχεία *Concept* και *Instance* βλέπουμε ότι όλες μας έδωσαν ένα μεγάλο πλήθος στοιχείων. Έχουμε δει βέβαια ήδη και θα το δούμε και σε επόμενη ενότητα για όλες τις κατηγορίες ότι ο μεγαλύτερος αριθμός αυτών των στοιχείων έχει πολύ μικρή τιμή πιθανότητας, οπότε ο σχετικός πειραματισμός με διαγραφή στοιχείων κάτω από ένα κατώφλι πιθανότητας χρειάζεται σε όλες τις περιπτώσεις ώστε να απαλλαγούμε από ένα μεγάλο

πλήθος μη αξιοποιήσιμων στοιχείων. Η διαδικασία των διαγραφών θα επηρεάσει και τα αντίστοιχα νούμερα του θεμελιακού στοιχείου *SubclassOf*. Συγκρίνοντας τις κατηγορίες που αφορούν κριτικές μεταξύ τους μπορούμε να συμπεράνουμε επίσης πως παρότι η κατηγορία *imdb_reviews* σε σχέση με τις κατηγορίες κριτικών της *Amazon* εμφανίζει για τα θεμελιακά στοιχεία *Concept*, *Instance* και *SubclassOf* μεγαλύτερο αριθμό στοιχείων αυτά, όπως θα δούμε και σε επόμενη ενότητα, έχουν στο σύνολο τους για τα δύο πρώτα θεμελιακά στοιχεία πάρα πολύ μικρή πιθανότητα. Αντίθετα ένα ποσοστό των στοιχείων που προκύπτουν από τις κατηγορίες κριτικών της *Amazon* εμφανίζουν μεγαλύτερες τιμές πιθανότητας κάτι που δείχνει ότι είναι πιθανόν η πιο αξιοποιήσιμη πηγή δεδομένων. Για το θεμελιακό στοιχείο *InstanceOf* και πάλι οι κατηγορίες κριτικών της *Amazon* μας δίνουν τον μεγαλύτερο αριθμό στοιχείων και με πολύ υψηλές πιθανότητας κάτι που επιβεβαιώνει το προηγούμενο συμπέρασμα σχετικά με την πηγή δεδομένων που φαίνεται να είναι η βέλτιστη. Τέλος σε ό, τι έχει να κάνει με το θεμελιακό στοιχείο *Relation* ο αριθμός των στοιχείων και από τις τρεις κατηγορίες που αφορούν κριτικές είναι του ίδιου μεγέθους και με αρκετά υψηλές τιμές πιθανότητας για το σύνολο των στοιχείων.

8.6 Σύγκριση ως προς τον αριθμό κοινών στοιχείων οντολογιών μεταξύ διαφορετικών κατηγοριών δεδομένων εισόδου

Μια δεύτερη κατεύθυνση ως προς την ανάλυση συγκρίσεων μεταξύ διαφορετικών κατηγοριών εισόδου αφορά τη σύγκριση ως προς τον αριθμό των κοινών στοιχείων οντολογιών μεταξύ διαφορετικών κατηγοριών εισόδου. Στην ενότητα αυτή ενδεικτικά αναλύεται το γράφημα που προκύπτει στην περίπτωση σύγκρισης της κατηγορίας δεδομένων εισόδου *amazon_random* με όλες τις υπόλοιπες κατηγορίες δεδομένων εισόδου. Τα δεδομένα εισόδου είναι της τάξεως των 1000, το είδος πειράματος που επιλέχθηκε είναι το πέμπτο και απεικονίζεται η σύγκριση αριθμού στοιχείων οντολογίας για κάθε θεμελιακό στοιχείο αντίστοιχα. Το γράφημα που προκύπτει φαίνεται παρακάτω στην εικόνα 8-27. Το αντίστοιχο γράφημα σύγκρισης μόνο των κατηγοριών δεδομένων εισόδου *amazon_random* και *imdb_reviews* με μεταβολή της ποσότητας των δεδομένων εισόδου παρατίθεται και αυτό στο παράρτημα.



Εικόνα 8-27: Σύγκριση ως προς τον αριθμό κοινών στοιχείων οντολογιών μεταξύ όλων των κατηγοριών δεδομένων εισόδου – είδος πειράματος 05

Παρατηρώντας αρχικά το τμήμα του γραφήματος για το θεμελιακό στοιχείο *Concept* βλέπουμε πως τα καλύτερα αποτελέσματα της σύγκρισης παρουσιάζονται για τη σύγκριση με την κατηγορία *amazon_same* που ήταν και σχετικά αναμενόμενο καθώς προέρχεται από την ίδια πηγή δεδομένων αλλά ο αριθμός των κοινών στοιχείων είναι σχετικά χαμηλός, κάτω του 40%. Ακολούθως η σύγκριση με την κατηγορία *imdb_reviews* μας δίνει αποτελέσματα κάτω του 30% και η σύγκριση με την κατηγορία *imdb_movies_data* μας έδωσε αποτελέσματα κοινών στοιχείων κοντά στο 2%. Η συμπεριφορά των αποτελεσμάτων είναι κάπως διαφορετική όσον αφορά το θεμελιακό στοιχείο *Instance*, με κανένα από τα αποτελέσματα των συγκρίσεων όμως να μην ξεπερνάει το 20%. Εδώ βλέπουμε πως τα καλύτερα αποτελέσματα μας έδωσε η σύγκριση με την κατηγορία *imdb_reviews* ενώ η σύγκριση με την κατηγορία *amazon_same* μας δίνει χειρότερα αποτελέσματα παρότι προέρχεται από την ίδια πηγή δεδομένων. Επιπρόσθετα βλέπουμε πως η σύγκριση με την κατηγορία *imdb_movies_data*, παρότι δίνει και πάλι τα χειρότερα αποτελέσματα, αυτά έχουν

αυξηθεί σε πάνω από 10%, σε αντίθεση με το θεμελιακό στοιχείο *Concept* που ήταν κοντά στο 2%. Στη συνέχεια για το θεμελιακό στοιχείο *SubclassOf* τα αποτελέσματα των συγκρίσεων είναι στο σύνολό τους σε ακόμα μικρότερα ποσοστά τα οποία κυμαίνονται κάτω του 10%. Η συμπεριφορά των αποτελεσμάτων είναι ανάλογη του θεμελιακού στοιχείου *Concept* καθώς η σύγκριση με την κατηγορία *amazon_same* δίνει τα καλύτερα αποτελέσματα που πλησιάζουν το 10%, η σύγκριση με την κατηγορία *imdb_reviews* μας δίνει αποτελέσματα λίγο πάνω από 5% και η σύγκριση με την κατηγορία *imdb_movies_data* μας έδωσε μηδενικά αποτελέσματα κοινών στοιχείων. Για τα υπόλοιπα δύο θεμελιακά στοιχεία *InstanceOf* και *Relation* όλες οι συγκρίσεις έδωσαν μηδενικά αποτελέσματα κοινών στοιχείων με μικρή εξαίρεση στο θεμελιακό στοιχείο *InstanceOf,* τη σύγκριση με την κατηγορία *amazon_same* που έδωσε αποτελέσματα κοντά στο 1%.

Ξεκινώντας από τη σύγκριση με την κατηγορία δεδομένων *imdb_movies_data*, τα πολύ χαμηλά έως μηδενικά κοινά στοιχεία της σύγκρισης για τα θεμελιακά στοιχεία οφείλονται καταρχάς σε ένα μεγάλο βαθμό στον πολύ μικρό αριθμό στοιχείων οντολογίας της συγκεκριμένης κατηγορίας σε σχέση με τις κριτικές, όπως είδαμε και στην προηγούμενη ενότητα. Μόνο όσον αφορά το θεμελιακό στοιχείο *Instance* που ο αριθμός των οντολογικών στοιχείων της κατηγορίας *imdb_movies_data* είναι αρκετά μεγαλύτερος σχετικά με τα υπόλοιπα θεμελιακά στοιχεία, ο αριθμός των κοινών στοιχείων ξεπέρασε το 10%. Για τις συγκρίσεις με τις κατηγορίες δεδομένων *imdb_reviews* και *amazon_same* και όσον αφορά τα θεμελιακά στοιχεία *Concept* και *SubclassOf* τα κοινά στοιχεία με την κατηγορία *amazon_same* είναι περισσότερα κάτι που είναι και λογικό καθώς προέρχεται από την ίδια πηγή δεδομένων. Για το θεμελιακό στοιχείο *Instance*, ο αριθμός των κοινών στοιχειών με την κατηγορία *imdb_review* είναι μεγαλύτερος κάτι που λογικά οφείλεται στον μεγάλο αριθμό οντολογικών στοιχείων που μας έδωσε αυτή η κατηγορία δεδομένων για το συγκεκριμένο οντολογικό στοιχείο. Για τα θεμελιακά στοιχεία *InstanceOf* και *Relation* βλέπουμε ότι τα αποτελέσματα είναι σχεδόν μηδενικά, πιθανόν λόγω του ότι τα συγκεκριμένα είναι τα πλέον δύσκολα ανιχνεύσιμα από τους αλγορίθμους θεμελιακά στοιχεία που πιθανώς να οδήγησε σε διαφορετικά αποτελέσματα για την κάθε κατηγορία δεδομένων. Παρατηρώντας πάντως ακόμα και στις περιπτώσεις που βρέθηκαν κοινά στοιχεία μεταξύ των διαφορετικών κατηγοριών δεδομένων αυτά κυμάνθηκαν σε σχετικά χαμηλά επίπεδα που στην καλύτερη των περιπτώσεων ξεπέρασαν το 35%. Οι χαμηλές τιμές πιθανότητας των

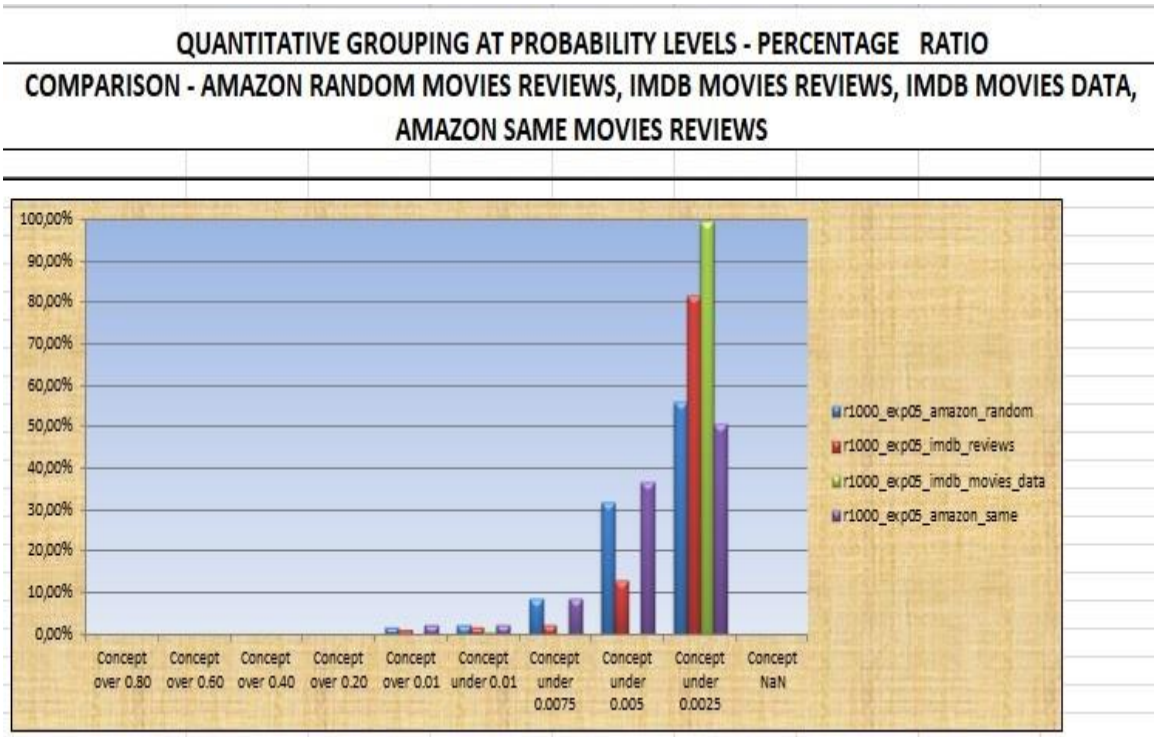
οντολογικών στοιχείων για τα δύο πρώτα θεμελιακά στοιχεία καθώς και η μεγαλύτερη πολυπλοκότητα των τριών επόμενων είναι το πιθανότερο αίτιο για τα συγκεκριμένα αποτελέσματα.

Αντίστοιχα με την σύγκριση ως προς τον αριθμό κοινών στοιχείων οντολογίας μιας κατηγορίας δεδομένων με ποσοτική μεταβολή των δεδομένων που περιγράφηκε σε προηγούμενη ενότητα και η εν λόγω σύγκριση μπορεί να αποβεί ιδιαίτερα χρήσιμη σε περαιτέρω διερεύνηση των οντολογικών αποτελεσμάτων με σκοπό τη βελτίωση της παραγόμενης οντολογίας. Τα κοινά στοιχεία που βρέθηκαν μεταξύ διαφορετικών κατηγοριών δεδομένων όπως είναι λογικό θα είναι και περισσότερο αξιοποιήσιμα στο σχηματισμό μιας οντολογίας. Πιθανοί πειραματισμοί σε συνδυασμό με διαγραφές οντολογικών στοιχείων κάτω από ένα κατώφλι πιθανότητας καθώς και σε συνδυασμό με την εύρεση των κοινών στοιχείων μιας κατηγορίας δεδομένων με ποσοτική μεταβολή των δεδομένων που περιγράφηκε σε προηγούμενη ενότητα θα μπορούσε να οδηγήσει σε περαιτέρω βελτίωση των οντολογικών αποτελεσμάτων.

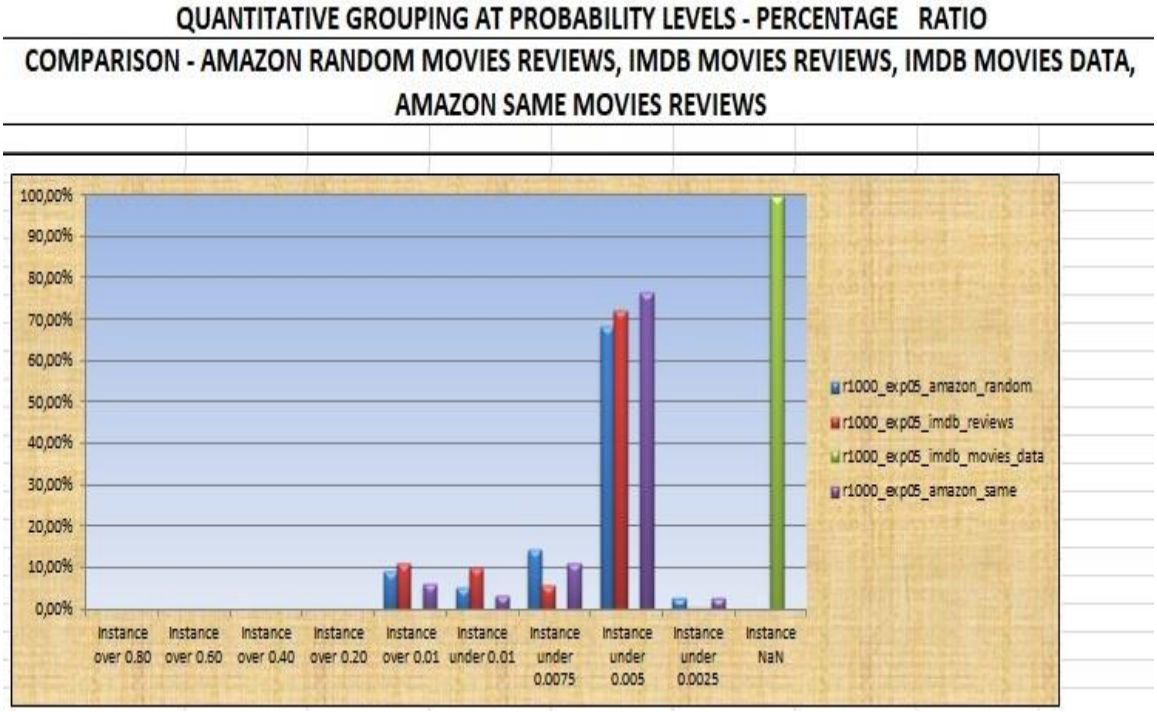
8.7 Σύγκριση ως προς την ποσοτική ομαδοποίηση στοιχείων οντολογίας με βάση την τιμή πιθανότητας μεταξύ διαφορετικών κατηγοριών δεδομένων εισόδου

Σε προηγούμενη ενότητα αναλύθηκε η ποσοτική ομαδοποίηση στοιχείων οντολογίας με βάση την τιμή της πιθανότητας σύμφωνα με κάποια προκαθορισμένα επίπεδα πιθανότητας για κάθε θεμελιακό στοιχείο. Η συγκεκριμένη ανάλυση κρίθηκε χρήσιμο να επεκταθεί και στην περίπτωση της σύγκρισης μεταξύ διαφορετικών κατηγοριών δεδομένων εισόδου.

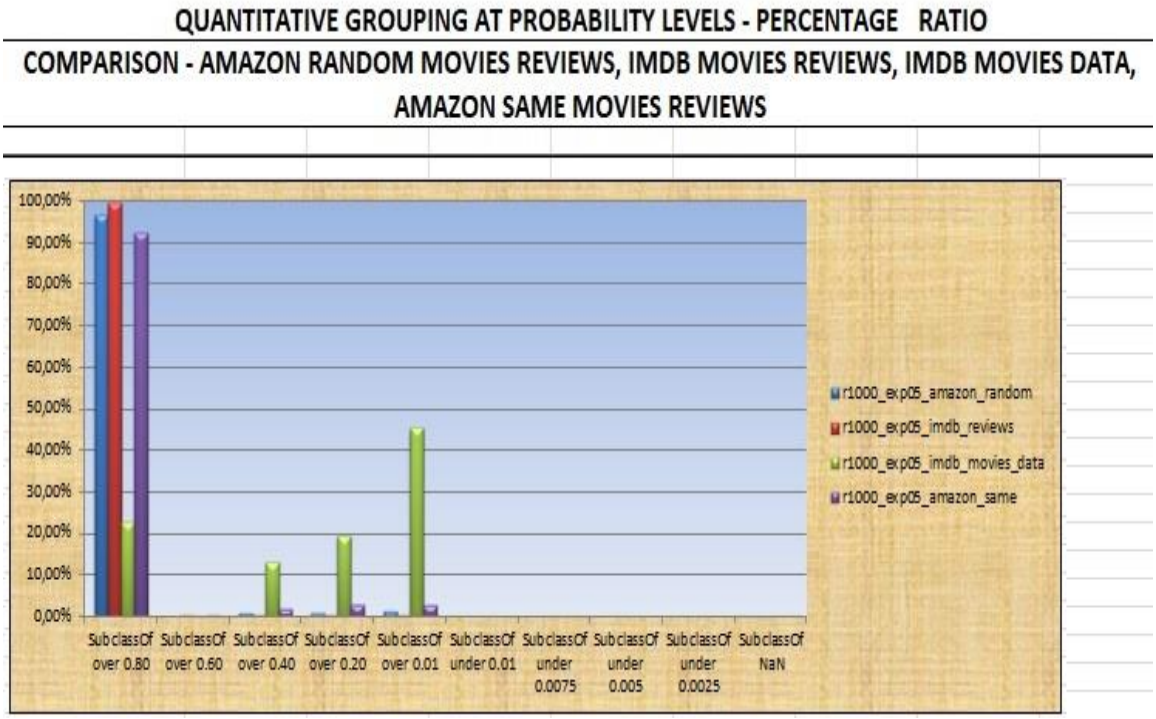
Στη συγκεκριμένη ενότητα αναλύονται γραφήματα ποσοτικής ομαδοποίησης ανά θεμελιακό στοιχείο που δείχνουν τη σύγκριση μεταξύ και των τεσσάρων κατηγοριών εισόδου. Τα δεδομένα εισόδου είναι της τάξεως των 1000 και το είδος πειράματος που επιλέχθηκε είναι το πέμπτο. Τα γραφήματα ανά θεμελιακό στοιχείο φαίνονται πιο κάτω στις εικόνες 8-28 έως 8-32. Αντίστοιχα γραφήματα σύγκρισης μόνο των κατηγοριών δεδομένων εισόδου *amazon_random* και *imdb_reviews* με μεταβολή της ποσότητας των δεδομένων εισόδου, παρατίθενται στο παράρτημα.



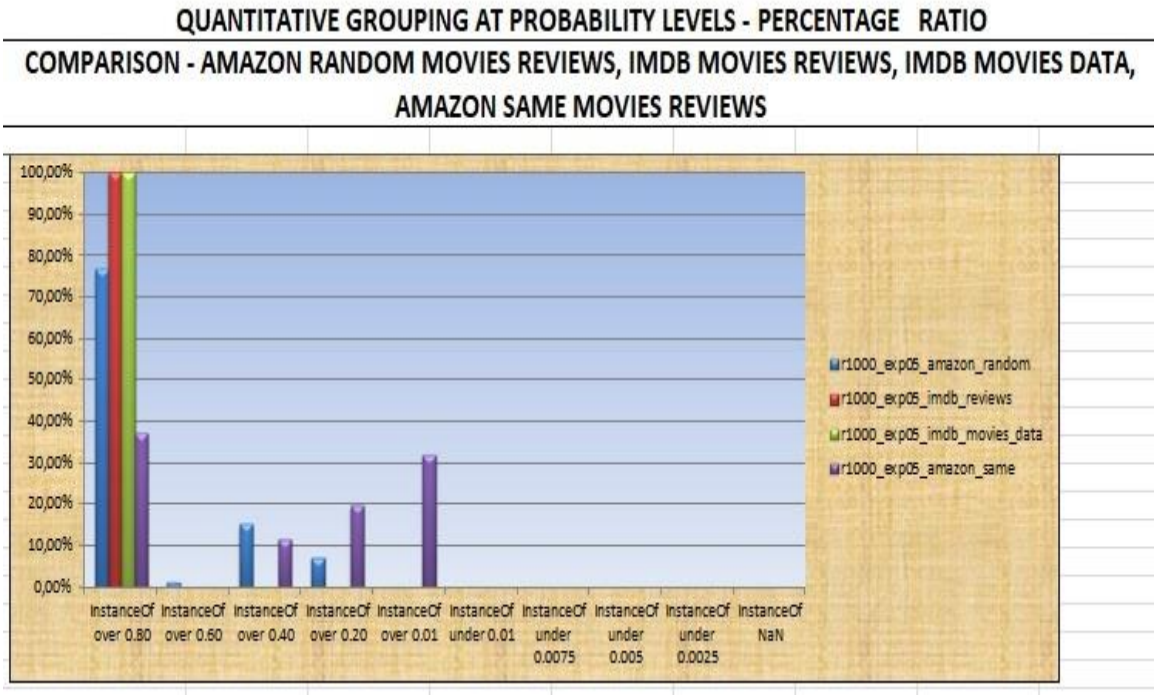
Εικόνα 8-28: Σύγκριση ως προς την ποσοτική ομαδοποίηση στοιχείων οντολογιών με βάση την τιμή πιθανότητας μεταξύ όλων των κατηγοριών δεδομένων εισόδου – Θεμελιακό στοιχείο Concept – είδος πειράματος 05



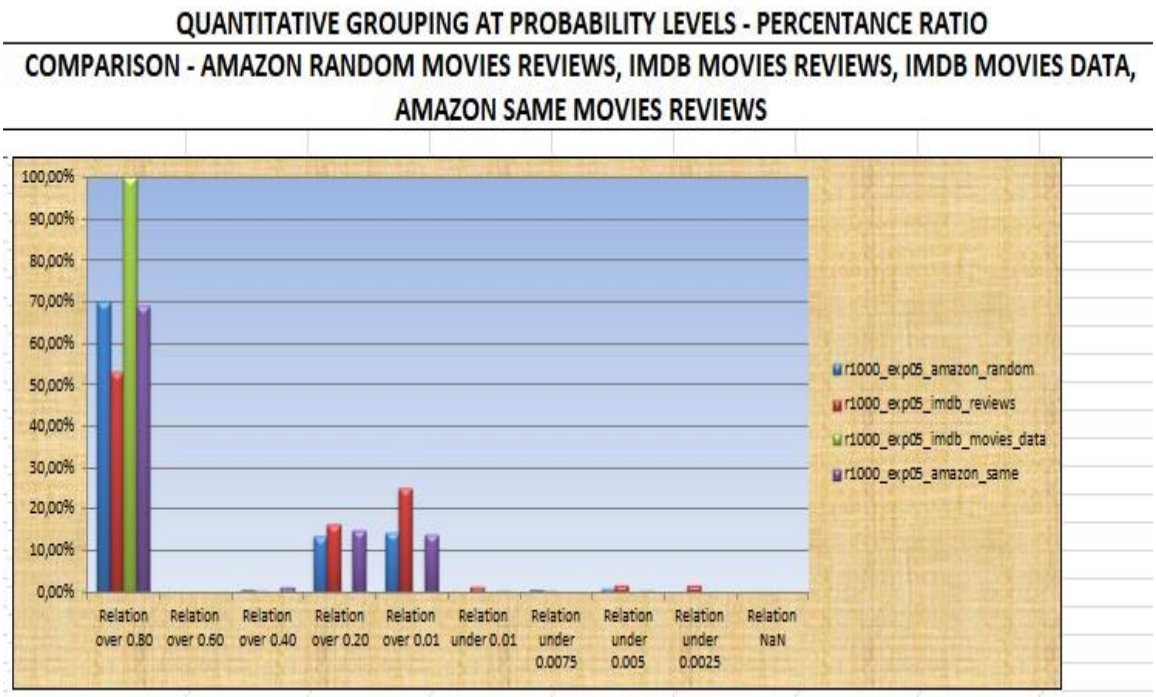
Εικόνα 8-29: Σύγκριση ως προς την ποσοτική ομαδοποίηση στοιχείων οντολογιών με βάση την τιμή πιθανότητας μεταξύ όλων των κατηγοριών δεδομένων εισόδου – Θεμελιακό στοιχείο Instance – είδος πειράματος 05



Εικόνα 8-30: Σύγκριση ως προς την ποσοτική ομαδοποίηση στοιχείων οντολογιών με βάση την τιμή πιθανότητας μεταξύ όλων των κατηγοριών δεδομένων εισόδου – Θεμελιακό στοιχείο SubclassOf – είδος πειράματος 05



Εικόνα 8-31: Σύγκριση ως προς την ποσοτική ομαδοποίηση στοιχείων οντολογιών με βάση την τιμή πιθανότητας μεταξύ όλων των κατηγοριών δεδομένων εισόδου – Θεμελιακό στοιχείο InstanceOf – είδος πειράματος 05



Εικόνα 8-32: Σύγκριση ως προς την ποσοτική ομαδοποίηση στοιχείων οντολογιών με βάση την τιμή πιθανότητας μεταξύ όλων των κατηγοριών δεδομένων εισόδου – Θεμελιακό στοιχείο Relation – είδος πειράματος 05

Ξεκινώντας τις παρατηρήσεις από το γράφημα του θεμελιακού στοιχείου *Concept*, βλέπουμε, όπως είχαμε παρατηρήσει και στην προηγούμενη σχετική ενότητα με τη συγκεκριμένη ανάλυση, ότι η μεγαλύτερη συγκέντρωση των στοιχείων για όλες τις κατηγορίες βρίσκεται στις χαμηλές στάθμες τιμής πιθανότητας. Αρχίζοντας τη σύγκριση μεταξύ των διαφόρων κατηγοριών δεδομένων εισόδου, για το συγκεκριμένο γράφημα βλέπουμε καταρχάς πως τα στοιχεία της κατηγορίας *imdb_movies_data* έχουν στο σύνολο τους πολύ χαμηλή τιμή πιθανότητας κάτω του 0,0025. Οι υπόλοιπες τρεις κατηγορίες παρατηρούμε πως εμφανίζουν και αυτές την υψηλότερη συγκέντρωση των στοιχείων τους σε πιθανότητες κάτω του 0,0025. Ένα ποσοστό όμως των στοιχείων για τις συγκεκριμένες κατηγορίες εμφανίζεται και σε υψηλότερες στάθμες τιμής πιθανότητας, με την κατηγορία *amazon_same* να εμφανίζει τα καλύτερα αποτελέσματα, την κατηγορία *amazon_random* με σχετικά μικρές διαφορές και τέλος την κατηγορία *imdb_reviews* να εμφανίζει σχετικά τα χειρότερα αποτελέσματα ως προς τις δύο προηγούμενες.

Κοιτώντας στη συνέχεια το γράφημα για το θεμελιακό στοιχείο *Instance* παρατηρούμε πως τα στοιχεία της κατηγορίας *imdb_movies_data* στο σύνολο τους εμφάνισαν τιμή πιθανότητας *NaN*. Ως προς τις άλλες τρεις κατηγορίες, όπως και πριν, η

μεγαλύτερη συγκέντρωση των στοιχείων τους εμφανίζεται σε αρκετά χαμηλές στάθμες τιμής πιθανότητας. Η διαφορά σε σχέση με πριν είναι ότι στη μεταξύ τους σύγκριση παρουσιάζουν ακριβώς αντίστροφη συμπεριφορά, με την κατηγορία *imdb_reviews* να εμφανίζει τα καλύτερα αποτελέσματα και τις κατηγορίες *amazon_random* και *amazon_same* να ακολουθούν κατά σειρά.

Συνεχίζοντας με το γράφημα για το θεμελιακό στοιχείο *SubclassOf* παρατηρούμε πως έχει αντιστραφεί η εικόνα σε σχέση με τα δύο προηγούμενα γραφήματα ως προς τη συγκέντρωση των στοιχείων οντολογίας όλων των κατηγοριών στις υψηλές στάθμες τιμής πιθανότητας. Τα αποτελέσματα για την κατηγορία *imdb_movies_data* παρουσιάζουν και πάλι τη χειρότερη συμπεριφορά σε σχέση με τις υπόλοιπες κατηγορίες με τη συγκέντρωση στη στάθμη πιθανότητας άνω του 0,80, να είναι λίγο πάνω από το 20% και η υψηλότερη συγκέντρωση αποτελεσμάτων να εμφανίζεται στη στάθμη πιθανότητας κάτω του 0,20. Αντιθέτως στις άλλες τρεις κατηγορίες η συγκέντρωση των αποτελεσμάτων στη στάθμη πιθανότητας άνω του 0,80 ξεπερνάει το 90%, με την κατηγορία *imdb_reviews* να φτάνει στο 100% των αποτελεσμάτων και τις κατηγορίες *amazon_random* και *amazon_same* να ακολουθούν κατά σειρά.

Στο επόμενο γράφημα που αφορά το θεμελιακό στοιχείο *InstanceOf* παρατηρούμε και πάλι υψηλές συγκεντρώσεις στις υψηλότερες στάθμες τιμής πιθανότητας ειδικότερα σε αυτήν άνω του 0,80. Τα αποτελέσματα της κατηγορίας *imdb_movies_data* βρίσκονται όπως βλέπουμε στο 100% άνω της στάθμης του 0,80, αλλά ουσιαστικά αυτό δεν είναι ενδεικτικό καθώς όπως είδαμε στην αντίστοιχη ενότητα των μετρήσεων είχαμε πάρει μόλις δύο αποτελέσματα για το συγκεκριμένο πείραμα. Όσον αφορά τις υπόλοιπες τρεις κατηγορίες, αυτές και πάλι εμφανίζουν τη μέγιστη συγκέντρωσή του στη στάθμη άνω του 0,80, με την κατηγορία *imdb_reviews* να φτάνει και πάλι στο 100% των αποτελεσμάτων και τις κατηγορίες *amazon_random* και *amazon_same* να υπολείπονται και πάλι, με τις συγκεντρώσεις τους όμως αυτή τη φορά, στη στάθμη άνω του 0,80 να είναι λίγο κάτω του 80% και του 40% αντίστοιχα.

Τέλος όσον αφορά το γράφημα του θεμελιακού στοιχείου *Relation* αυτό ακολουθεί τη συμπεριφορά των δύο προηγούμενων γραφημάτων με υψηλές συγκεντρώσεις όλων των κατηγοριών και πάλι στις υψηλές στάθμες τιμής πιθανότητας. Για τα αποτελέσματα της κατηγορίας *imdb_movies_data* ισχύει ό, τι ακριβώς και στην προηγούμενη περίπτωση με μόλις δύο αποτελέσματα με τιμές πιθανότητας άνω του 0,80. Ως προς τις άλλες τρεις κατηγορίες η συγκέντρωση των αποτελεσμάτων της καθεμίας

στην υψηλότερη στάθμη πιθανότητας αυτή τη φορά δεν ξεπερνάει το 70%, με την κατηγορία *imdb_reviews* να δίνει τα χειρότερα αποτελέσματα λίγο άνω του 50% και τις δύο κατηγορίες των κριτικών της *Amazon* να αγγίζουν το 70%, με την κατηγορία *amazon_random* ελαφρώς υψηλότερα.

Συνδυάζοντας τις παρατηρήσεις του γραφήματος της σύγκρισης μετρήσεων αριθμού στοιχείων οντολογιών, μεταξύ όλων των κατηγοριών δεδομένων εισόδου που εξετάστηκε σε προηγούμενη ενότητα, και των γραφημάτων αυτής της ενότητας, μπορούμε να καταλήξουμε σε κάποια γενικά συμπεράσματα που αφορούν τις κατηγορίες δεδομένων εισόδου και τις αντίστοιχες εξαγόμενες οντολογίες.

Ένα πρώτο βασικό συμπέρασμα είναι πως οι τρεις κατηγορίες δεδομένων εισόδου που αφορούν κριτικές δίνουν πολύ καλύτερα αποτελέσματα από την κατηγορία δεδομένων εισόδου που αφορά στοιχεία κινηματογραφικών ταινιών για όλα τα θεμελιακά στοιχεία των εξαγόμενων οντολογιών. Ένα δεύτερο συμπέρασμα που προκύπτει από τη σύγκριση των κατηγοριών δεδομένων εισόδου που αφορούν κριτικές από την *Amazon* είναι πως η επιλογή κριτικών από ίδιες κινηματογραφικές ταινίες δεν επέφερε βελτιωμένα αποτελέσματα σε σχέση με την επιλογή κριτικών από τυχαίες ταινίες. Ένα τρίτο συμπέρασμα που αφορά τη σύγκριση μεταξύ της κατηγορίας δεδομένων εισόδου με κριτικές από την *IMDb* σε σχέση με τις κατηγορίες κριτικών της *Amazon* είναι πως για τα θεμελιακά στοιχεία Instance και SubclassOf οι κριτικές της *IMDb* έδωσαν καλύτερα οντολογικά αποτελέσματα ενώ για τα θεμελιακά στοιχεία *Concept, InstanceOf* και *Relation* καλύτερα αποτελέσματα πήραμε από τις κριτικές της *Amazon*.

8.8 Σύγκριση ως προς τον αριθμό κοινών στοιχείων οντολογιών με πρότυπο οντολογίας κινηματογραφικών ταινιών

Όπως αναφέρθηκε και στο προηγούμενο κεφάλαιο ένας από τους αρχικούς στόχους της μελέτης ήταν η σύγκριση εξαγόμενων οντολογιών από τα πειράματα με ένα πρότυπο οντολογίας σχετικά με κινηματογραφικές ταινίες (Amancio Bouza (2010)) ως προς τον αριθμό των κοινών στοιχείων των οντολογιών για κάθε θεμελιακό στοιχείο αντίστοιχα.

Επιλέχθηκαν πειράματα και από τις τέσσερις κατηγορίες δεδομένων εισόδου. Συγκεκριμένα επιλέχθηκε για όλες τις κατηγορίες το πέμπτο είδος πειράματος με τα δεδομένα εισόδου να είναι της τάξεως των 1000. Τα αρχεία *owl* των παραπάνω

πειραμάτων μέσω της διαδικασίας μετατροπής θεμελιακών στοιχείων, που περιγράφηκε στο προηγούμενο κεφάλαιο, μετατράπηκαν σε κατάλληλη μορφή για τη σύγκριση με το αρχείο *owl* του πρότυπου οντολογίας προς εύρεση κοινών στοιχείων. Η σύγκριση έγινε μέσω της διαδικασίας σύγκρισης με πρότυπο οντολογίας που και αυτή περιγράφηκε στο προηγούμενο κεφάλαιο.

Η σύγκριση αφορά τα τρία βασικά από τα πέντε θεμελιακά στοιχεία: *Concept*, *Instance* και *Relation*. Αρχικά βρέθηκε ο αριθμός των στοιχείων οντολογίας των παραπάνω θεμελιακών στοιχείων για το πρότυπο οντολογίας που φαίνεται παρακάτω στην εικόνα 8-33.

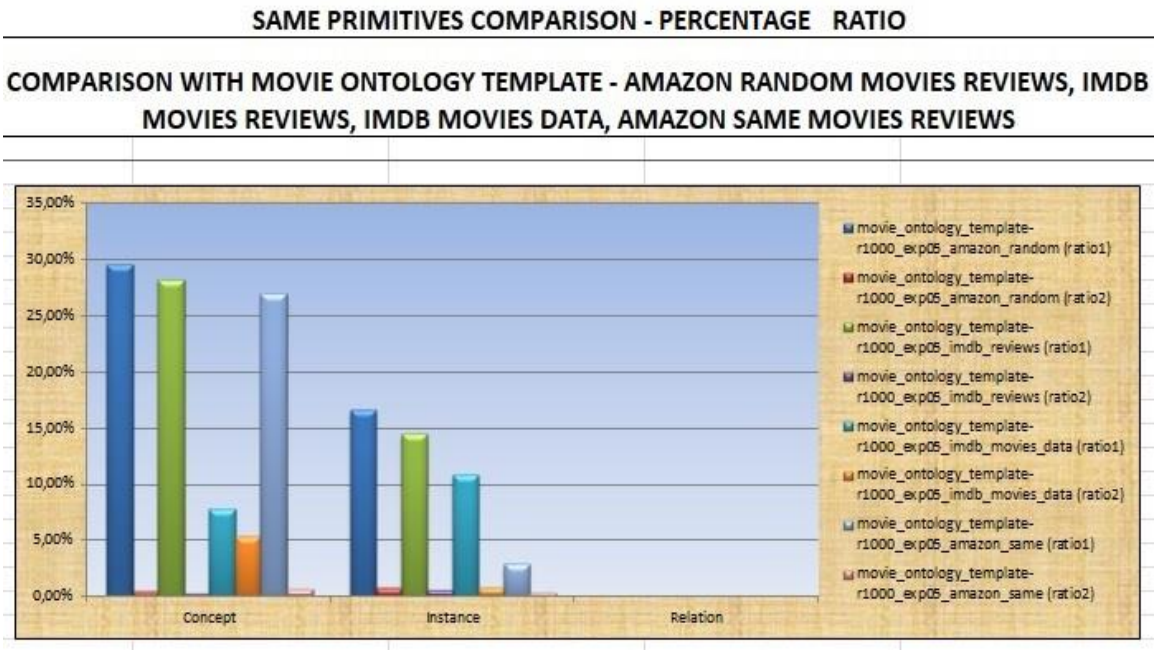
	movie_ontology_template
Concept	78
Instance	139
Relation	38

Εικόνα 8-33: Αριθμός στοιχείων οντολογίας του προτύπου οντολογίας σχετικά με κινηματογραφικές ταινίες

Οι συγκρίσεις που πραγματοποιήθηκαν μεταξύ του προτύπου οντολογίας και των πειραμάτων των τεσσάρων κατηγοριών δεδομένων εισόδου έδωσαν ένα γράφημα το οποίο φαίνεται παρακάτω στην εικόνα 8-34. Στις συγκρίσεις μεταξύ κάθε πειράματος και του προτύπου οντολογίας υπολογίζεται για κάθε θεμελιακό στοιχείο ο ποσοστιαίος λόγος που δίνει τον αριθμό των κοινών στοιχείων ως προς τον αριθμό των στοιχείων του προτύπου οντολογίας που παραμένει σταθερός. Ο λόγος αυτός που θα ονομαστεί *ratio1* έχει την παρακάτω μορφή όπου ***Κοινά*** ο αριθμός των κοινών στοιχείων οντολογίας και ***Στοιχεία Προτύπου*** ο αριθμός των στοιχείων του προτύπου οντολογίας:

$$\frac{\boldsymbol{Κοινά}}{\boldsymbol{Στοιχεία\ Προτύπου}}$$

Υπολογίζεται επίσης για κάθε θεμελιακό στοιχείο ο συντελεστής ομοιότητας *Dice* όπως περιγράφηκε στην ενότητα 8.3 ως ο αριθμός των κοινών στοιχείων οντολογίας των δύο οντολογιών επί δύο προς το άθροισμα των στοιχείων της κάθε οντολογίας αντίστοιχα. Ο συντελεστής αυτός θα ονομαστεί *ratio2* στην προκειμένη περίπτωση.

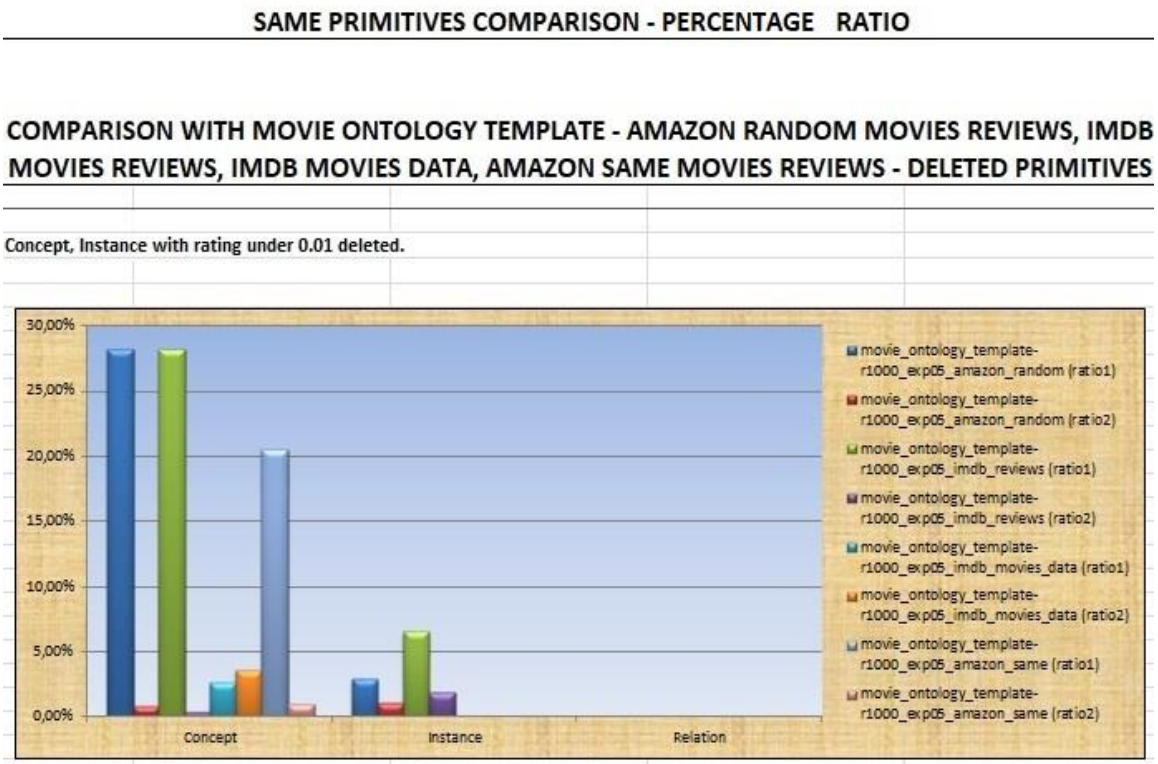


Εικόνα 8-34: Σύγκριση ως προς τον αριθμό κοινών στοιχείων οντολογιών με πρότυπο οντολογίας κινηματογραφικών ταινιών – είδος πειράματος 05

Ξεκινώντας την παρατήρηση από το θεμελιακό στοιχείο *Concept* βλέπουμε πως ο ποσοστιαίος λόγος *ratio1* για τις τρεις κατηγορίες δεδομένων εισόδου εκτός της *imdb_movies_data*, μας δίνει ένα ποσοστό κοινών στοιχείων ανάμεσα σε 25% και 30%, Η κατηγορία *amazon_random* εμφανίζει τα καλύτερα αποτελέσματα με τις κατηγορίες *imdb_reviews* και *amazon_same* να έπονται κατά σειρά. Αντίθετα τα αποτελέσματα του ποσοστιαίου λόγου *ratio2* παραμένει σε σχεδόν μηδενικά επίπεδα. Για την κατηγορία *imdb_movies_data* βλέπουμε πως ο λόγος *ratio1* είναι γύρω στο 7,5% ενώ ο *ratio2* αντίθετα με τις άλλες κατηγορίες είναι στο 5%. Ως προς το θεμελιακό στοιχείο *Instance*, για τον λόγο *ratio1*, τα ποσοστά των τεσσάρων κατηγοριών κυμαίνονται από 17% έως 3% για τις κατηγορίες *amazon_random, imdb_reviews, imdb_movies_data* και *amazon_same* κατά σειρά. Αντίθετα ο λόγος *ratio2* παραμένει σε σχεδόν μηδενικά επίπεδα για όλες τις κατηγορίες. Αυτό συμβαίνει λόγω του μεγάλου παρανομαστή καθώς ενώ βρίσκονται τα κοινά στοιχεία, λόγω του μεγάλου αριθμού στοιχείων οντολογίας που μας δίνουν τα αποτελέσματα των πειραμάτων υπάρχει και πολύ μεγάλος αριθμός μη κοινών. Τέλος όσον αφορά το θεμελιακό στοιχείο *Relation* δεν βρέθηκε κανένα κοινό στοιχείο στις συγκρίσεις του προτύπου και με τα τέσσερα πειράματα.

Στη συνέχεια θα ήταν χρήσιμο να δούμε πώς αλλάζει η συμπεριφορά του γραφήματος αν για τα πειράματα και των τεσσάρων κατηγοριών γίνει διαγραφή στοιχείων οντολογίας με χαμηλή τιμή πιθανότητας κάτω από ένα κατώφλι. Στην

περίπτωση αυτή επιλέχθηκε για τα θεμελιακά στοιχεία το κατώφλι τιμής πιθανότητας να είναι 0,01. Παρακάτω στην εικόνα 8-35 βλέπουμε το νέο γράφημα που προκύπτει.



Εικόνα 8-35: Σύγκριση ως προς τον αριθμό κοινών στοιχείων οντολογιών με πρότυπο οντολογίας κινηματογραφικών ταινιών με διαγραφή θεμελιακών στοιχείων – είδος πειράματος 05

Ξεκινώντας από το θεμελιακό στοιχείο *Concept* βλέπουμε πως για τον ποσοστιαίο λόγο *ratio1* οι κατηγορίες *amazon_random* και *imdb_reviews* σχεδόν διατήρησαν τα ποσοστά τους, αντίθετα η κατηγορία *amazon_same* έχει μια πτώση του ποσοστού της στην τιμή του 20%. Το ποσοστό της κατηγορίας *imdb_movies_data* για τον λόγο *ratio1* επίσης παρουσιάζει πτώση στην τιμή κοντά στο 3%, με την τιμή του λόγου *ratio2* να γίνεται τώρα ελαφρώς υψηλότερη παρότι παρουσιάζει και αυτή πτώση. Οι τιμές του λόγου *ratio2* για τις άλλες τρεις κατηγορίες παραμένουν σε σχεδόν μηδενικά επίπεδα. Για το θεμελιακό στοιχείο *Instance* για τον λόγο *ratio1* βλέπουμε πως η πτώση των ποσοστών είναι ακόμα μεγαλύτερες με τις κατηγορίες *imdb_movies_data* και *amazon_same* να παρουσιάζουν πια μηδενικά ποσοστά. Οι κατηγορίες *amazon_random* και *imdb_reviews* εμφανίζουν και αυτές πτώση των ποσοστών τους, με την πρώτη να εμφανίζει εμφανώς μεγαλύτερη πτώση στην τιμή κοντά στο 3% με τη δεύτερη να μένει κοντά στο 6%.

8.9 Είσοδος εξαγόμενης οντολογίας από την εφαρμογή Text2Onto στο περιβάλλον οντολογιών Protégé

Ένας ακόμα βασικός στόχος της μελέτης ήταν να μπορέσουμε να χρησιμοποιήσουμε τα αρχεία *owl* των εξαγόμενων οντολογιών από τα πειράματα με το *Text2Onto* ως είσοδο στο περιβάλλον οντολογιών *Protégé* για δυνατότητα περαιτέρω επεξεργασίας και οπτικοποίησης. Στην ενότητα αυτή δίνονται τα βασικά βήματα αυτής της διαδικασίας.

Αρχικά δημιουργήθηκε ένα νέο είδος πειράματος, το έκτο κατά σειρά. Επιλέχθηκε η κατηγορία δεδομένων εισόδου *amazon_same* που αφορά κριτικές της *Amazon* από ίδιες ταινίες. Η ποσότητα των δεδομένων εισόδου είναι 1000 κριτικές. Ως προς την επιλογή των αλγορίθμων επιλέχθηκαν οι αλγόριθμοι που μας έδωσαν τα καλύτερα αποτελέσματα ανά θεμελιακό στοιχείο όπως αναλύθηκε σε προηγούμενες ενότητες. Οι αλγόριθμοι αυτοί συνοψίζονται παρακάτω, στον πίνακα 8-1.

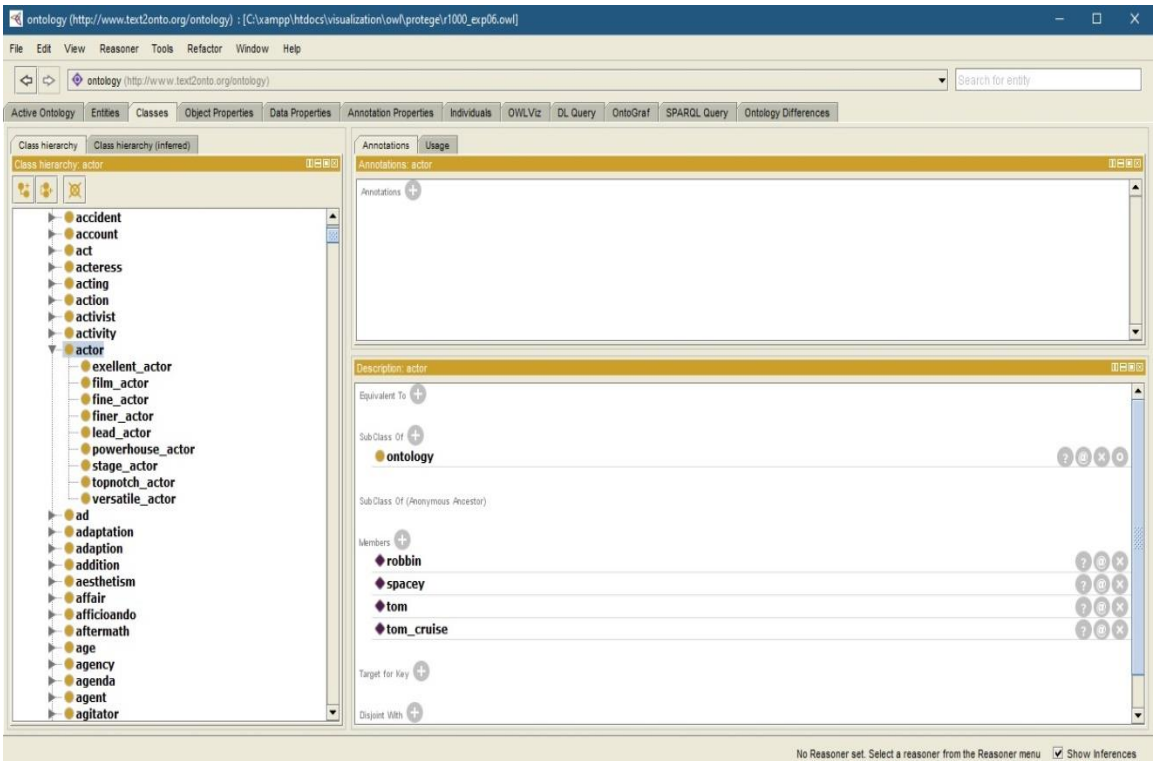
Πείραμα 01	
Concept	*TFIDFConceptExtraction*
Instance	*TFIDFInstanceExtraction*
SubclassOf	*VerticalRelationsConceptClassification*
InstanceOf	*PatternInstanceClassification*
Relation	*SubcatRelationExtraction*

Πίνακας 8-1: Οι αλγόριθμοι του έκτου είδους πειράματος ανά θεμελιακό στοιχείο

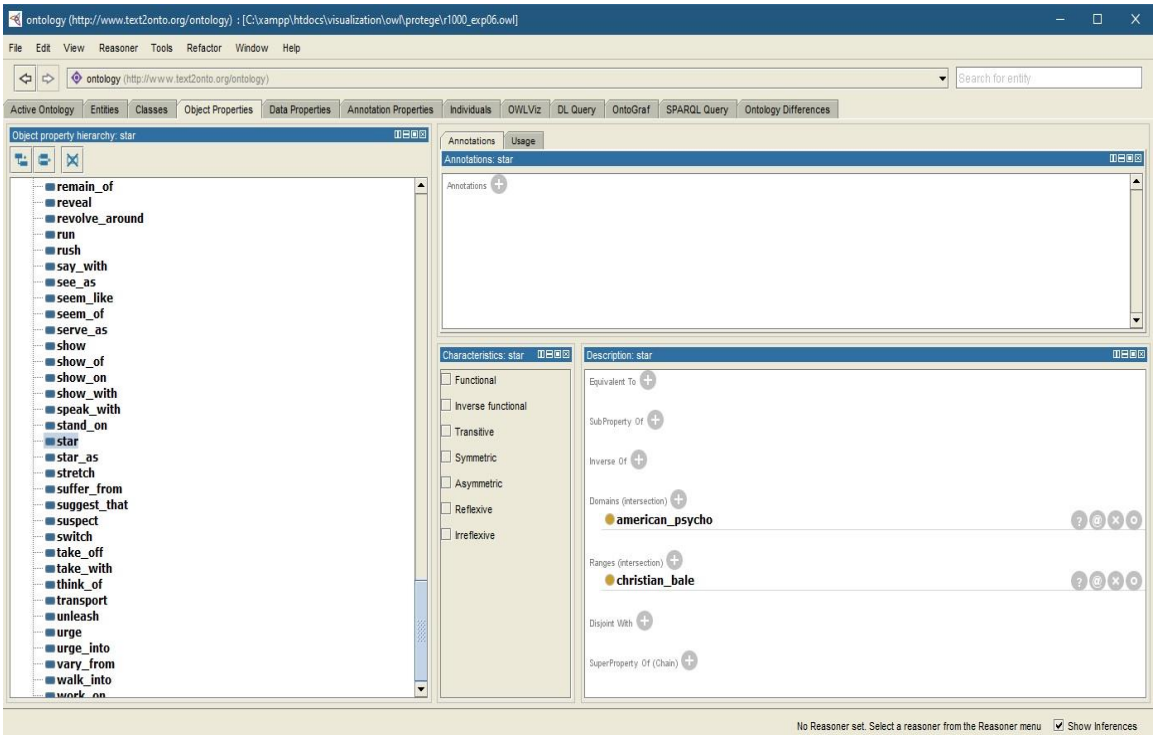
Από το αρχείο *owl* που εξάχθηκε από το *Text2Onto* ως αποτέλεσμα του πειράματος επιλέχθηκε να διαγραφούν στοιχεία οντολογίας για κάθε θεμελιακό στοιχείο κάτω από ένα κατώφλι πιθανότητας μέσω της διαδικασίας διαγραφής θεμελιακών στοιχείων που περιγράφηκε στο 7^{ο} κεφάλαιο. Ως κατώφλι πιθανότητας για τα θεμελιακά στοιχεία *Concept* και *Instance* επιλέχθηκε η τιμή 0,1 και για τα θεμελιακά στοιχεία *SubclassOf, InstanceOf* και *Relation*, η τιμή 0,4. Αφού εξαχθεί το νέο *owl* αρχείο μετά τις διαγραφές, μέσω της διαδικασίας μετατροπής θεμελιακών στοιχείων που επίσης περιγράφηκε στο 7^{ο} κεφάλαιο, μετατρέπεται σε *owl* αρχείο που είναι πλέον κατάλληλο για είσοδο στο περιβάλλον οντολογιών *Protégé*.

Παρακάτω στις εικόνες 8-36 έως 8-41 ακολουθούν εικόνες από τις καρτέλες *Classes, Object Properties, Individuals* και *OWL Viz* του *Protégé* που αφορούν την οντολογία που εξάχθηκε από το παραπάνω πείραμα. Στις καρτέλες αυτές φαίνονται

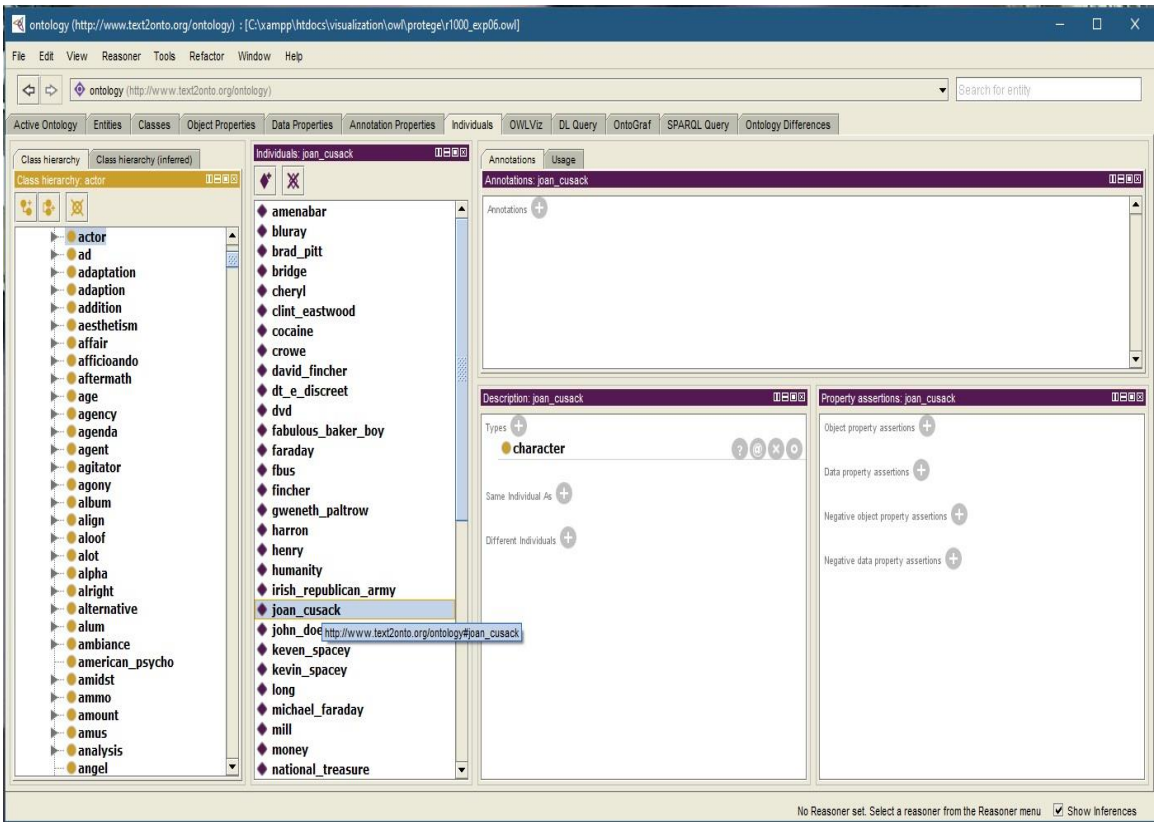
αντίστοιχα οι κλάσεις, οι σχέσεις, τα άτομα και μια οπτικοποίηση της ιεραρχίας των κλάσεων της οντολογίας.



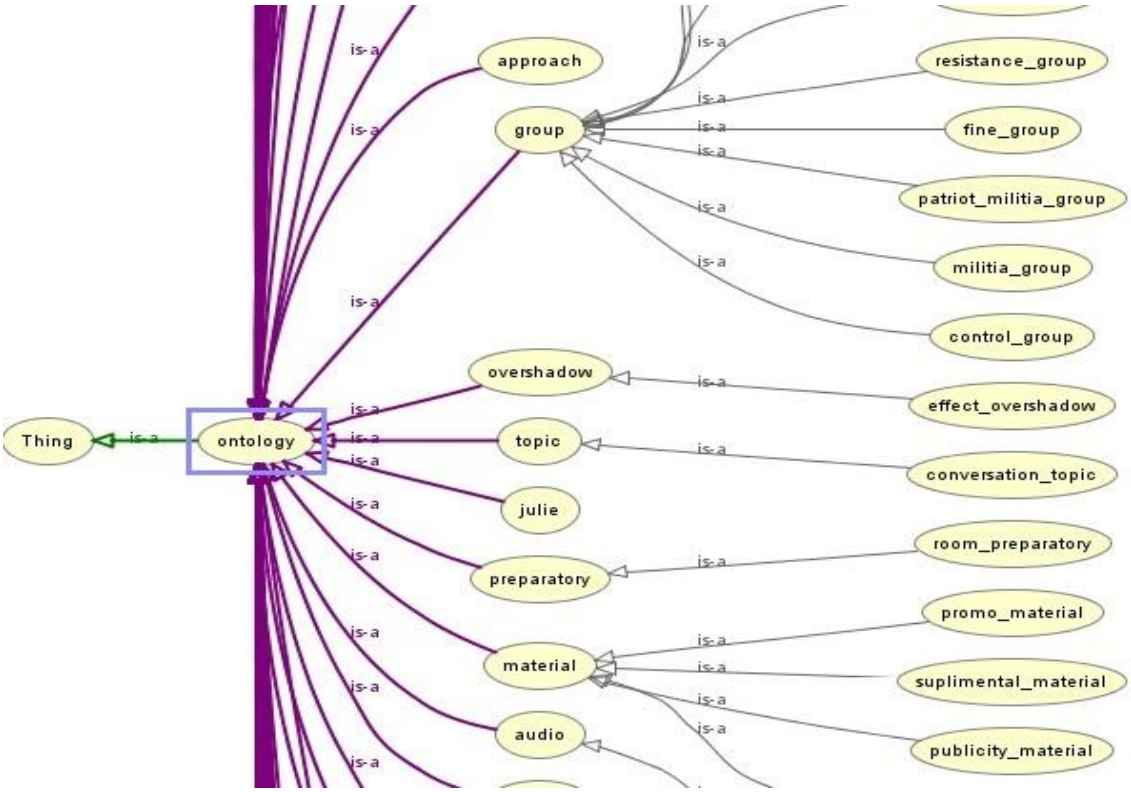
Εικόνα 8-36: Η καρτέλα "Classes" της οντολογίας του πειράματος r1000_exp06_amazon_same



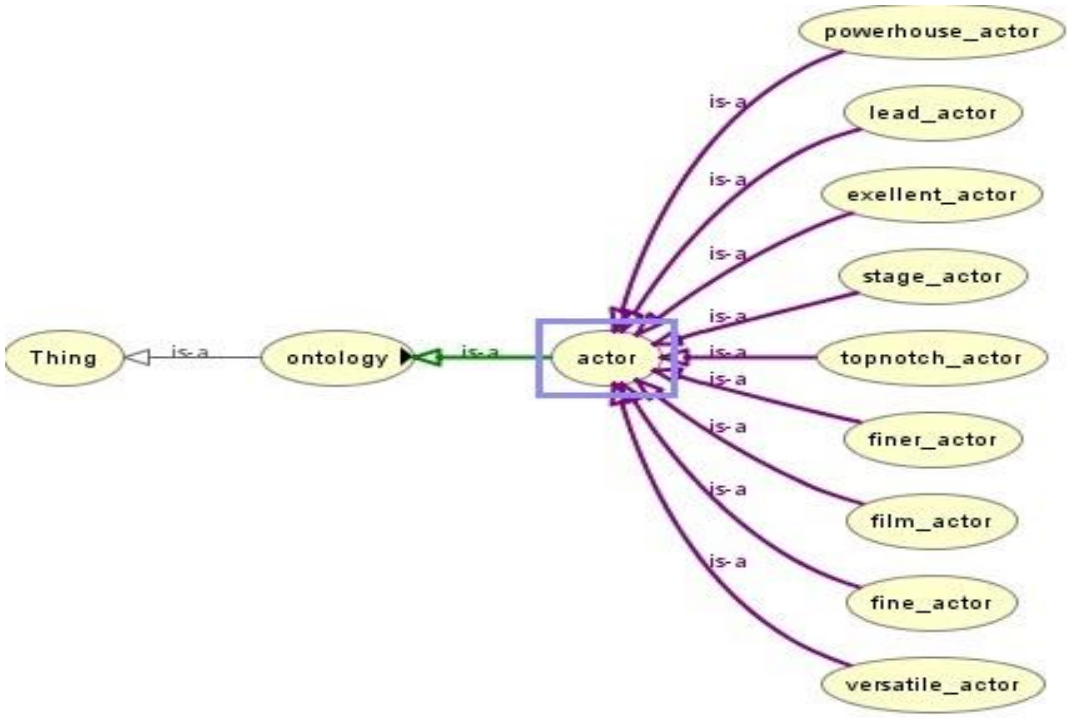
Εικόνα 8-37: Η καρτέλα "Object Properties" της οντολογίας του πειράματος r1000_exp06_amazon_same



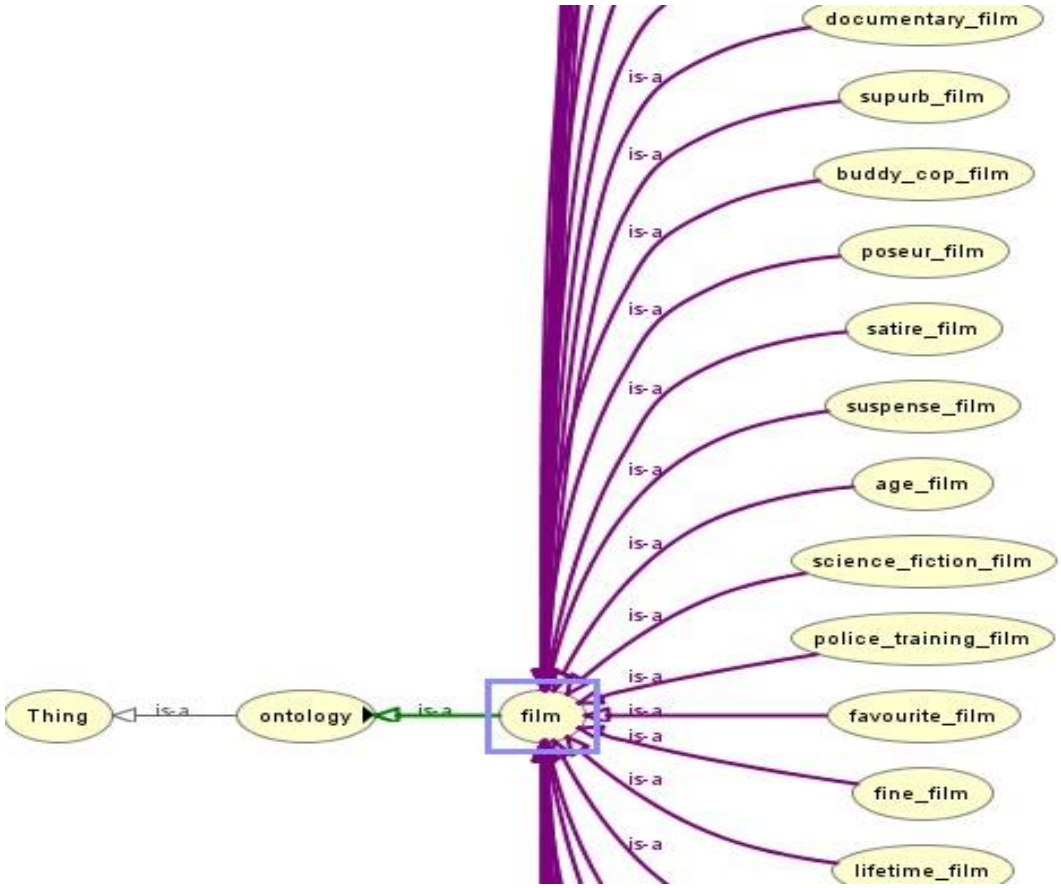
Εικόνα 8-38: Η καρτέλα "Individuals" της οντολογίας του πειράματος r1000_exp06_amazon_same



Εικόνα 8-39: Η καρτέλα "OWL Viz" της οντολογίας του πειράματος r1000_exp06_amazon_same



Εικόνα 8-40: Η καρτέλα "OWL Viz" της οντολογίας του πειράματος r1000_exp06_amazon_same ως προς την κλάση actor



Εικόνα 8-41: Η καρτέλα "OWL Viz" της οντολογίας του πειράματος r1000_exp06_amazon_same ως προς την κλάση film

9 Επίλογος

9.1 Σύνοψη και συμπεράσματα

Όπως αναφέρθηκε και στο εισαγωγικό κεφάλαιο της εργασίας βασικός στόχος της ήταν η πειραματική μελέτη της εφαρμογής εκμάθησης οντολογιών *Text2Onto* για την αυτόματη εξαγωγή οντολογικής γνώσης από αδόμητο κείμενο. Έγινε προσπάθεια μέσω πειραμάτων διαφόρων ειδών και μέσω της ποσοτικής και ποσοτικής ανάλυσης των αποτελεσμάτων τους να εξαχθούν όσο το δυνατόν περισσότερα και ασφαλή συμπεράσματα σχετικά με την έκταση των δυνατοτήτων και τους τρόπους εφαρμογής του εργαλείου *Text2Ont* στον τομέα των οντολογιών καθώς και για το βαθμό αξιοποίησης κριτικών κινηματογραφικών ταινιών στη δημιουργία οντολογίας.

Συνοπτικά αυτά που ήταν ήδη γνωστά από προηγούμενες μελέτες και εφαρμογές σχετικά με τις δυνατότητες της εφαρμογής *Text2Onto* ως εργαλείο εκμάθησης οντολογιών είναι τα εξής (Philip Cimiano, Johanna Volker (2015, σ.44)):

- Η εφαρμογή *Text2Onto* δεν μπορεί να κατασκευάσει σε πλήρη έκταση αυτόματα μια οντολογία μέσω εκμάθησης επάνω σε ένα υποκείμενο σώμα δεδομένων αδόμητου κειμένου.
- Η εφαρμογή *Text2Onto* μπορεί να λειτουργήσει ως βοηθητικό εργαλείο του χρήστη στην κατασκευή μιας οντολογίας αλλά χρειάζεται βελτίωση.

Τα παραπάνω συμπεράσματα επιβεβαιώθηκαν και από τη συγκεκριμένη μελέτη. Η εφαρμογή *Text2Onto* έχει ορισμένες δυνατότητες που μπορούν να βοηθήσουν ως βασικός σκελετός στην κατεύθυνση της κατασκευής μιας οντολογίας αλλά παρουσιάζει και αρκετές αδυναμίες. Όσον αφορά αυτές τις αδυναμίες, ειδικότερα επάνω στον τομέα των κινηματογραφικών ταινιών που πραγματοποιήθηκαν τα πειράματα, παρατηρήθηκαν τα εξής:

- Τα αποτελέσματα για τα θεμελιακά στοιχεία *Concep*t και *Instance* σε όλα τα πειράματα που πραγματοποιήθηκαν έδιναν στοιχεία οντολογίας μεγάλου πλήθους με μικρές τιμές πιθανότητας. Ειδικότερα το πλήθος τους αυξανόταν πολύ καθώς αυξανόταν και η ποσότητα των δεδομένων εισόδου χωρίς να αυξάνονται ιδιαίτερα οι τιμές πιθανότητάς τους.
- Για το θεμελιακό στοιχείο *Similarity* η εφαρμογή δεν έδωσε αποτελέσματα σε κανένα από τα πειράματα που πραγματοποιήθηκαν.

- Το αρχείο *owl* που εξαγόταν από τα αποτελέσματα των πειραμάτων είχε μορφή που δεν το καθιστούσε κατάλληλο για να χρησιμοποιηθεί ως είσοδος στο περιβάλλον οντολογιών *Protégé* για δυνατότητα περαιτέρω επεξεργασίας της οντολογίας.

Τα θετικά στοιχεία που παρατηρήθηκαν από το σύνολο των πειραμάτων μπορούν να συνοψισθούν στα εξής:

- Τα αποτελέσματα για τα θεμελιακά στοιχεία *SubclassOf, InstanceOf* και *Relation* έδιναν στοιχεία οντολογίας με αρκετά μεγάλες τιμές πιθανότητας.
- Η σύγκριση των εξαγόμενων οντολογιών των πειραμάτων ως προς τον αριθμό των κοινών στοιχείων με ένα πρότυπο οντολογίας σχετικά με κινηματογραφικές ταινίες έδωσε σε κάποιες περιπτώσεις πειραμάτων θετικά αποτελέσματα.

Όσον αφορά το βαθμό αξιοποίησης των κριτικών κινηματογραφικών ταινιών για τη δημιουργία οντολογίας τα συμπεράσματα προκύπτουν σε συνδυασμό με αυτά που ειπώθηκαν πιο πάνω σχετικά με τις δυνατότητες του Text2Onto. Αυτό που μπορεί να λεχθεί είναι πως πράγματι οι κριτικές κινηματογραφικών ταινιών μπορούν να χρησιμοποιηθούν για τη δημιουργία οντολογίας. Οι περιορισμοί και ο βαθμός αξιοποίησης τους έχουν να κάνουν κυρίως με τις δυνατότητες του εργαλείου που δεν δίνουν τη δυνατότητα δημιουργίας της οντολογίας αυτόματα και σε πλήρη έκταση αλλά μπορεί να δημιουργηθεί ένας σκελετός που θα βοηθήσει στην περαιτέρω κατασκευή και βελτίωση της οντολογίας.

Μετά την ολοκλήρωση των πειραμάτων, μέσω των διαδικασιών κώδικα που αναπτύχθηκαν για την ποσοτική και ποιοτική ανάλυση των αποτελεσμάτων των πειραμάτων, έγινε μια προσπάθεια βελτίωσης των δυνατοτήτων της εφαρμογής *Text2Onto* ως βοηθητικού εργαλείου για την κατασκευή μιας οντολογίας. Μέσω της διαδικασίας διαγραφής θεμελιακών στοιχείων δόθηκε η δυνατότητα μαζικής διαγραφής από την εξαγόμενη οντολογία, στοιχείων οντολογίας με τιμές πιθανότητας κάτω από ένα κατώφλι που μπορούμε να ορίσουμε με σκοπό τη βελτίωση της εξαγόμενης οντολογίας. Επίσης μέσω της διαδικασίας μετατροπής θεμελιακών στοιχείων δόθηκε η δυνατότητα η εξαγόμενη οντολογία να μπορεί να εισαχθεί ως είσοδος στο περιβάλλον οντολογιών *Protégé* για δυνατότητα περαιτέρω επεξεργασίας.

9.2 Μελλοντικές Επεκτάσεις

Στη συγκεκριμένη μελέτη έγινε πειραματισμός με δεδομένα εισόδου ενός συγκεκριμένου τομέα ενδιαφέροντος, των κινηματογραφικών ταινιών και με ποσοτική αύξηση των δεδομένων εισόδου μέχρι ενός συγκεκριμένου ορίου.

Πιθανός πειραματισμός με δεδομένα διαφορετικού τομέα ενδιαφέροντος ή και με περαιτέρω ποσοτική αύξηση των δεδομένων εισόδου σε συνδυασμό με τις διαδικασίες που αναπτύχθηκαν για την ποσοτική και ποιοτική ανάλυση των αποτελεσμάτων των πειραμάτων θα μπορούσε στο μέλλον να δώσει επιπλέον συμπεράσματα και βελτιώσεις στο ρόλο της εφαρμογής *Text2Onto* ως βοηθητικό εργαλείο για την εξαγωγή οντολογικής γνώσης από αδόμητο κείμενο. Οι παραπάνω διαδικασίες κώδικα θα μπορούσαν και αυτές με τη σειρά τους να βελτιωθούν και να εμπλουτιστούν προς αυτήν την κατεύθυνση.

Βιβλιογραφία

Dean Allemang, James Hendler (2011). *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL*. Morgan Kaufmann.

Γιάου Ευαγγελία (2013). *Ο Σημασιολογικός Ιστός και η σχέση του με την Λογική Επιχειρηματολογία.* Πανεπιστήμιο Πελοποννήσου. [pdf] Διαθέσιμο: <http://amitos.library.uop.gr/xmlui/bitstream/handle/123456789/975/358_000028m.pdf?sequence=1&isAllowed=y> [03 Απριλίου 2020].

Τσουκαλά Αναστασία (2018). *Οντολογίες*. Πανεπιστήμιο Θεσσαλίας. [pdf] Διαθέσιμο: <https://www.academia.edu/38713246/Ontologies_and_the_Semantic_Web-_Greek> [31 Μαρτίου 2020].

Matthew Horridge (2011). *A Practical Guide To Building OWL Ontologies Using Protégé 4 and CO-ODE Tools*. The University of Manchester. [pdf] Διαθέσιμο: <http://mowl-power.cs.man.ac.uk/protegeowltutorial/resources/ProtegeOWLTutorialP4_v1_3.pdf> [14 Οκτωβρίου 2019].

Philip Cimiano, Johanna Volker (2005). *Text2Onto A Framework for Ontology Learning and Data-driven Change Discovery*. Conference: Natural Language Processing and Information Systems, 10th International Conference on Applications of Natural Language to Information Systems, NLDB 2005, Alicante, Spain, June 15-17, 2005. [pdf] Διαθέσιμο: <https://www.researchgate.net/publication/221474301_Text2onto> [14 Οκτωβρίου 2019].

Philip Cimiano, Johanna Volker (2015). *Text2Onto A Framework for Ontology Learning and Data-driven Change Discovery*. University of Saarland. [pdf] Διαθέσιμο: <http://docplayer.net/39151026-Text2onto-a-framework-for-ontology-learning-and-data-driven-change-discovery-philipp-cimiano-johanna-volker.html> [14 Οκτωβρίου 2019].

Sonam Mittal, Nupur Mittal (2013). *Tools for Ontology Building from Texts. Analysis and Improvement of the Results of Text2Onto.* IOSR Journal of Computer Engineering. e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 11, Issue 2 (May. - Jun. 2013), PP 101-117. [pdf] Διαθέσιμο: <https://www.researchgate.net/publication/221474301_Text2onto> [14 Οκτωβρίου 2019].

Johanna Volker, York Sure (2005). *D3.3.1 Data-driven Change Discovery*. EU-IST Integrated Project (IP) IST-2003-506826 SEKT. [pdf] Διαθέσιμο: <http://www.sekt-project.com/rd/deliverables/wp03/sekt-d-3-3-1-Data-driven%20Change%20Discovery.pdf > [15 Οκτωβρίου 2019].

Johanna Volker, Denny Vrandecic, York Sure (2005). *D3.3.2 Data-driven Change Discovery - Evaluation.* EU-IST Integrated Project (IP) IST-2003-506826 SEKT. [pdf] Διαθέσιμο: <http://www.sekt-project.com/rd/deliverables/wp03/sekt-d-3-3-2-Data-driven%20Change%20Discovery%20Evaluation.pdf> [15 Οκτωβρίου 2019].

Johanna Volker, Denny Vrandecic, York Sure (2006). *D3.3. Data-driven Change Discovery.* EU-IST Integrated Project (IP) IST-2003-506826 SEKT. [pdf] Διαθέσιμο: <https://pdfs.semanticscholar.org/0bc1/97161e5f39b683c65b5b5dd0d25536ce9aa1.pdf?_ga=2.85745755.316916459.1585503076-1729321252.1585503076> [15 Οκτωβρίου 2019].

Peter Haase, Johanna Volker (2005). *Ontology Learning and Reasoning - Dealing with Uncertainty and Inconsistency.* International Workshop on Uncertainty Reasoning for the Semantic Web 2005. [pdf] Διαθέσιμο: <https://link.springer.com/chapter/10.1007%2F978-3-540-89765-1_21> [16 Οκτωβρίου 2019].

Han-Hsiang Wang, Frank Boukamp (2008). *A Context Ontology Development Process for Construction Safety.* Joint CIB Conference W102 Information and Knowledge Management in Building, Helsinki, 3–4 June 2008. [pdf] Διαθέσιμο: <https://pdfs.semanticscholar.org/7f4e/2cf552ecb16572b14808106d3e104fcd87bd.pdf?_ga=2.72027217.316916459.1585503076-1729321252.1585503076> [16 Οκτωβρίου 2019].

Jianmo Ni. *Amazon Review Data (2018).* Διαθέσιμο: <https://nijianmo.github.io/amazon/index.html> [20 Νοεμβρίου 2019].

Lakshmipathi N. *IMDB Dataset of 50K Movie Reviews. Large Movie Review Dataset.* Διαθέσιμο: <https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews> [16 Δεκεμβρίου 2019].

Daniel Grijalva. *Movie Industry. Three decades of movies.* Διαθέσιμο: <https://www.kaggle.com/danielgrijalvas/movies> [16 Δεκεμβρίου 2019].

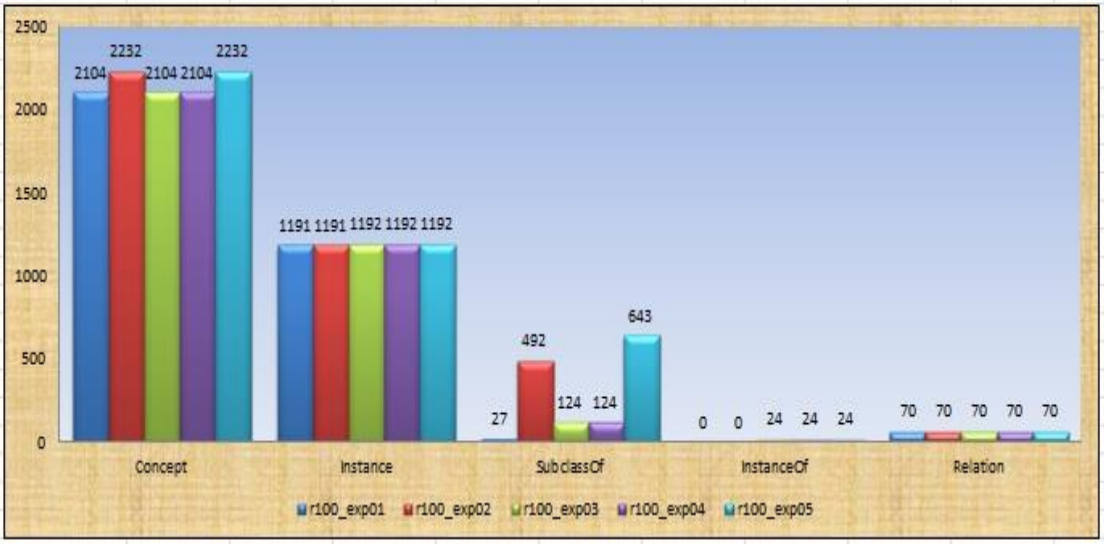
Amancio Bouza (2010). *The movie ontology MO*. Διαθέσιμο: <http://www.movieontology.org/2010/01/movieontology.owl> [16 Δεκεμβρίου 2019].

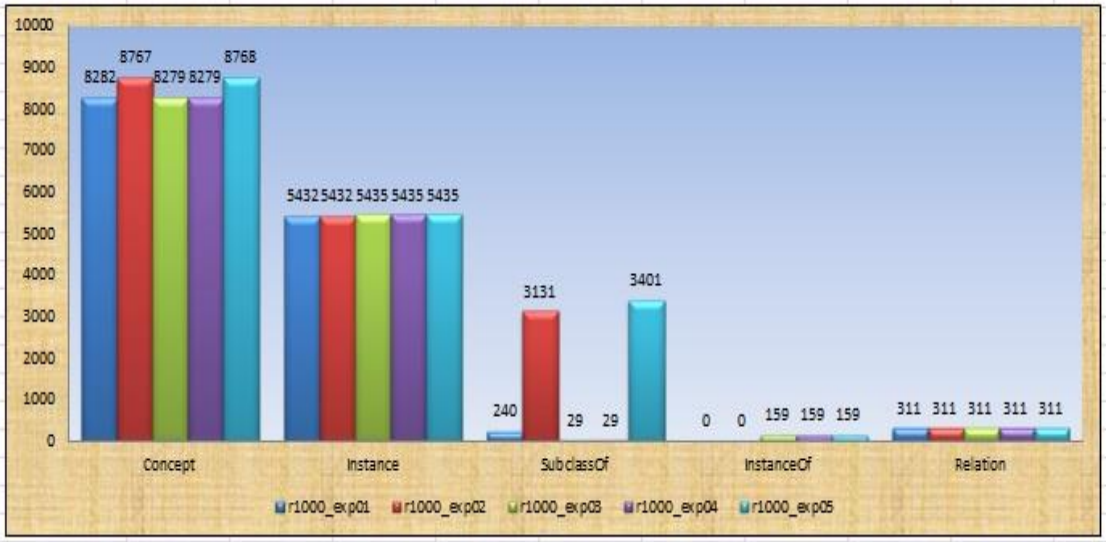
Παράρτημα Α - Γραφήματα

Σε αυτό το παράρτημα παρατίθενται τα αποτελέσματα όλων των αναλύσεων αποτελεσμάτων των πειραμάτων που πραγματοποιήθηκαν σε μορφή γραφημάτων. Η κωδικοποίηση της ονομασίας των πειραμάτων έχει τη μορφή ***rX_expY_Z***, όπου *X* παίρνει τιμές 20, 100, 1000 ανάλογα με την ποσότητα των δεδομένων (κριτικών, στοιχείων ταινιών), όπου *Y* παίρνει τιμές 01, 02, 03, 04, 05 ανάλογα με το είδος πειραμάτων και όπου *Z* να παίρνει τιμές *amazon_random*, *amazon_same*, *imdb_reviews*, *imdb_movies_data* ανάλογα με την κατηγορία των δεδομένων. Στις περιπτώσεις όπου τα γραφήματα είναι για μια συγκεκριμένη κατηγορία δεδομένων που αναγράφεται στον τίτλο του γραφήματος, στην κωδικοποίηση ονομασίας των πειραμάτων αναγράφονται μόνο τα δύο πρώτα μέρη σε μορφή ***rX_expY***.

ENUMERATION

AMAZON RANDOM MOVIES REVIEWS (PER NUMBER OF REVIEWS)

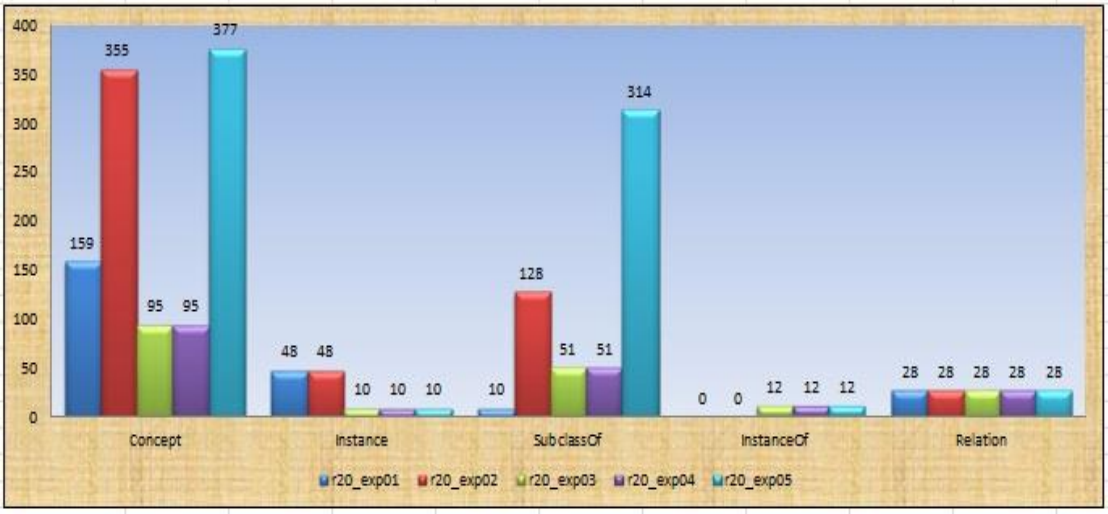


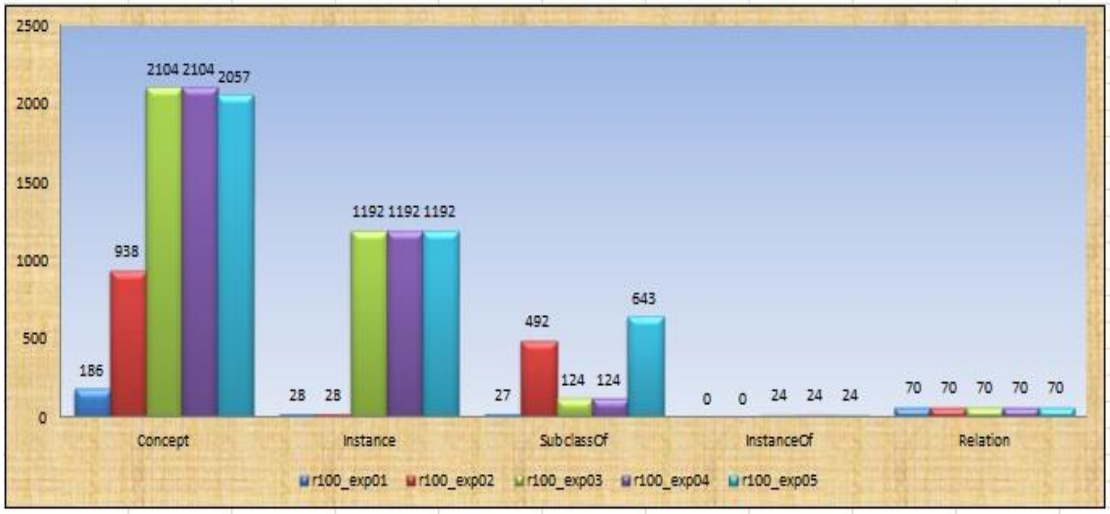


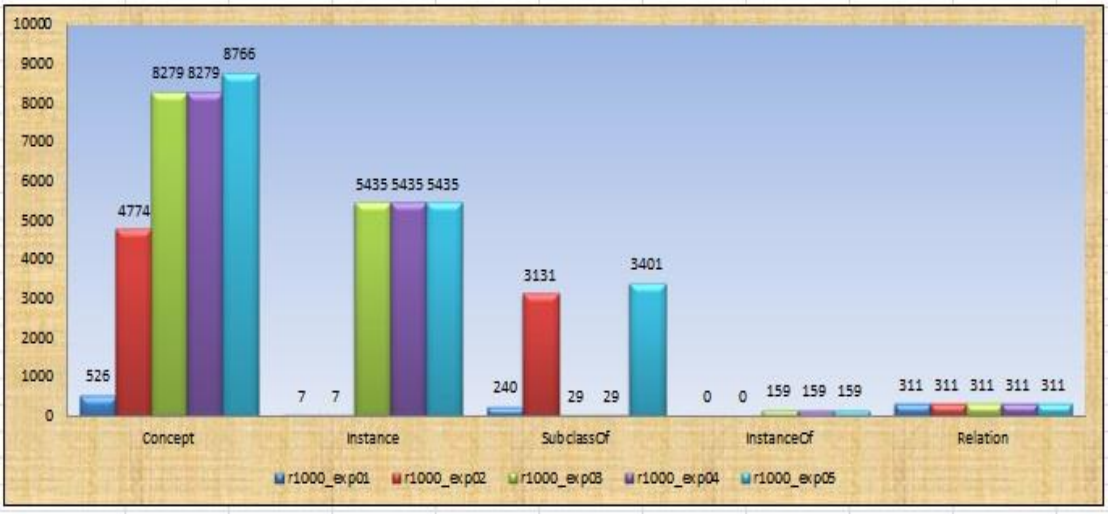
ENUMERATION

AMAZON RANDOM MOVIES REVIEWS - DELETED PRIMITIVES (PER NUMBER OF REVIEWS)

Concept, Instance with rating under 0.001 deleted.

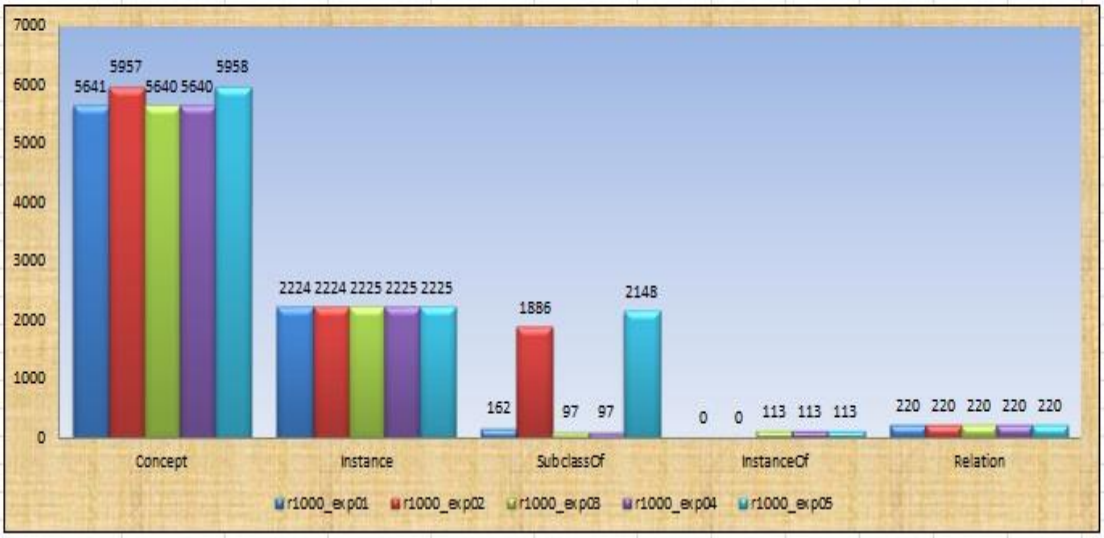






ENUMERATION

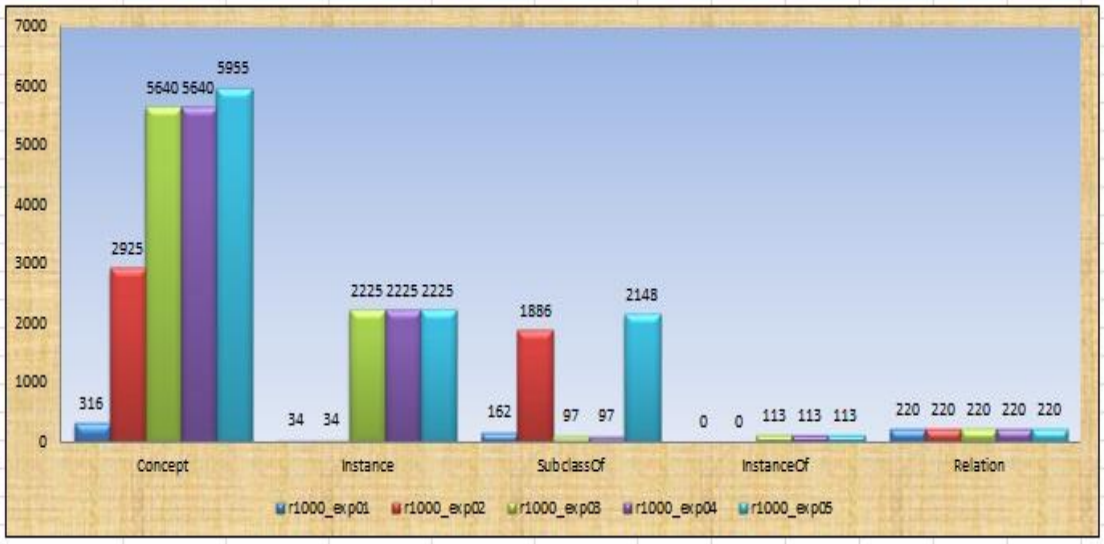
AMAZON SAME MOVIES REVIEWS (PER NUMBER OF REVIEWS)



ENUMERATION

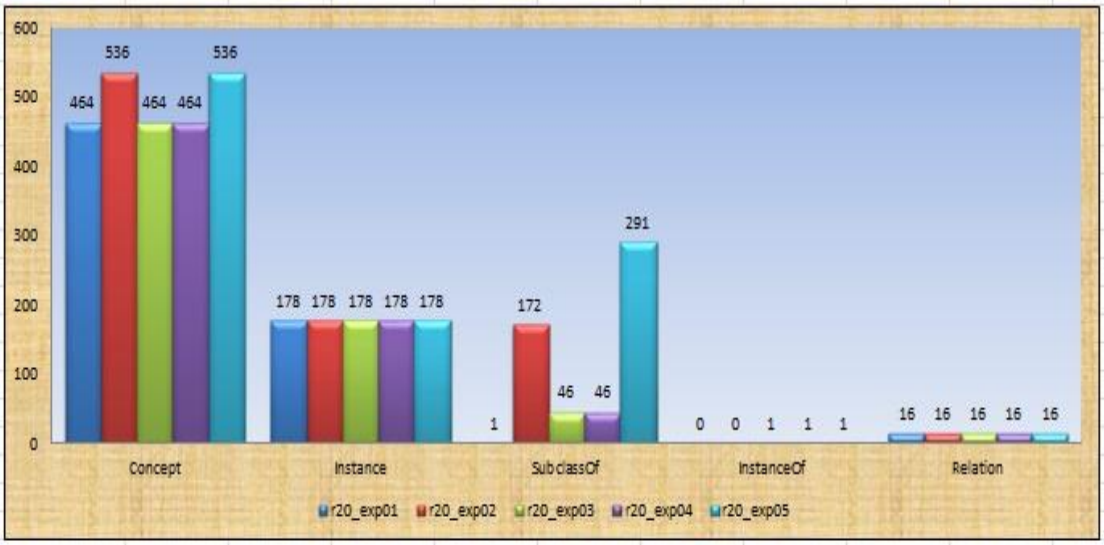
AMAZON SAME MOVIES REVIEWS - DELETED PRIMITIVES (PER NUMBER OF REVIEWS)

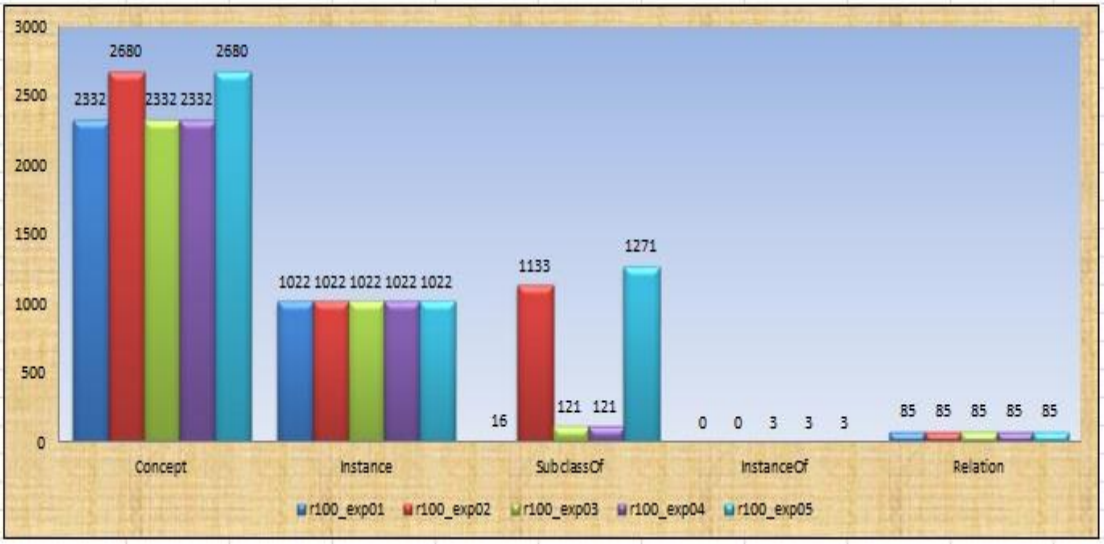
Concept, Instance with rating under 0.001 deleted.



ENUMERATION

IMDB MOVIES REVIEWS (PER NUMBER OF REVIEWS)

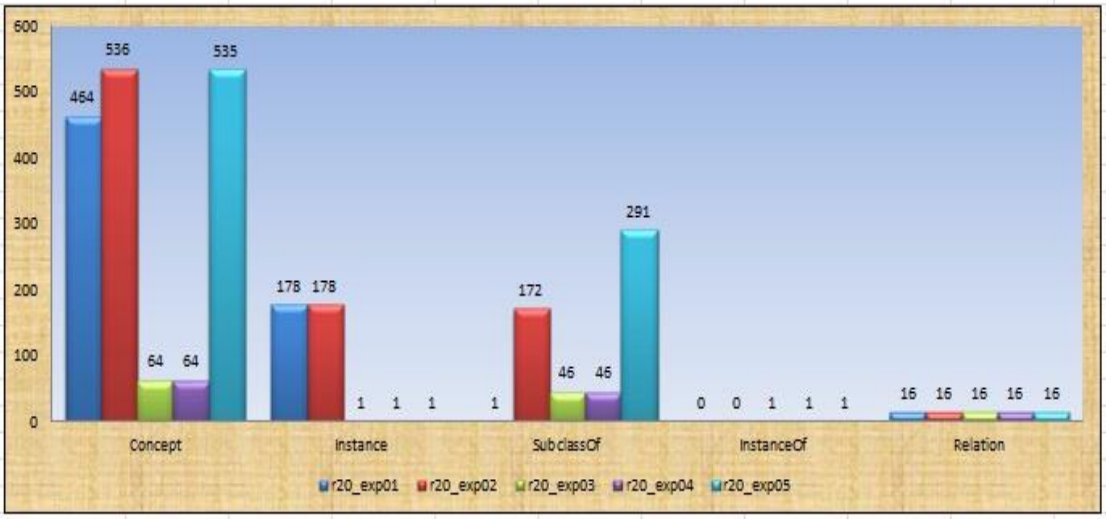


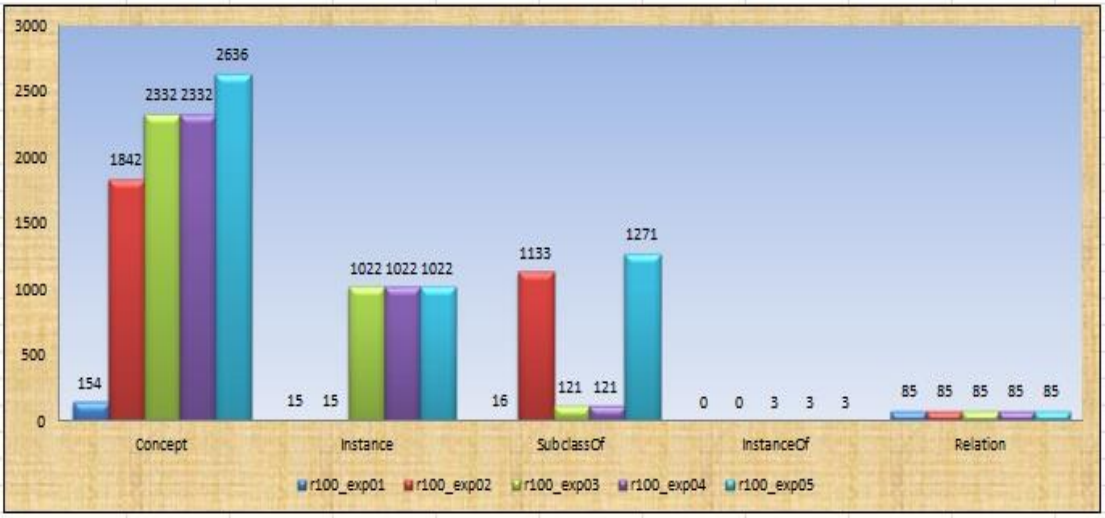


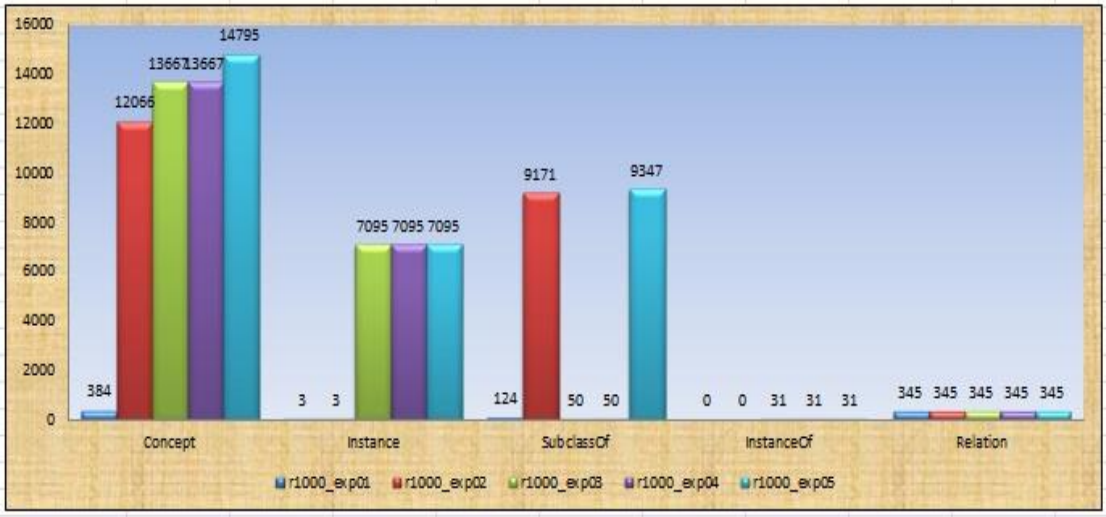
ENUMERATION

IMDB MOVIES REVIEWS - DELETED PRIMITIVES (PER NUMBER OF REVIEWS)

Concept, Instance with rating under 0.001 deleted.

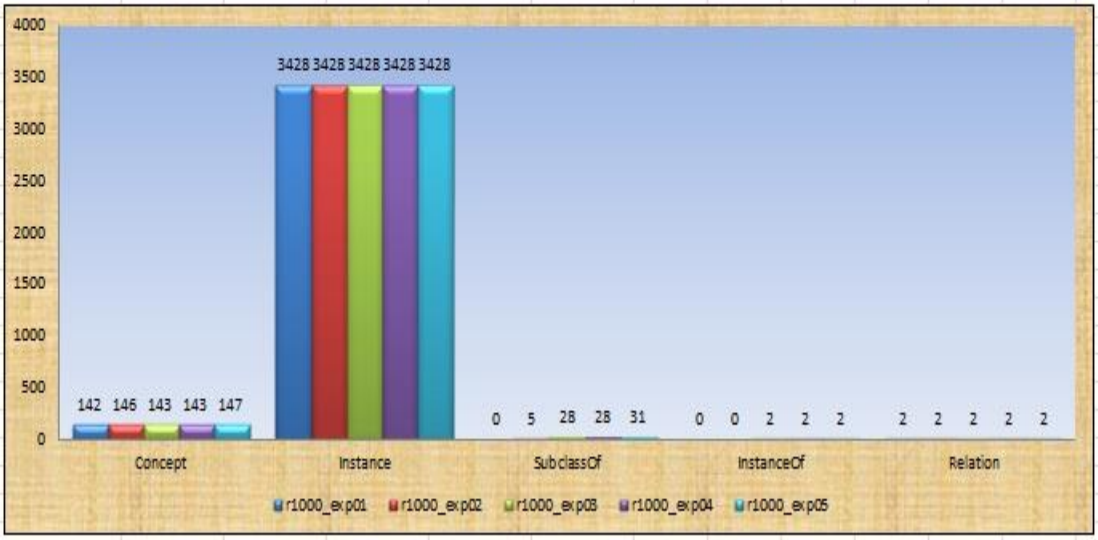






ENUMERATION

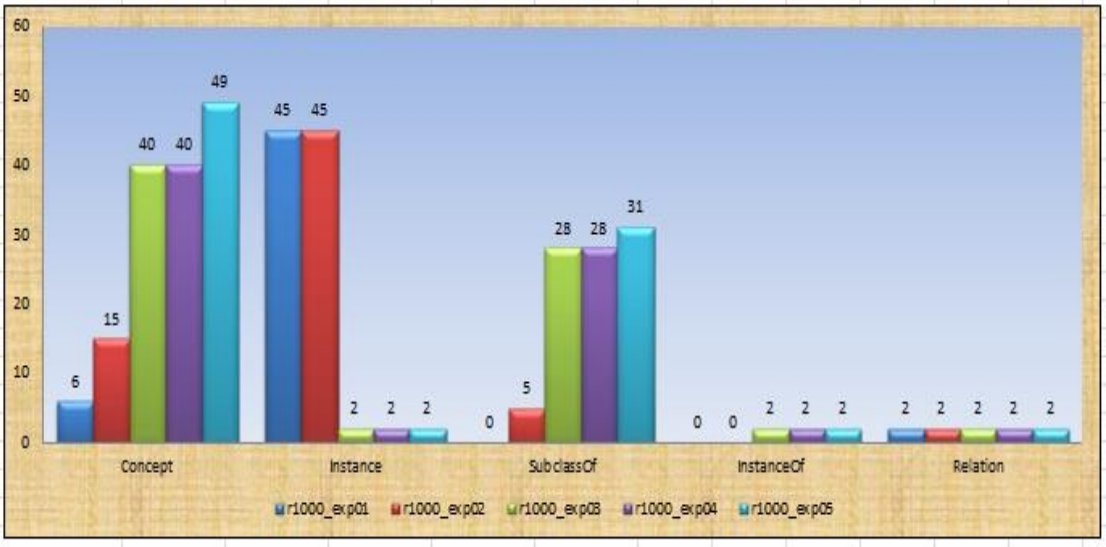
IMDB MOVIES DATA (PER NUMBER OF DATA RECORDS)



ENUMERATION

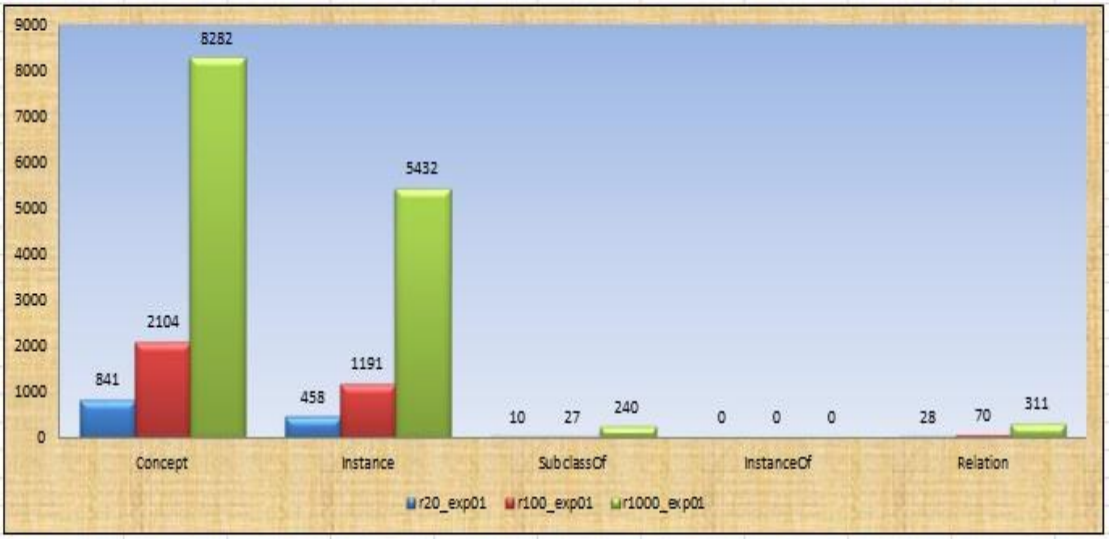
IMDB MOVIES DATA - DELETED PRIMITIVES (PER NUMBER OF DATA RECORDS)

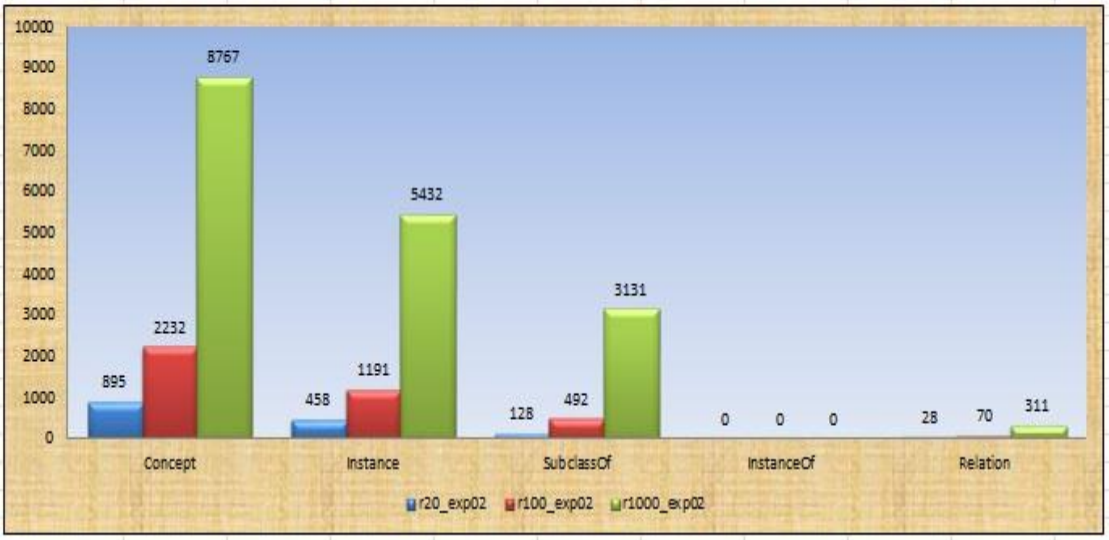
Concept, Instance with rating under 0.001 deleted.

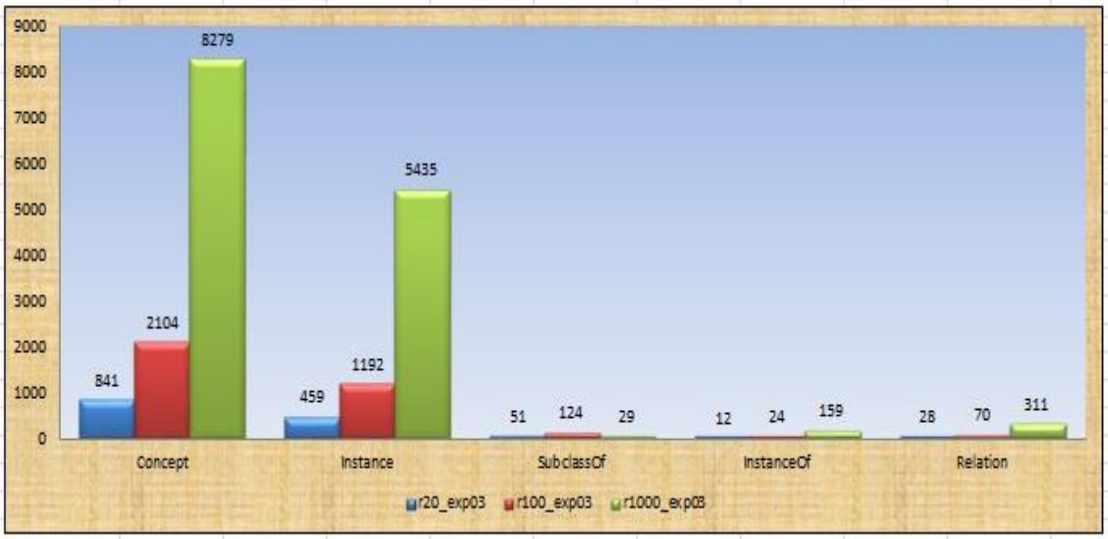


ENUMERATION

AMAZON RANDOM MOVIES REVIEWS (PER EXPERIMENT)

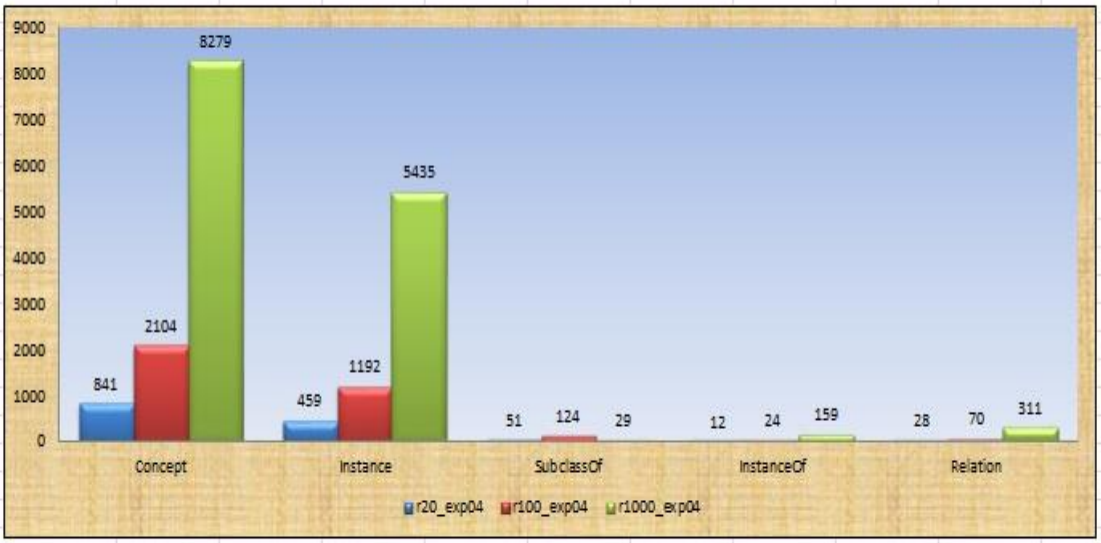


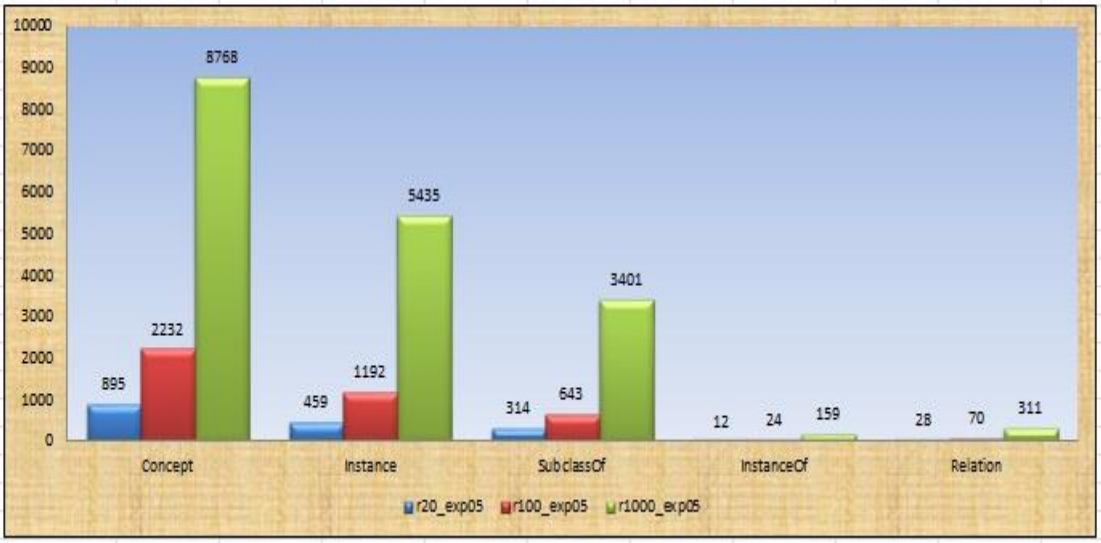




ENUMERATION

AMAZON RANDOM MOVIES REVIEWS (PER EXPERIMENT)

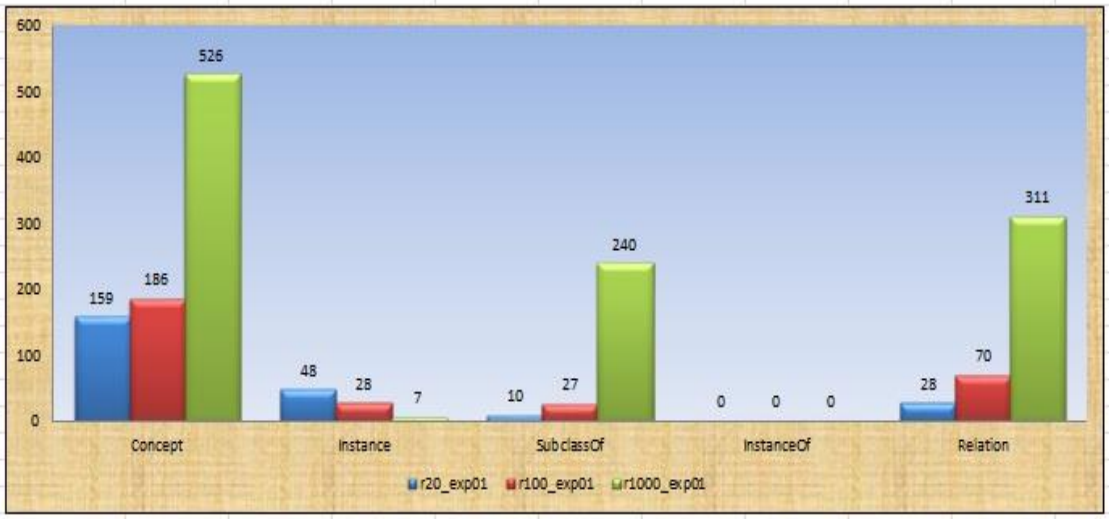


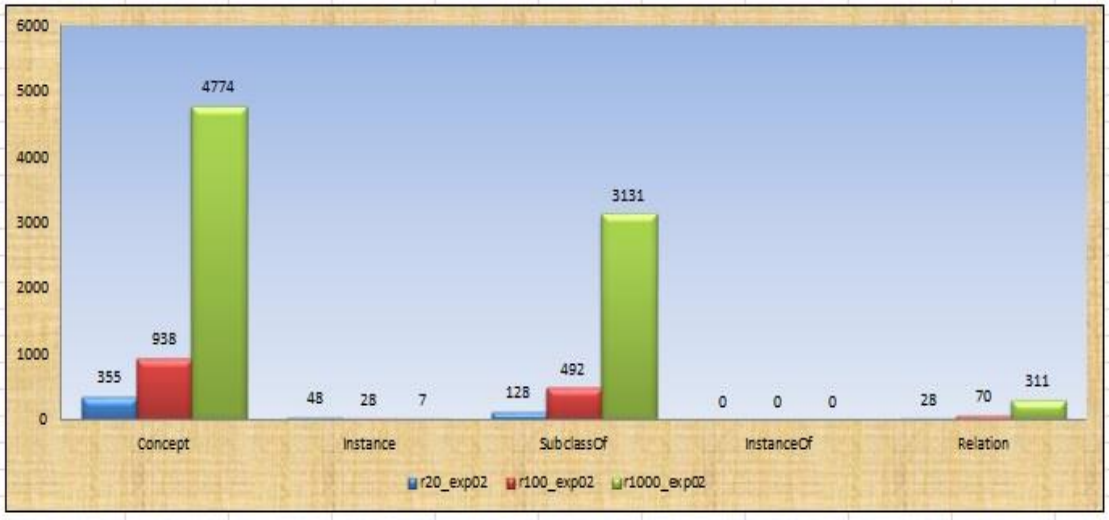


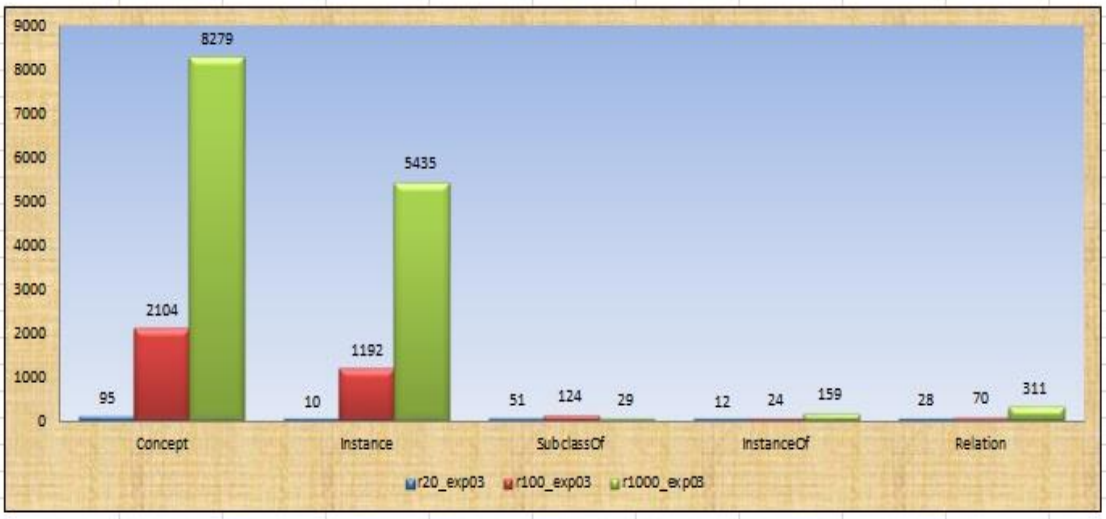
ENUMERATION

AMAZON RANDOM MOVIES REVIEWS - DELETED PRIMITIVES (PER EXPERIMENT)

Concept, Instance with rating under 0.001 deleted.

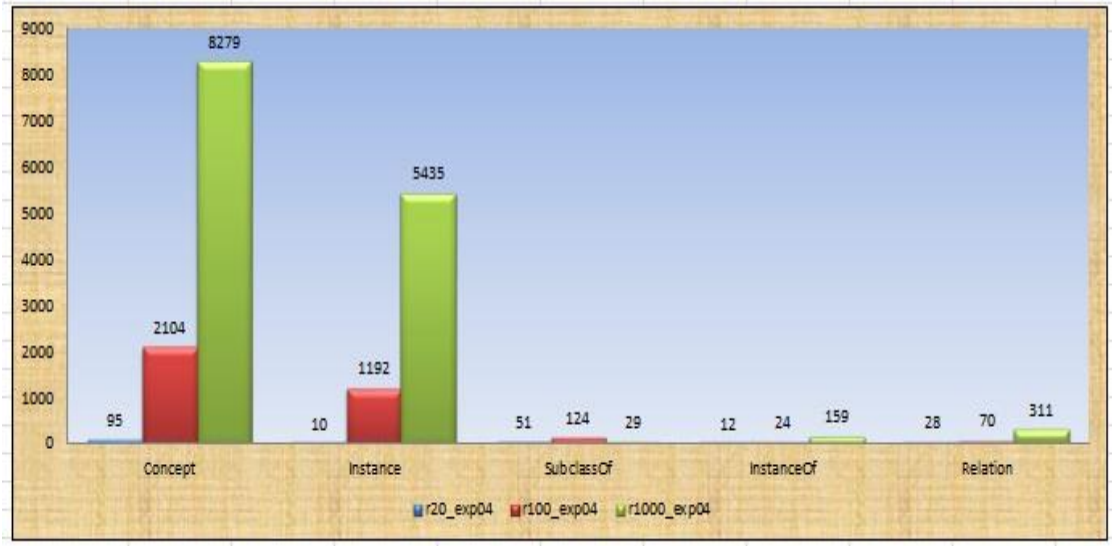


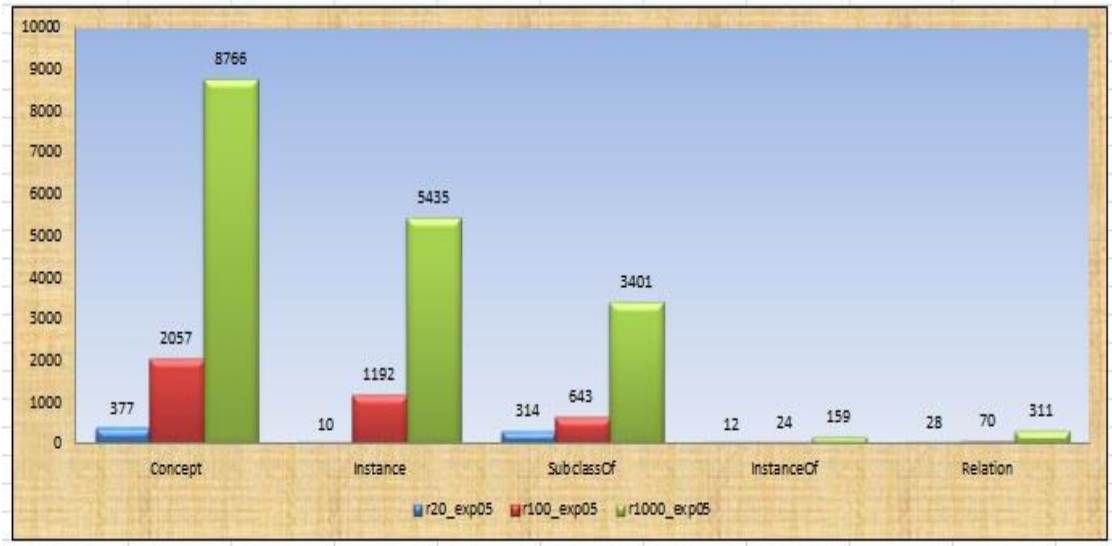




ENUMERATION

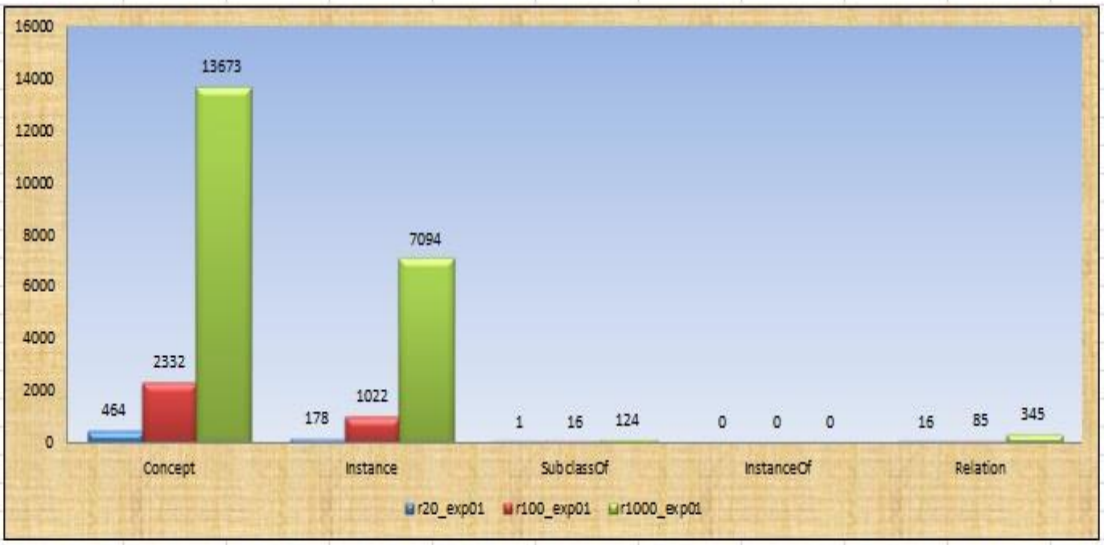
AMAZON RANDOM MOVIES REVIEWS - DELETED PRIMITIVES (PER EXPERIMENT)

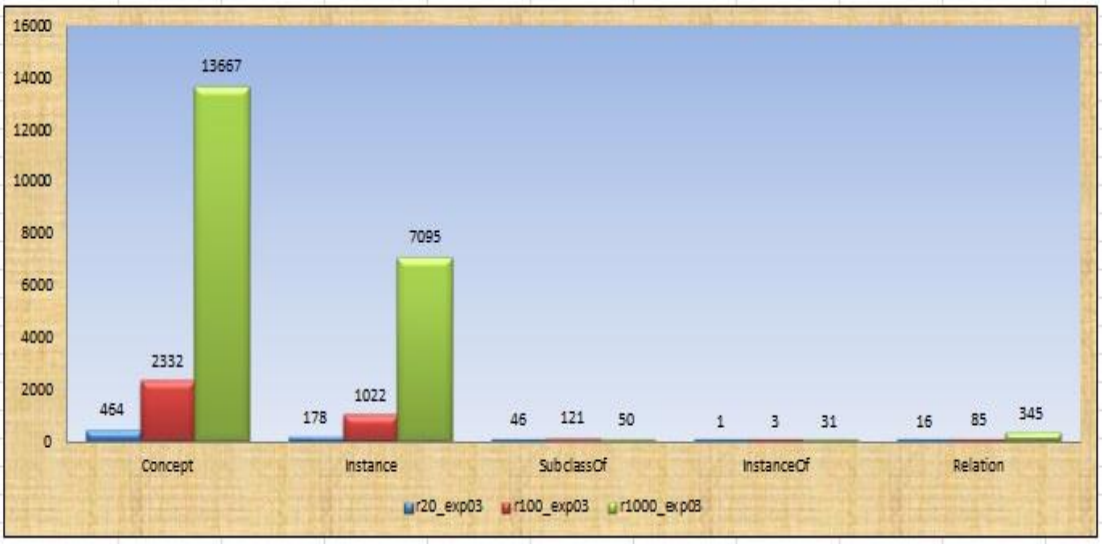




ENUMERATION

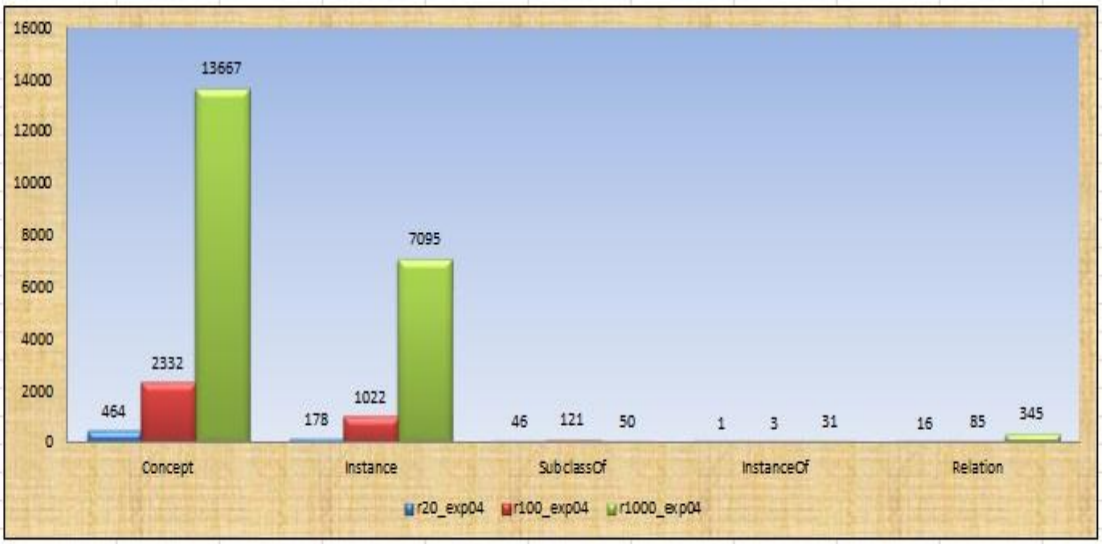
IMDB MOVIES REVIEWS (PER EXPERIMENT)

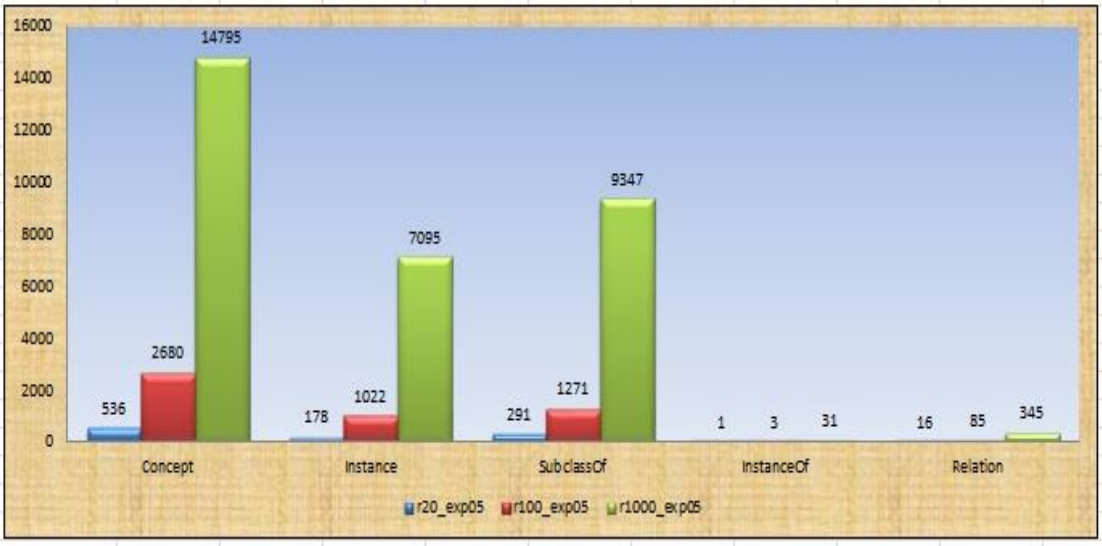




ENUMERATION

IMDB MOVIES REVIEWS (PER EXPERIMENT)

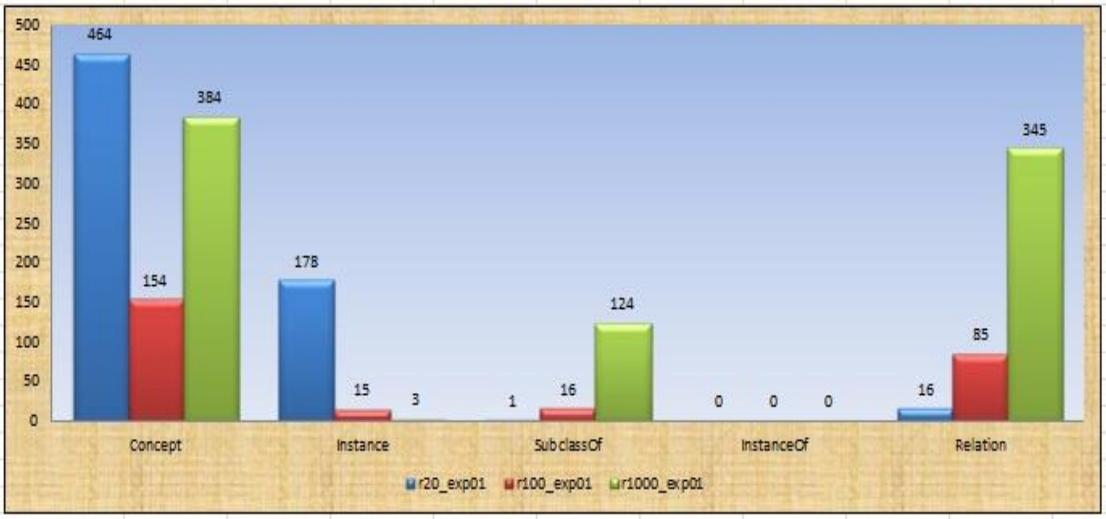


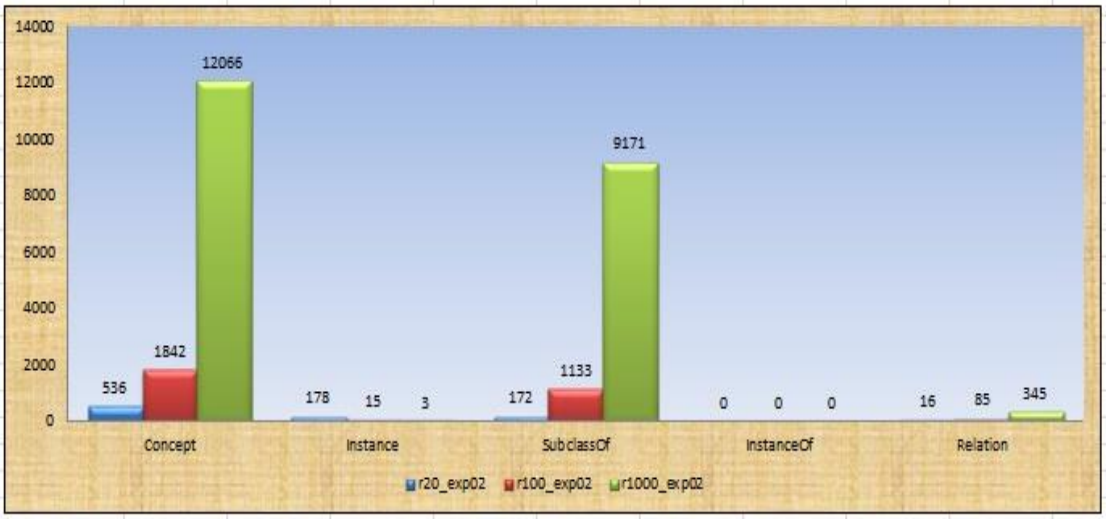


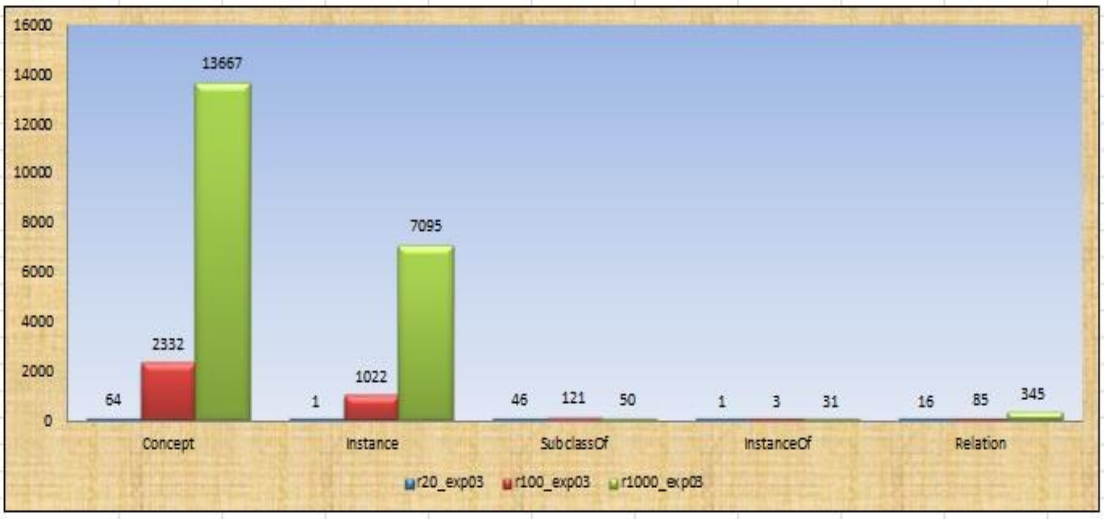
ENUMERATION

IMDB MOVIES REVIEWS - DELETED PRIMITIVES (PER EXPERIMENT)

Concept, Instance with rating under 0.001 deleted.

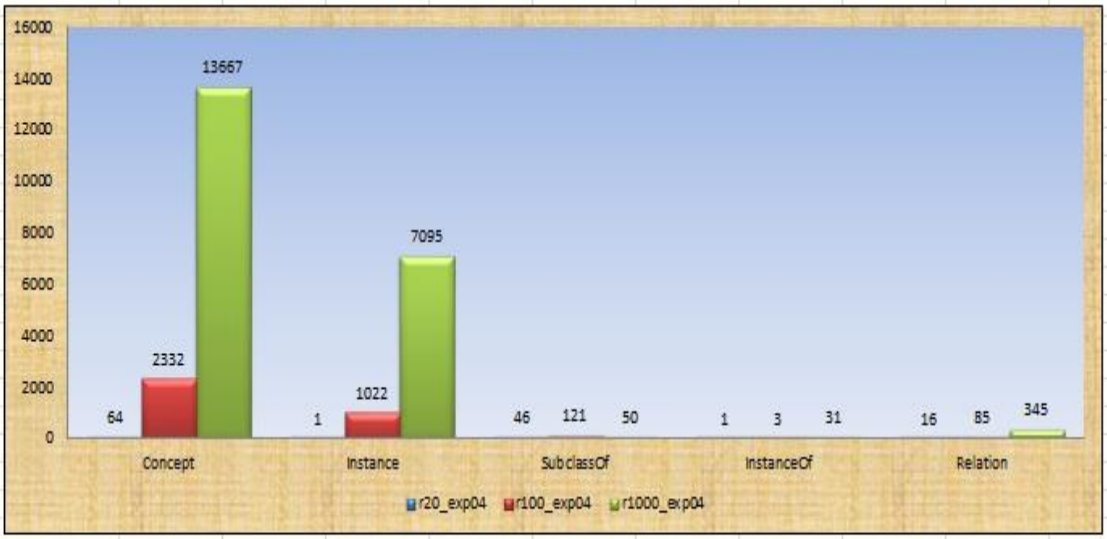






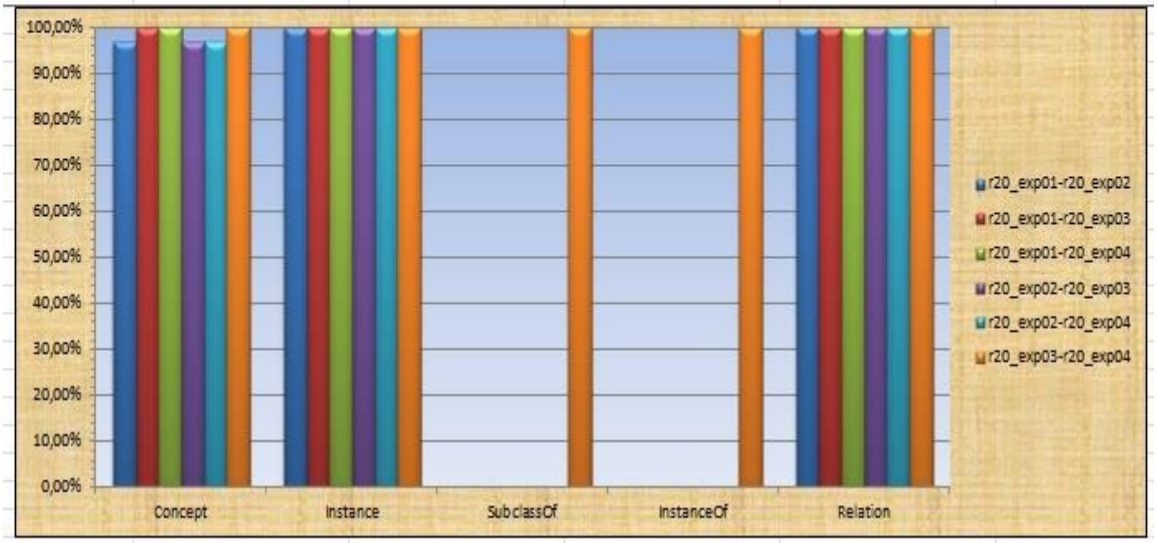
ENUMERATION

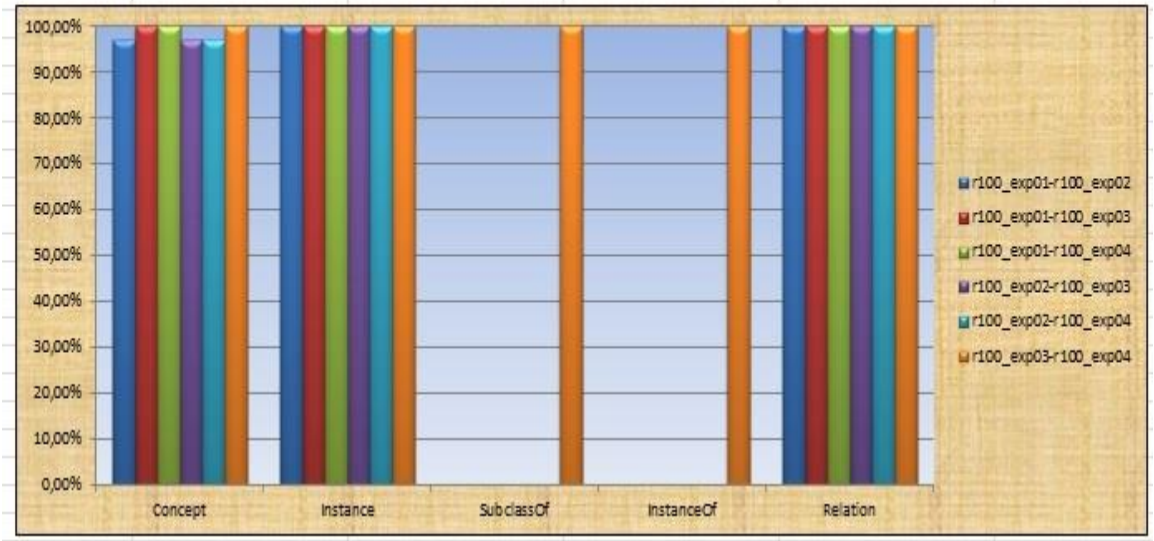
IMDB MOVIES REVIEWS - DELETED PRIMITIVES (PER EXPERIMENT)

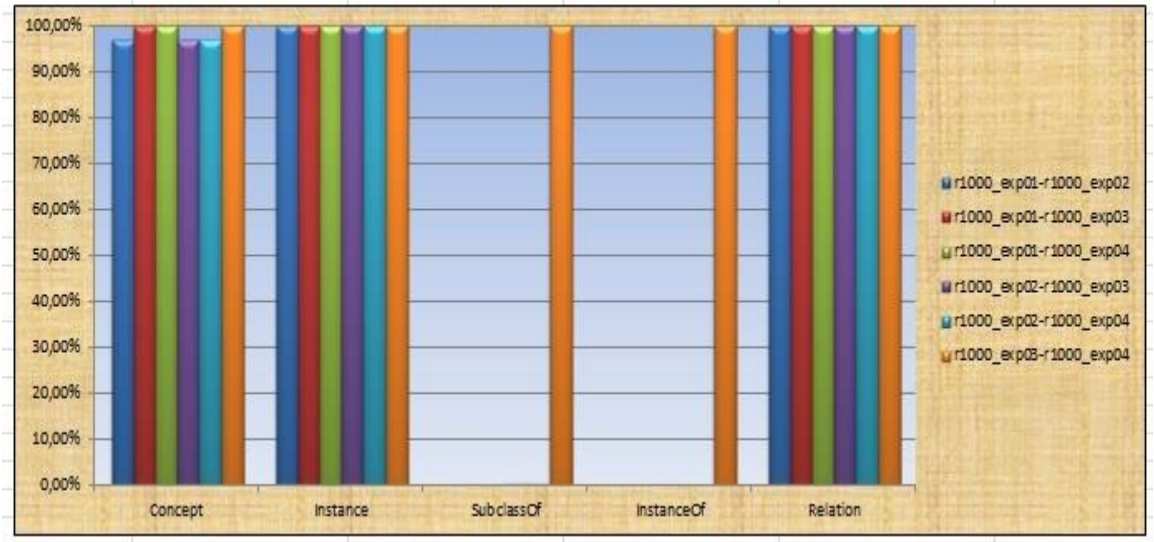


SAME PRIMITIVES COMPARISON - PERCENTAGE RATIO

AMAZON RANDOM MOVIES REVIEWS (PER NUMBER OF REVIEWS)



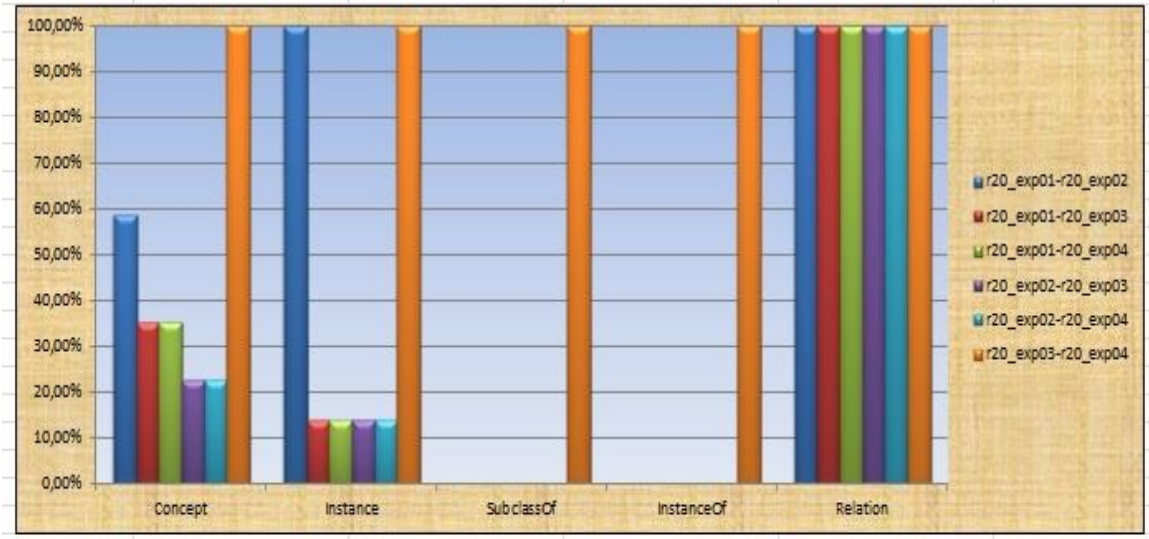


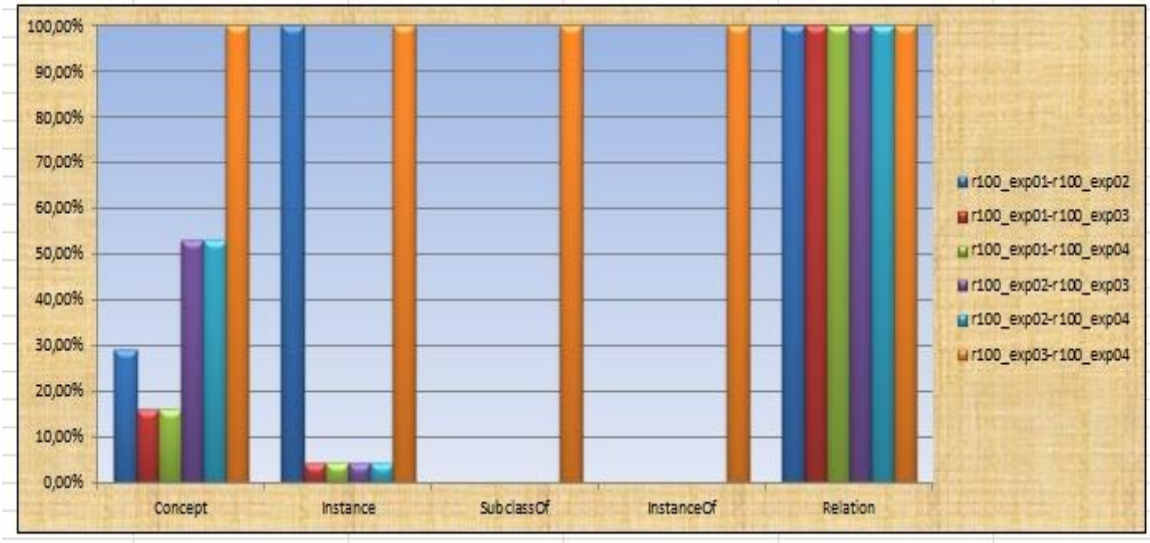


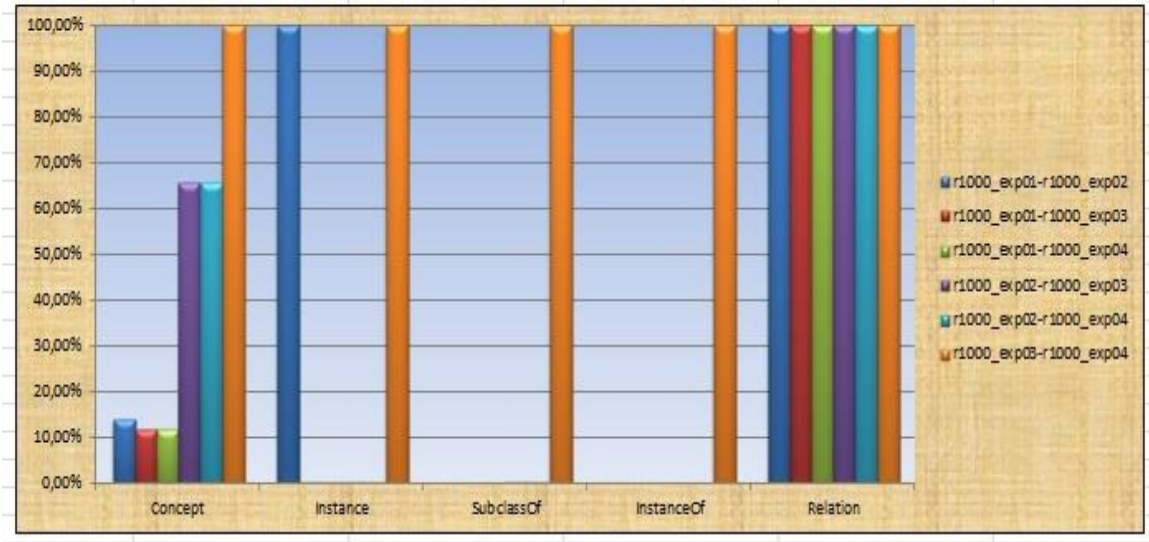
SAME PRIMITIVES COMPARISON - PERCENTAGE RATIO

AMAZON RANDOM MOVIES REVIEWS - DELETED PRIMITIVES (PER NUMBER OF REVIEWS)

Concept, Instance with rating under 0.001 deleted.

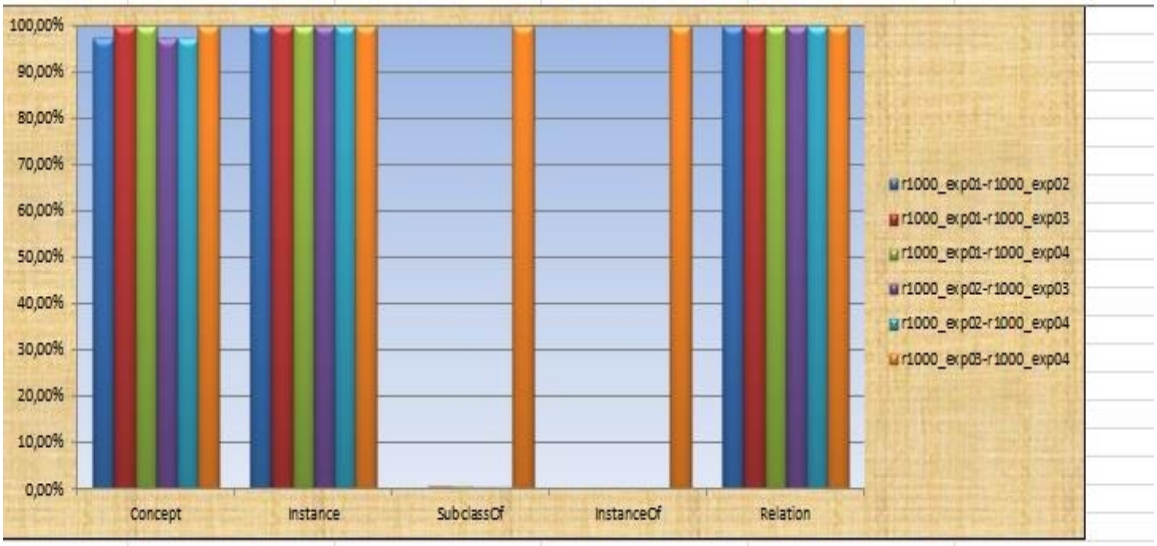






SAME PRIMITIVES COMPARISON - PERCENTAGE RATIO

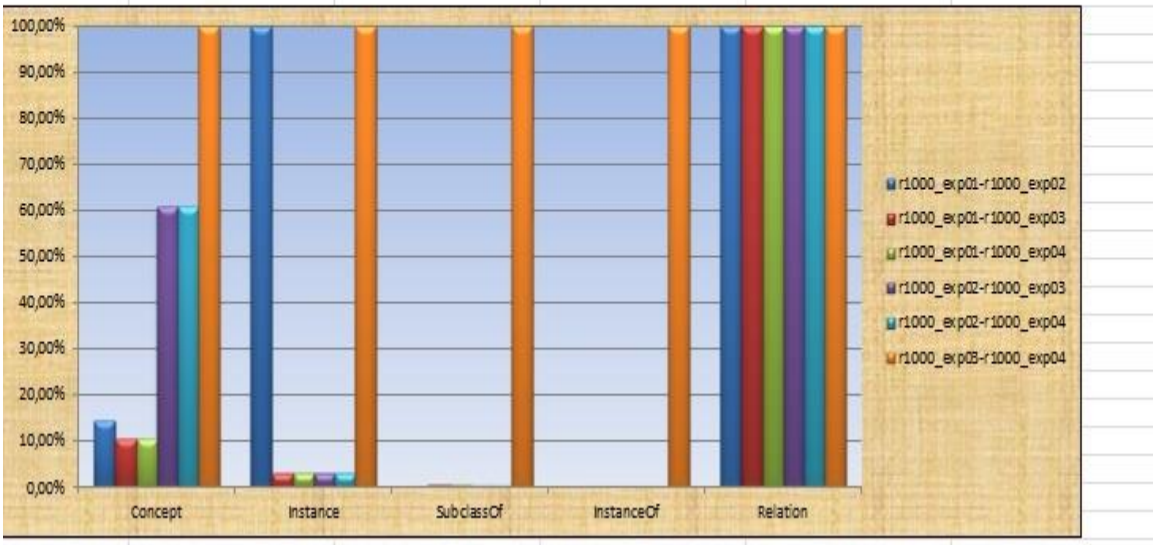
AMAZON SAME MOVIES REVIEWS (PER NUMBER OF REVIEWS)



SAME PRIMITIVES COMPARISON - PERCENTAGE RATIO

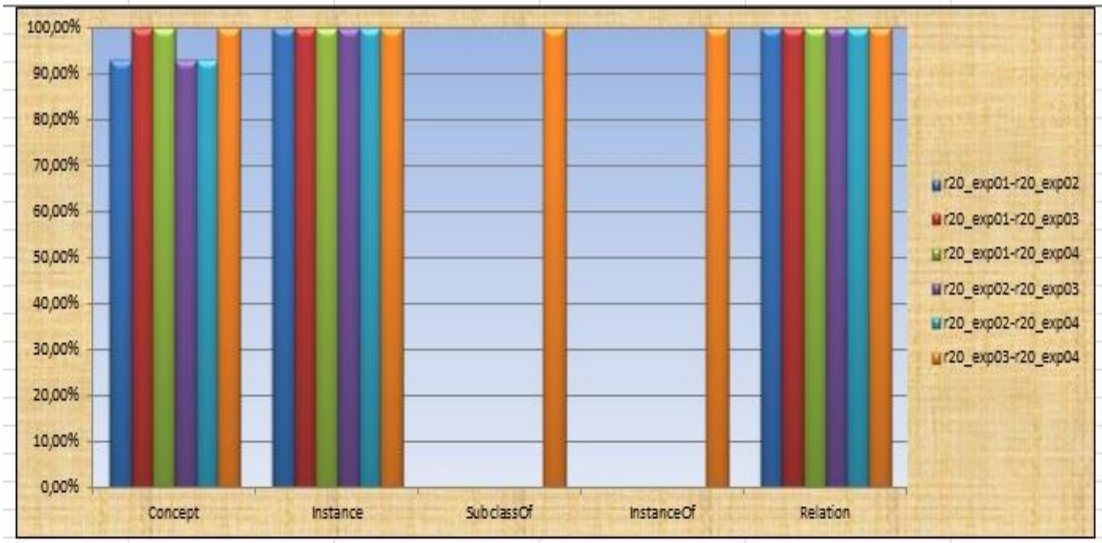
AMAZON SAME MOVIES REVIEWS - DELETED PRIMITIVES (PER NUMBER OF REVIEWS)

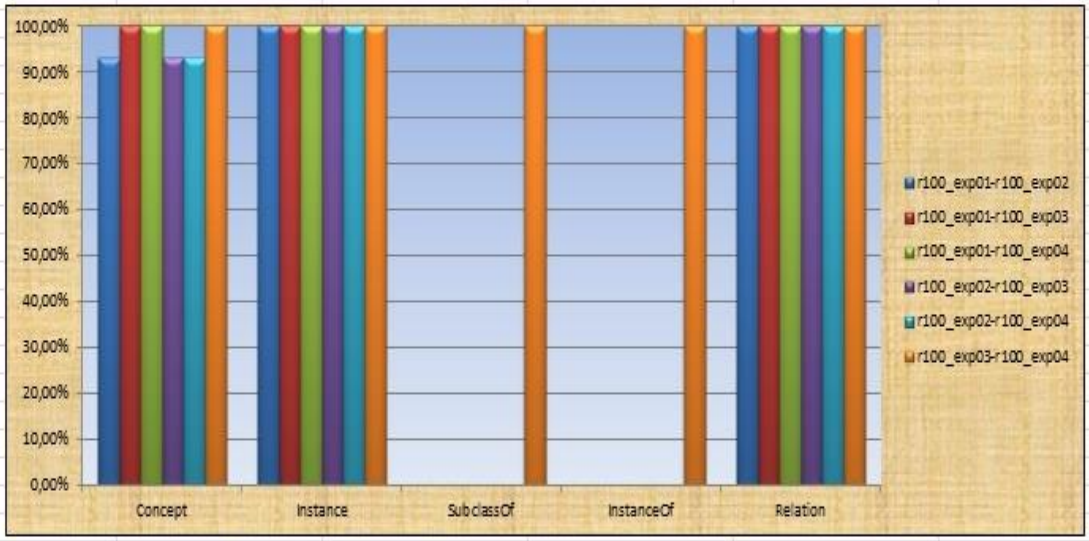
Concept, Instance with rating under 0.001 deleted.

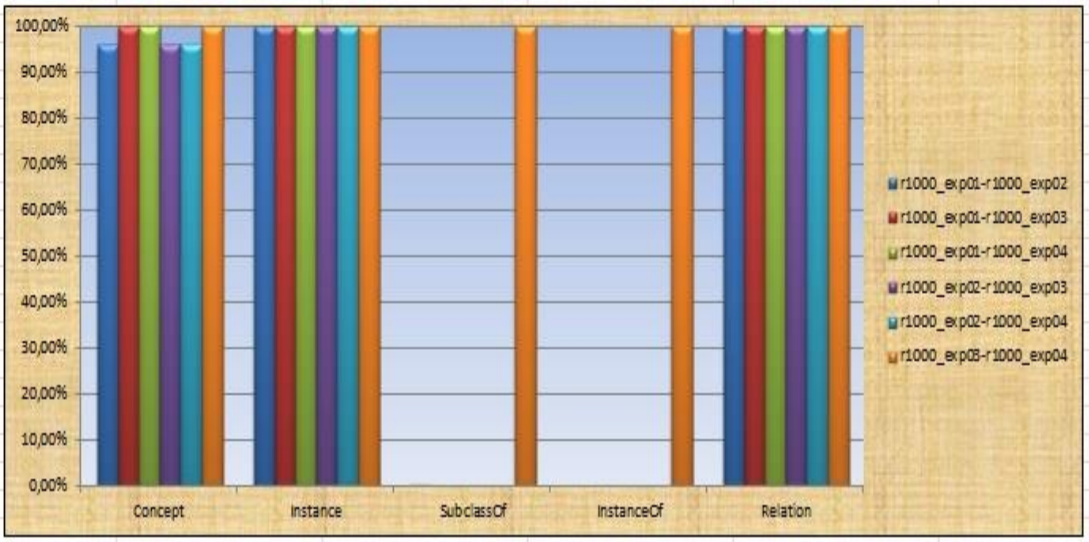


SAME PRIMITIVES COMPARISON - PERCENTAGE RATIO

IMDB MOVIES REVIEWS (PER NUMBER OF REVIEWS)



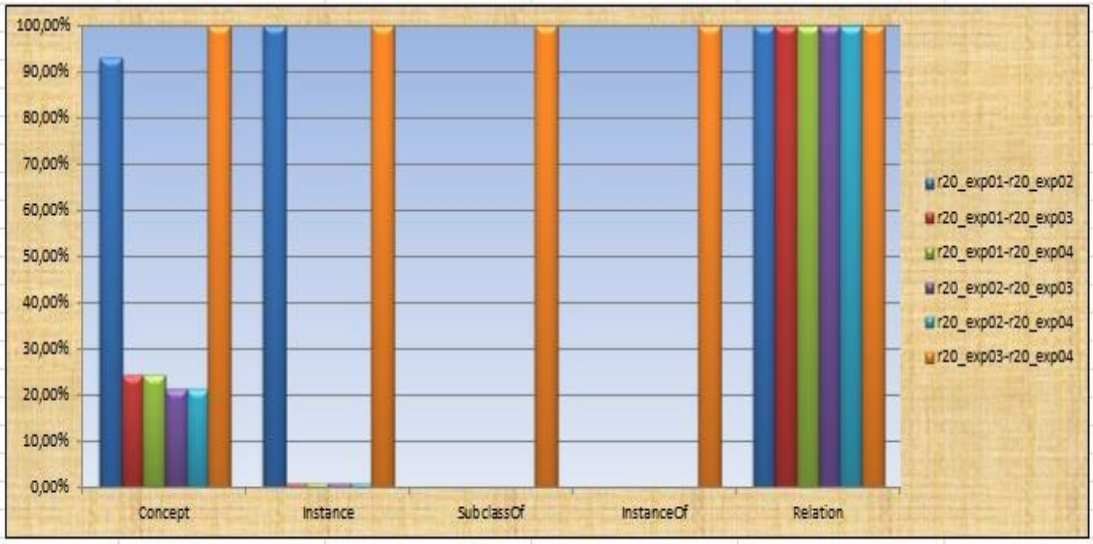


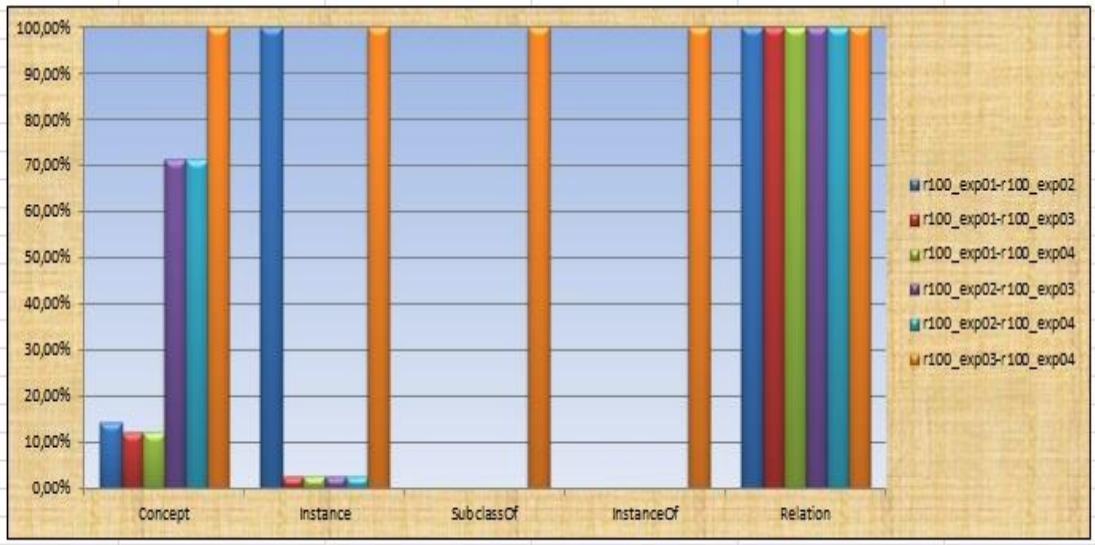


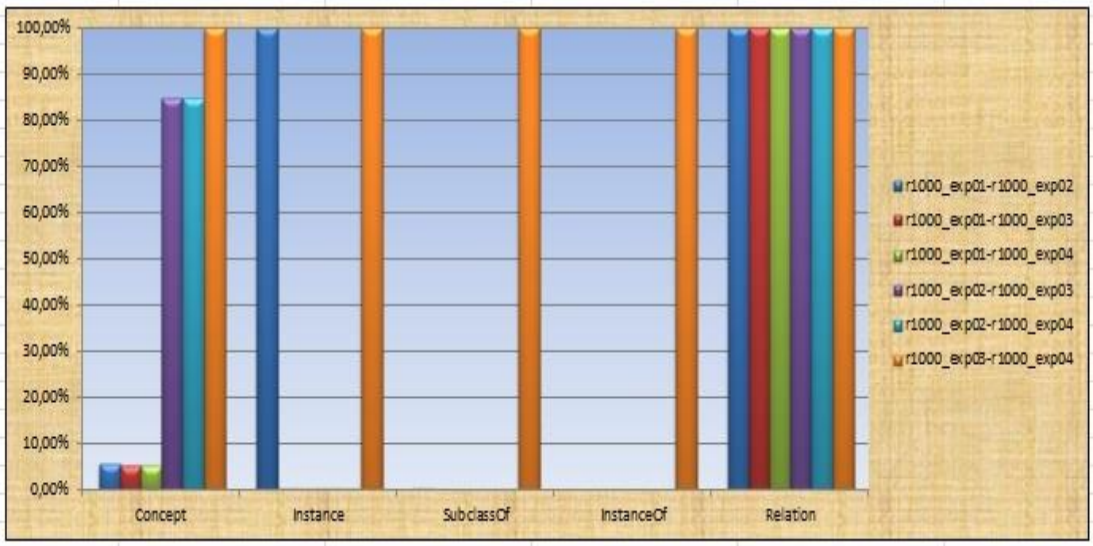
SAME PRIMITIVES COMPARISON - PERCENTAGE RATIO

IMDB MOVIES REVIEWS - DELETED PRIMITIVES (PER NUMBER OF REVIEWS)

Concept, Instance with rating under 0.001 deleted.

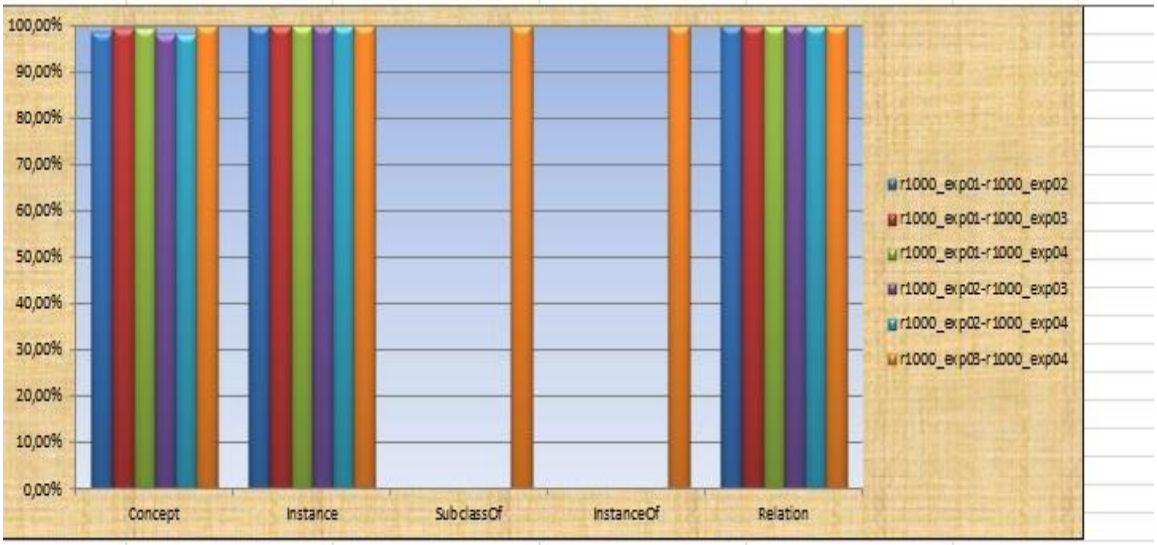






SAME PRIMITIVES COMPARISON - PERCENTAGE RATIO

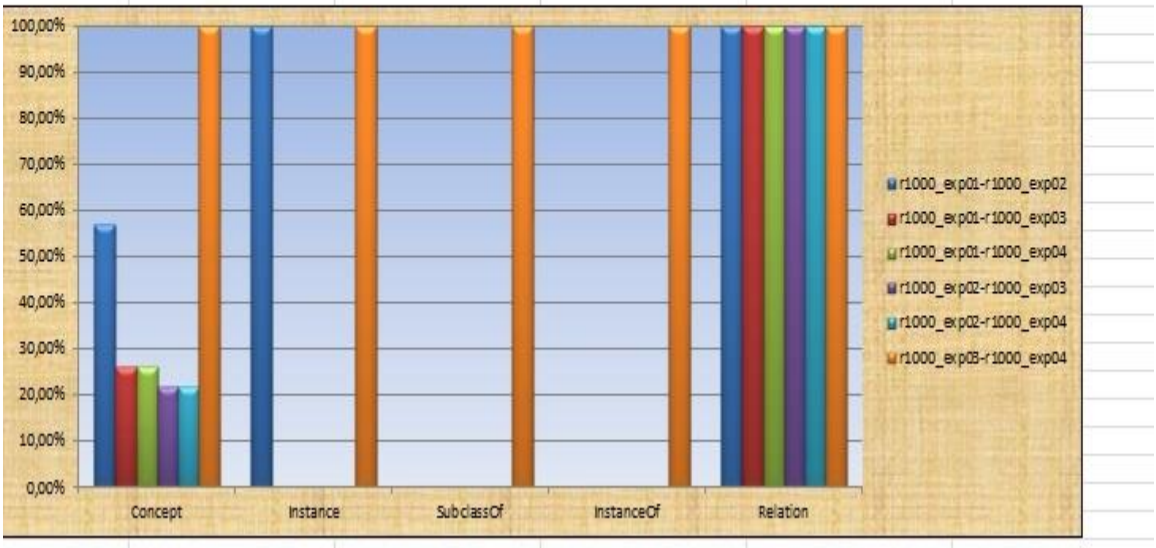
IMDB MOVIES DATA (PER NUMBER OF DATA RECORDS)



SAME PRIMITIVES COMPARISON - PERCENTAGE RATIO

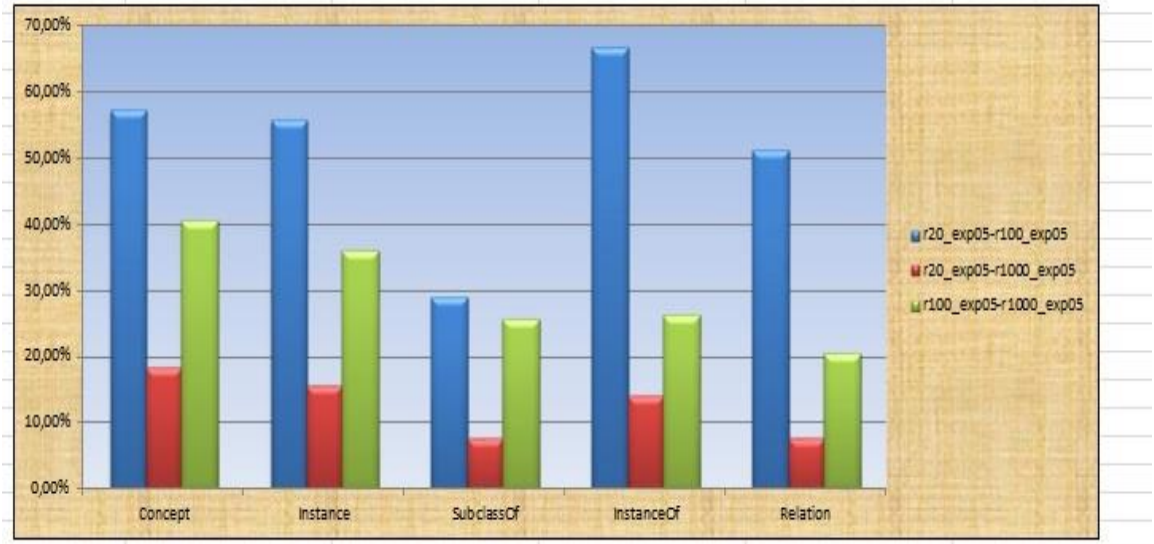
IMDB MOVIES DATA - DELETED PRIMITIVES (PER NUMBER OF DATA RECORDS)

Concept, Instance with rating under 0.001 deleted.



SAME PRIMITIVES COMPARISON - PERCENTAGE RATIO

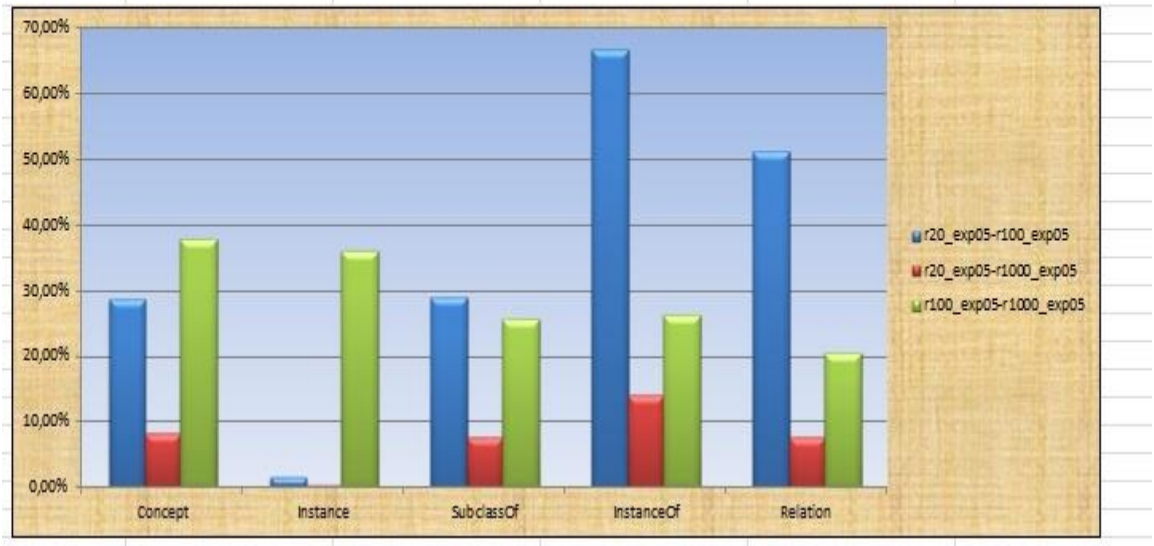
AMAZON RANDOM MOVIES REVIEWS (PER EXPERIMENT)



SAME PRIMITIVES COMPARISON - PERCENTAGE RATIO

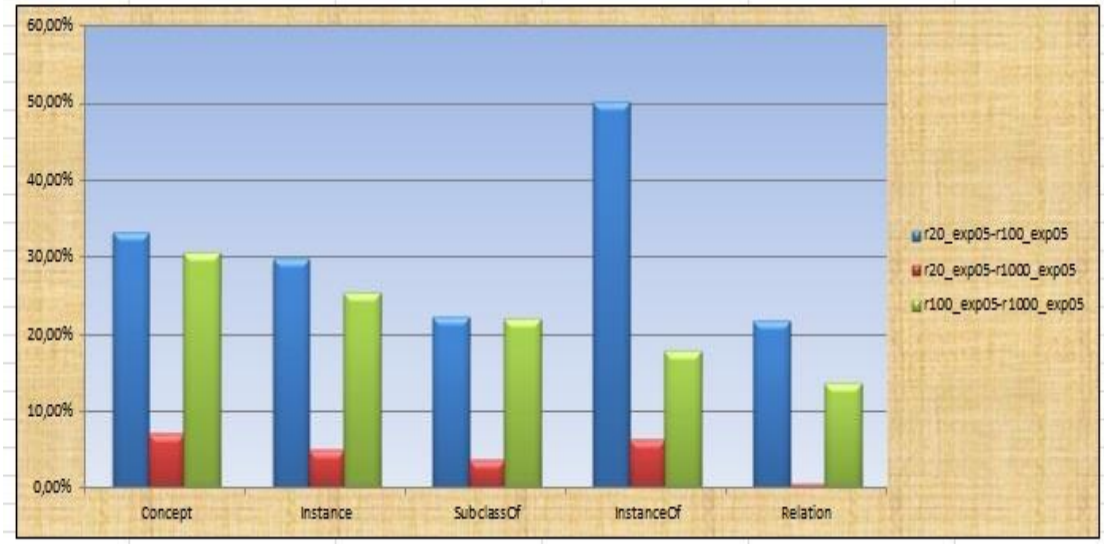
AMAZON RANDOM MOVIES REVIEWS - DELETED PRIMITIVES (PER EXPERIMENT)

Concept, Instance with rating under 0.001 deleted.



SAME PRIMITIVES COMPARISON - PERCENTAGE RATIO

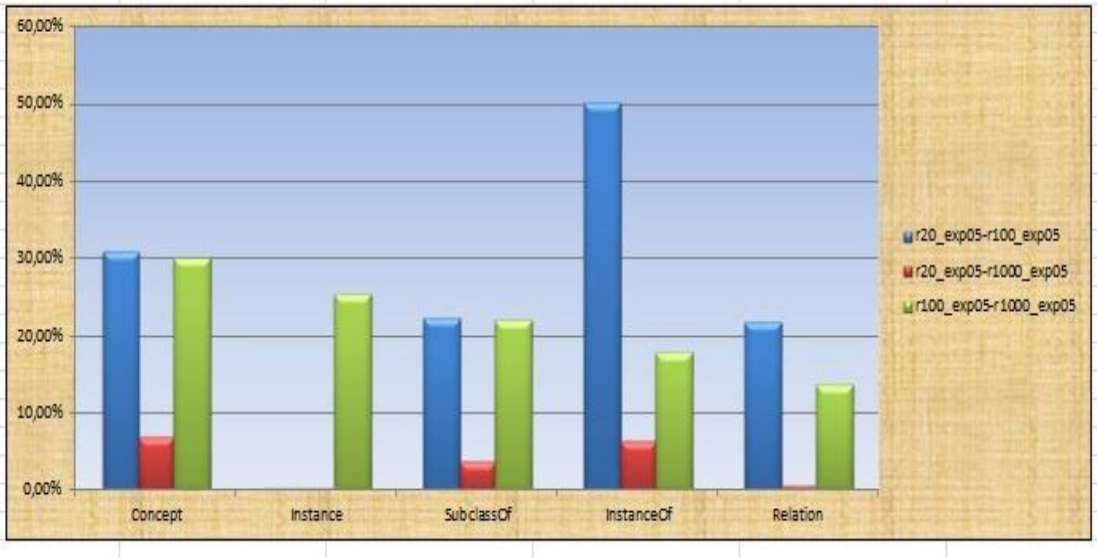
IMDB MOVIES REVIEWS - (PER EXPERIMENT)



SAME PRIMITIVES COMPARISON - PERCENTAGE RATIO

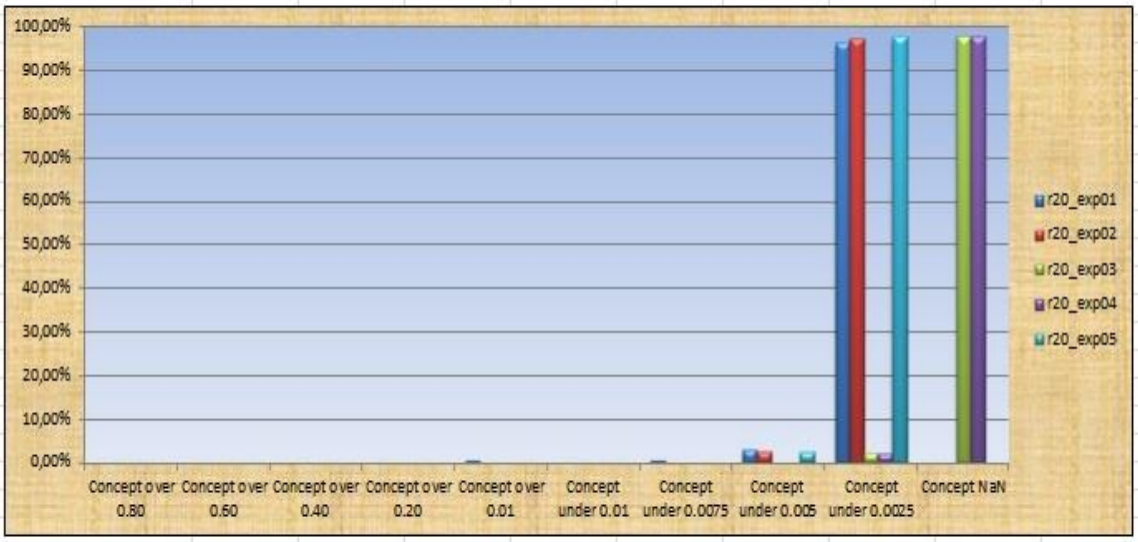
IMDB MOVIES REVIEWS - DELETED PRIMITIVES (PER EXPERIMENT)

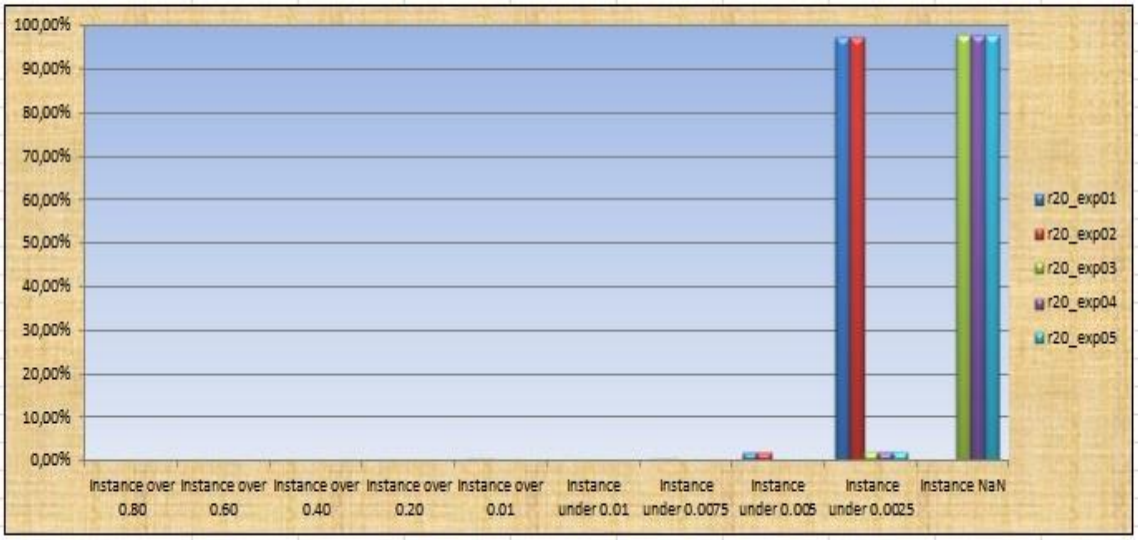
Concept, Instance with rating under 0.001 deleted.

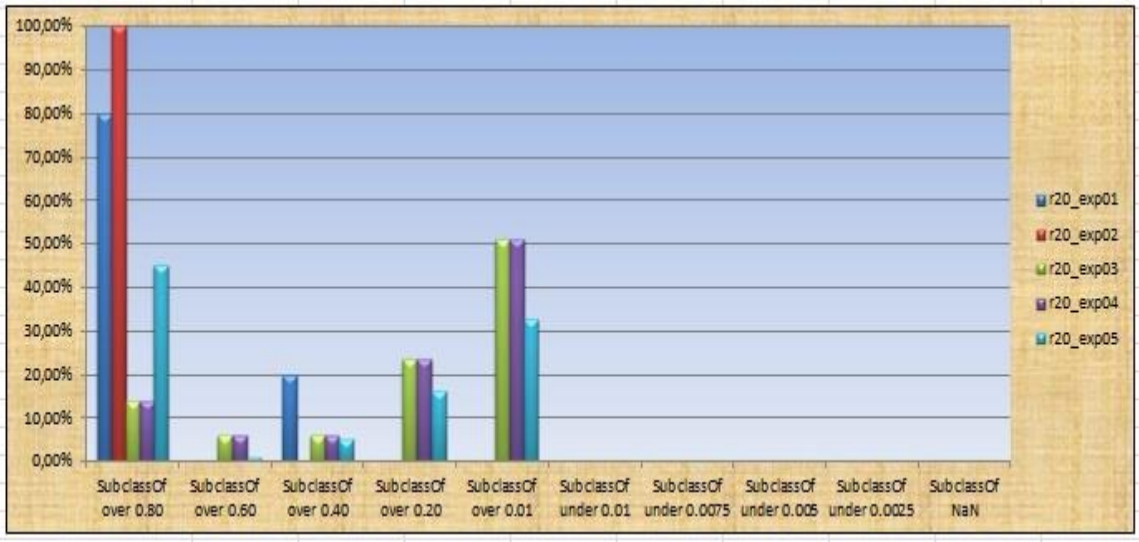


QUANTITATIVE GROUPING AT PROBABILITY LEVELS - PERCENTAGE RATIO

AMAZON RANDOM MOVIES REVIEWS (PER NUMBER OF REVIEWS)

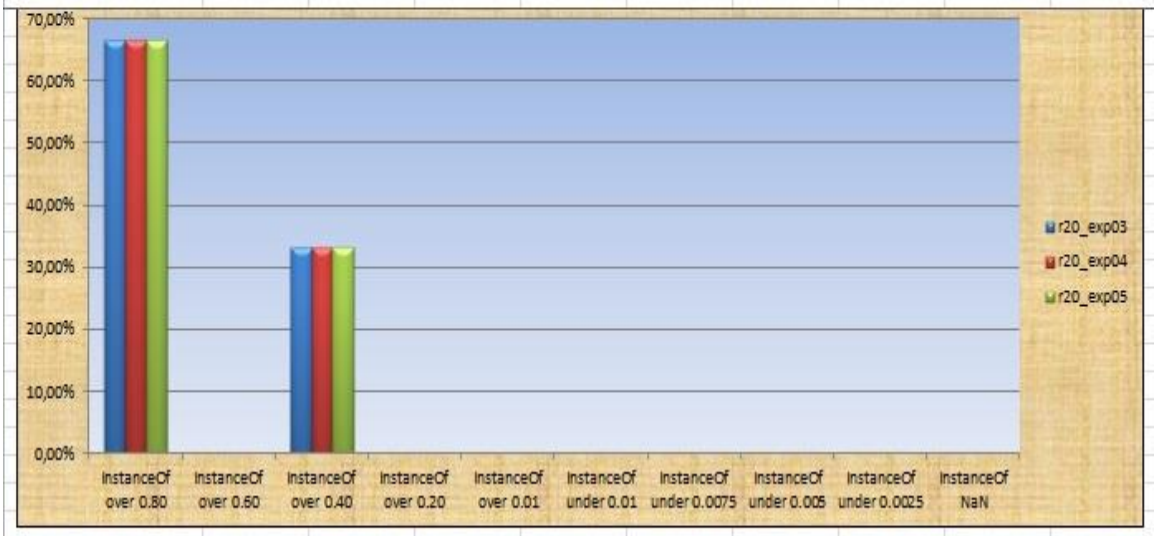


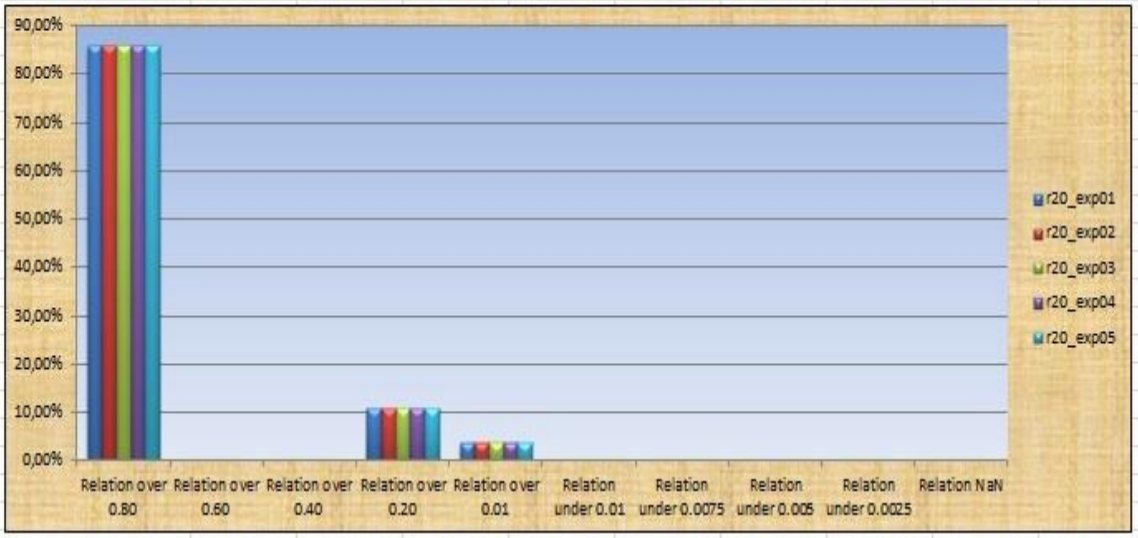


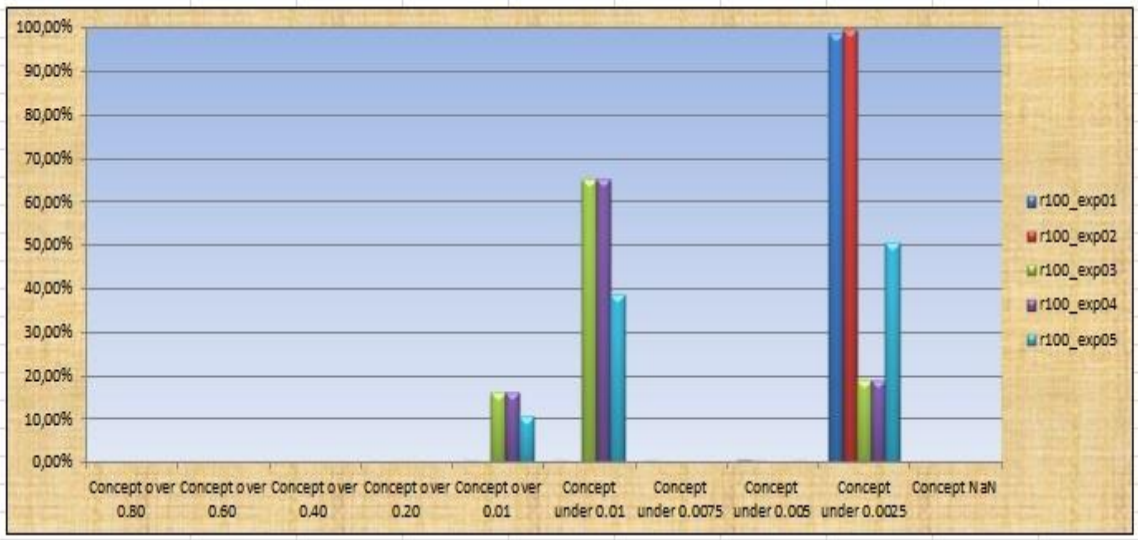


QUANTITATIVE GROUPING AT PROBABILITY LEVELS - PERCENTAGE RATIO

AMAZON RANDOM MOVIES REVIEWS (PER NUMBER OF REVIEWS)

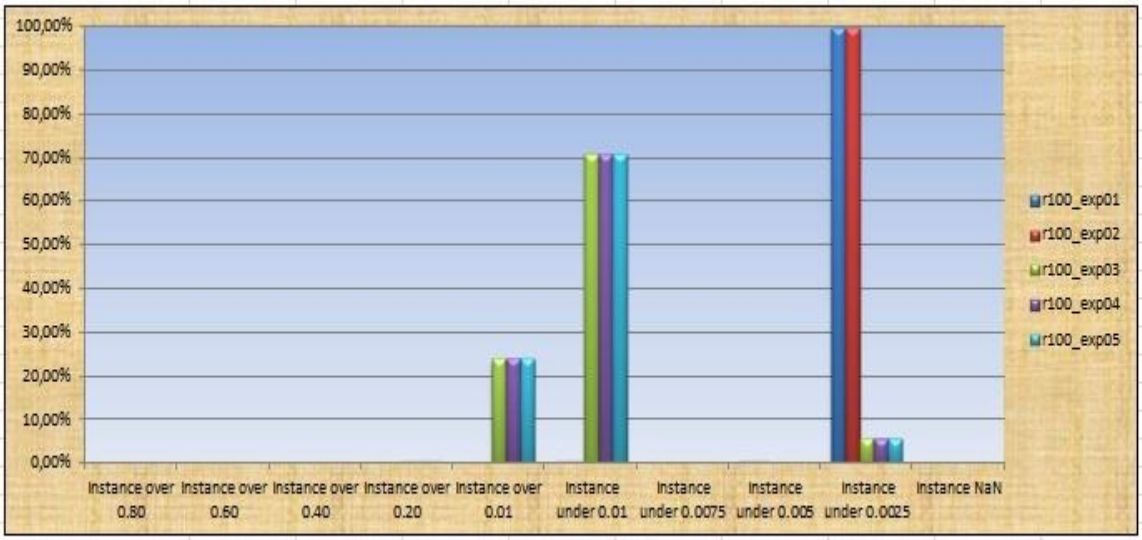


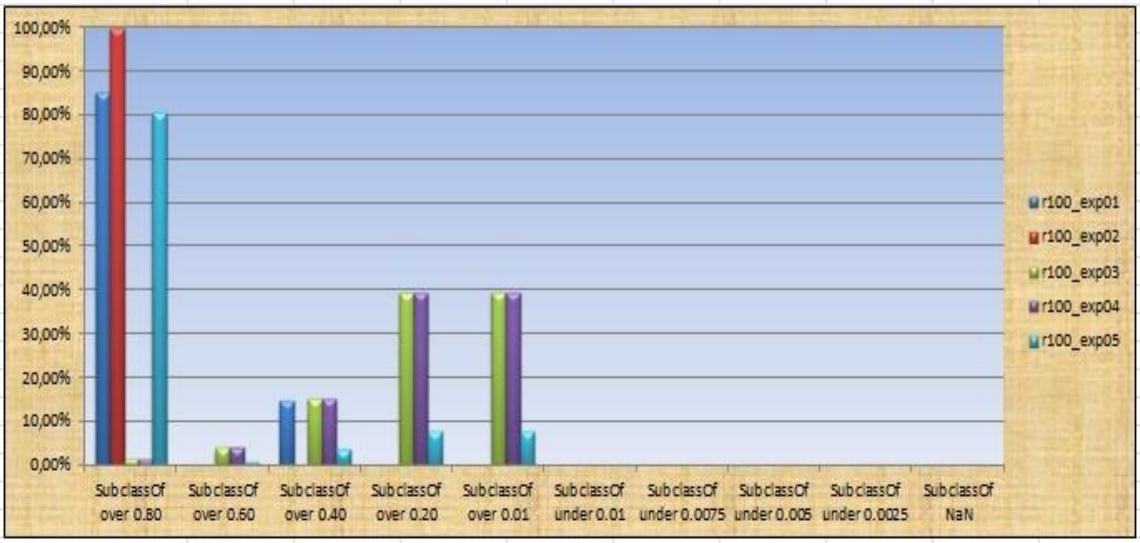


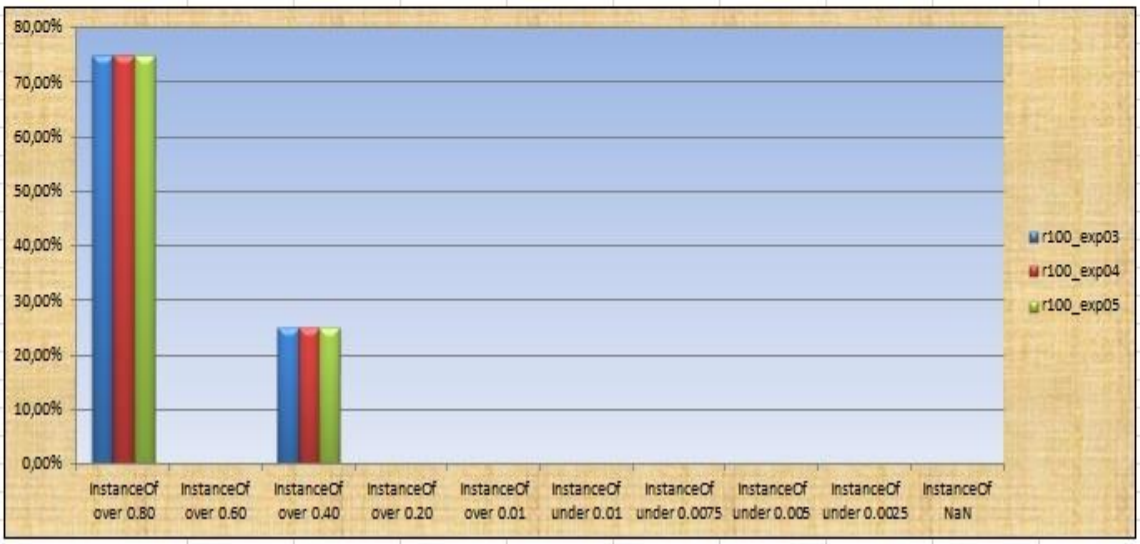


QUANTITATIVE GROUPING AT PROBABILITY LEVELS - PERCENTAGE RATIO

AMAZON RANDOM MOVIES REVIEWS (PER NUMBER OF REVIEWS)

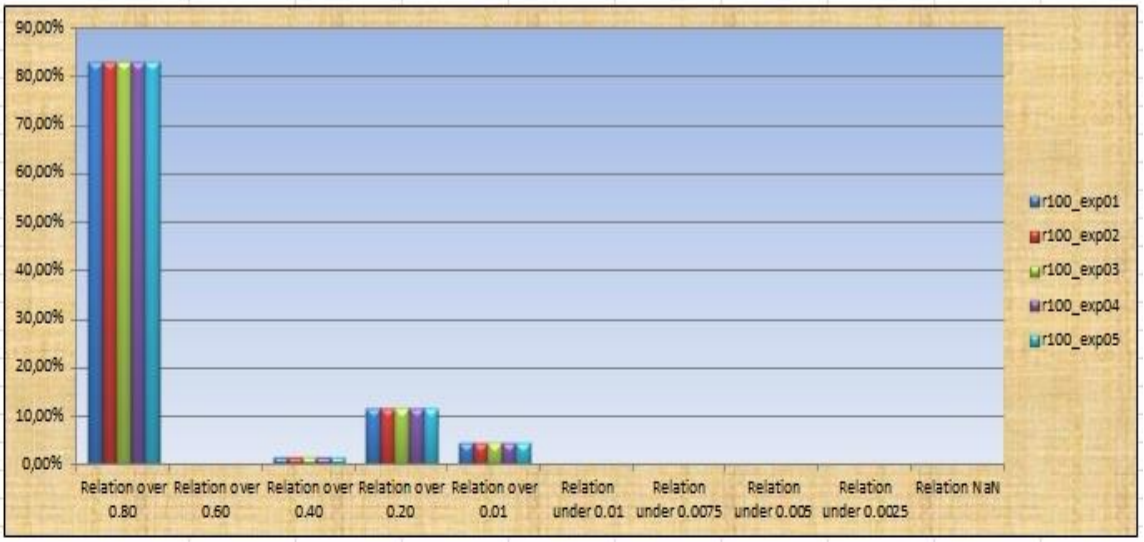


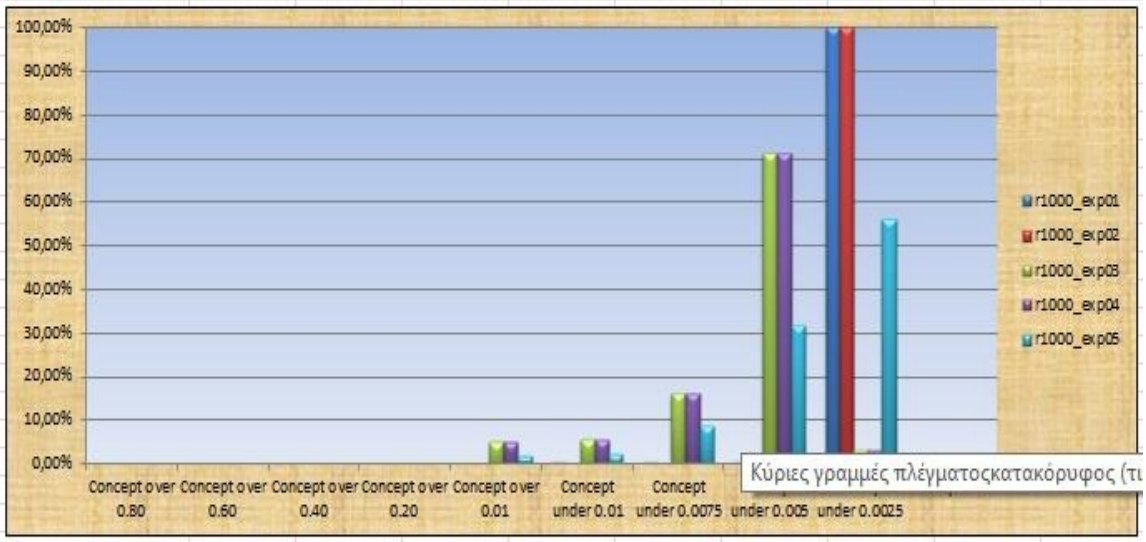


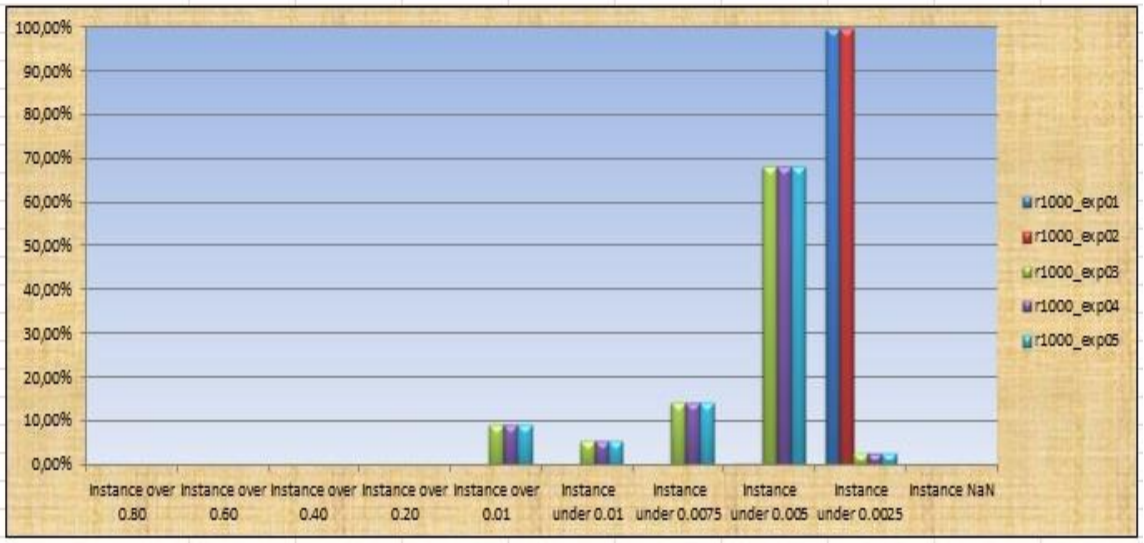


QUANTITATIVE GROUPING AT PROBABILITY LEVELS - PERCENTAGE RATIO

AMAZON RANDOM MOVIES REVIEWS (PER NUMBER OF REVIEWS)

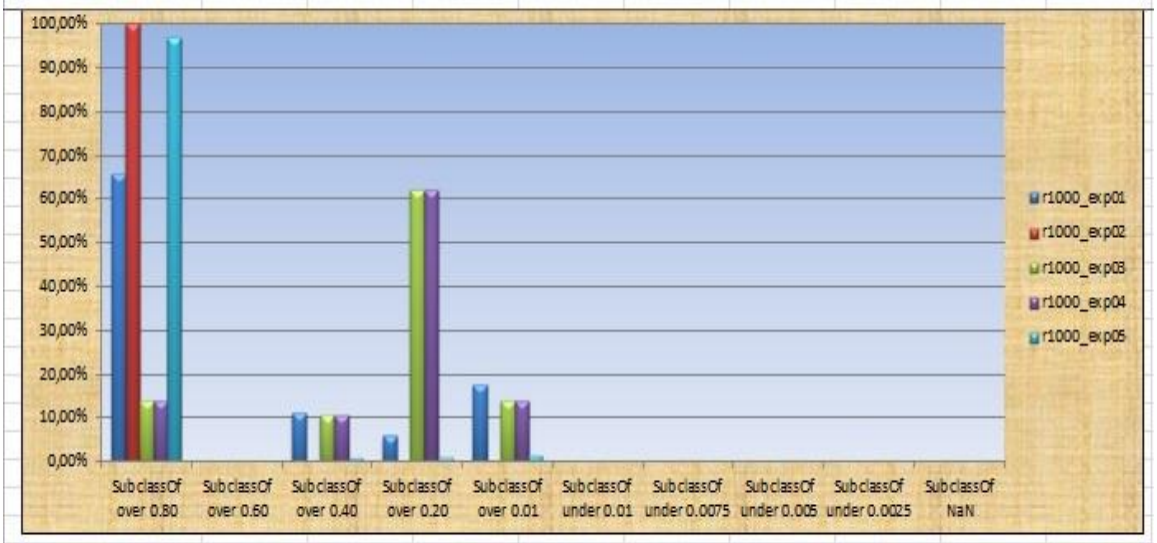


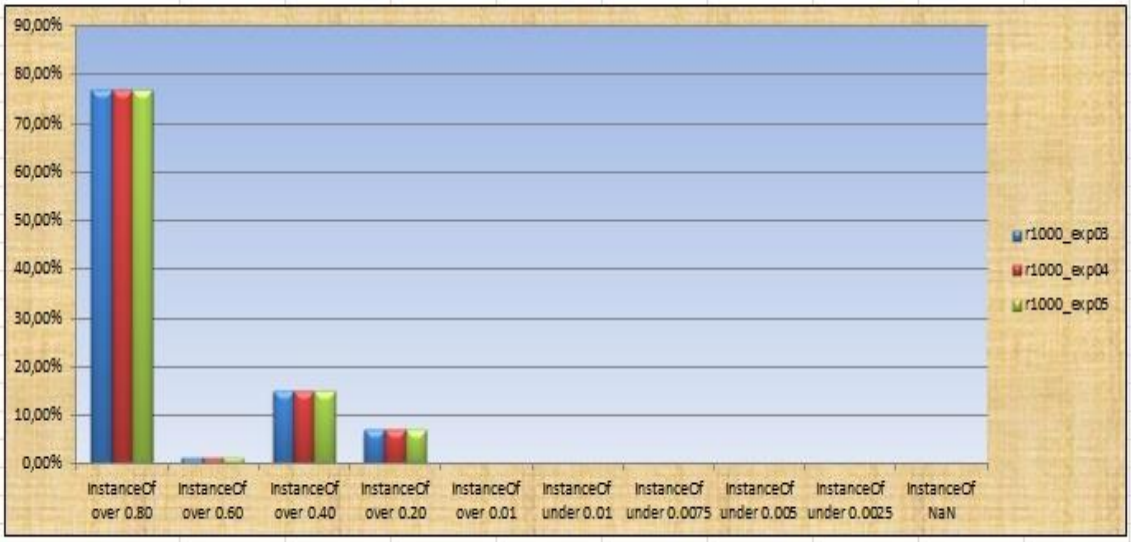


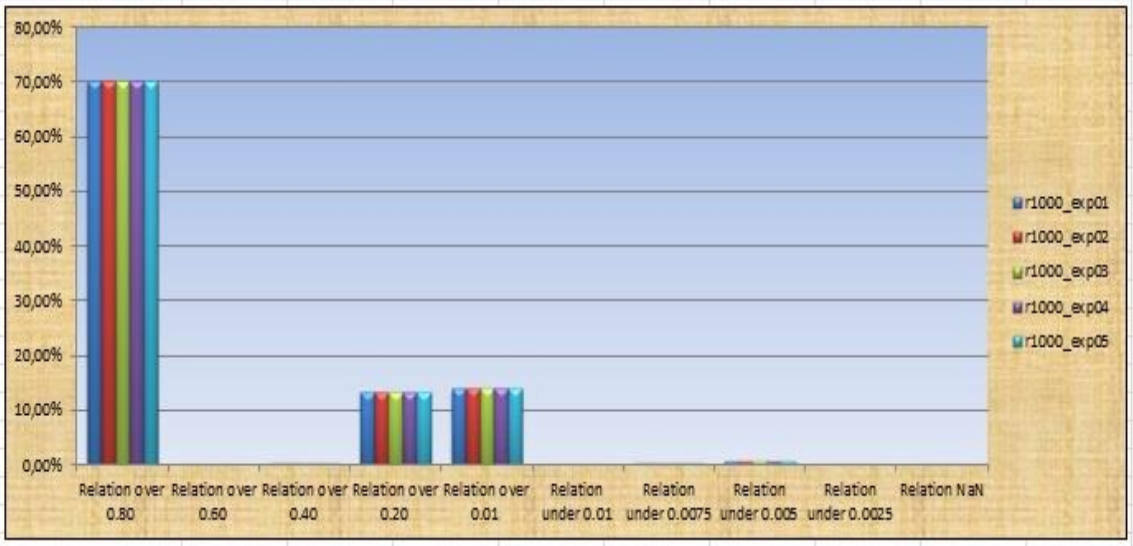


QUANTITATIVE GROUPING AT PROBABILITY LEVELS - PERCENTAGE RATIO

AMAZON RANDOM MOVIES REVIEWS (PER NUMBER OF REVIEWS)

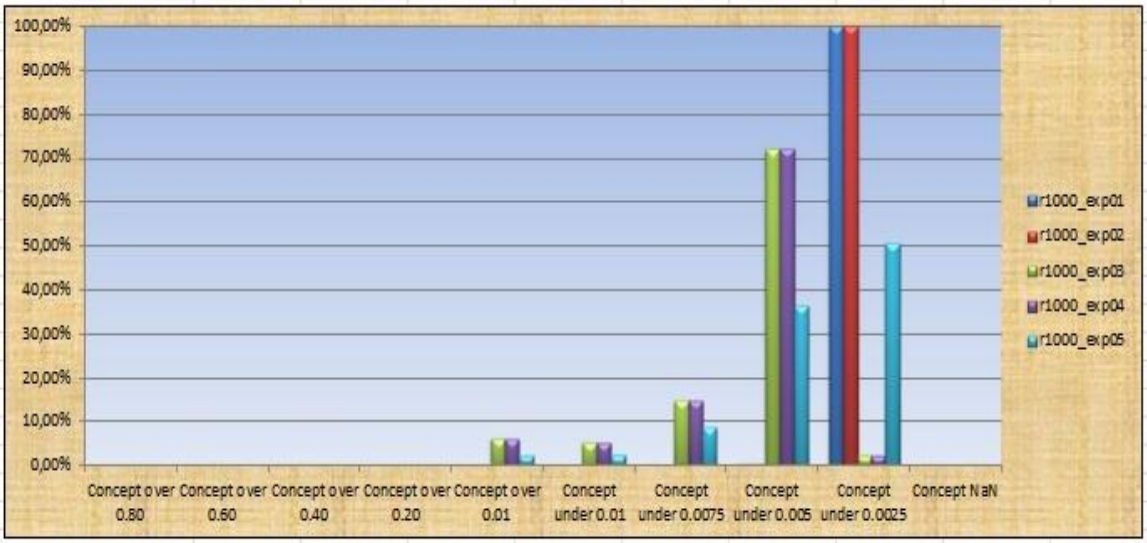


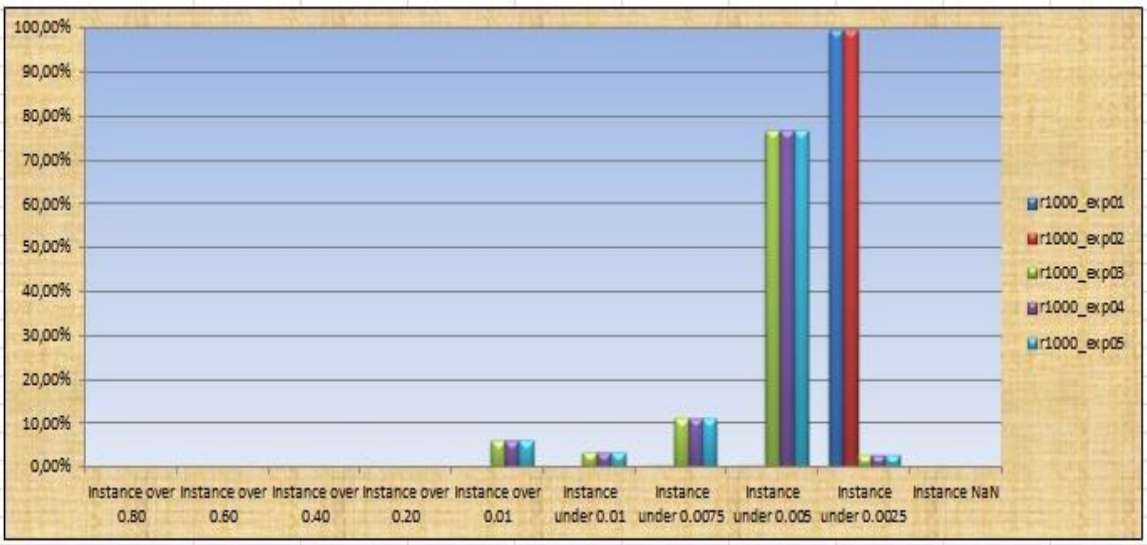


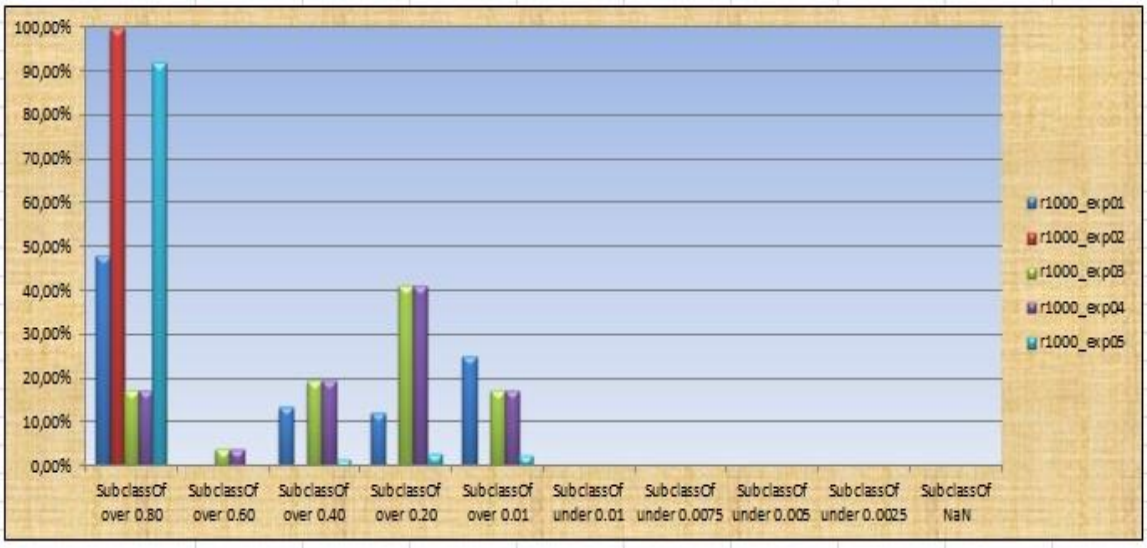


QUANTITATIVE GROUPING AT PROBABILITY LEVELS - PERCENTAGE RATIO

AMAZON SAME MOVIES REVIEWS (PER NUMBER OF REVIEWS)

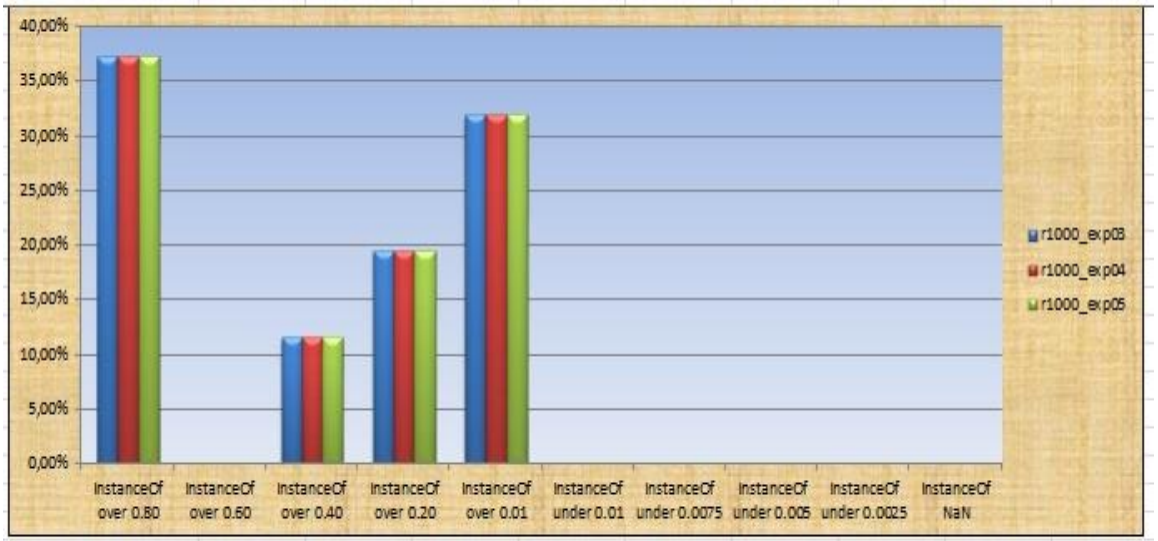


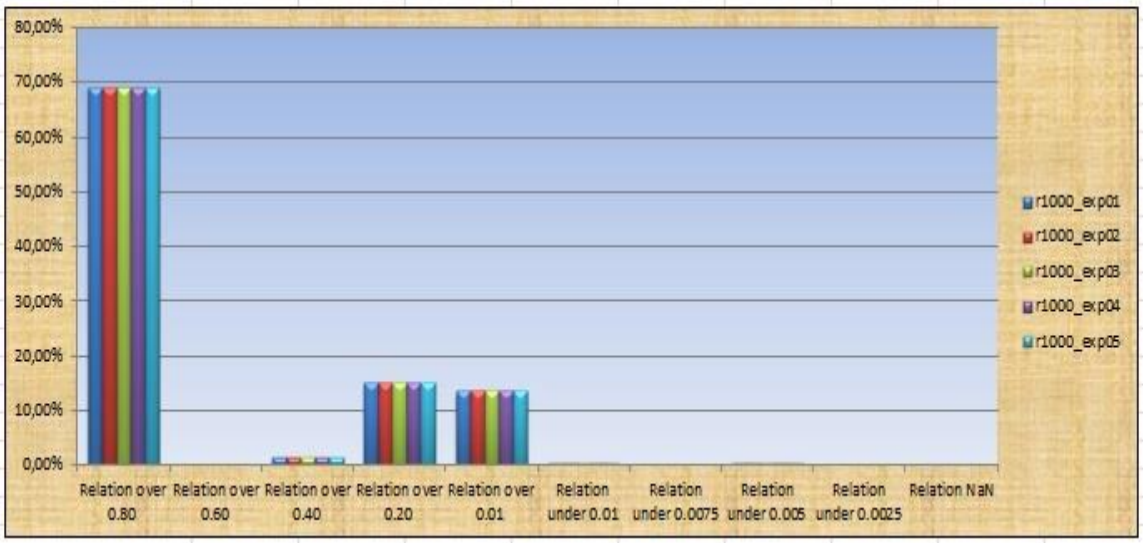




QUANTITATIVE GROUPING AT PROBABILITY LEVELS - PERCENTAGE RATIO

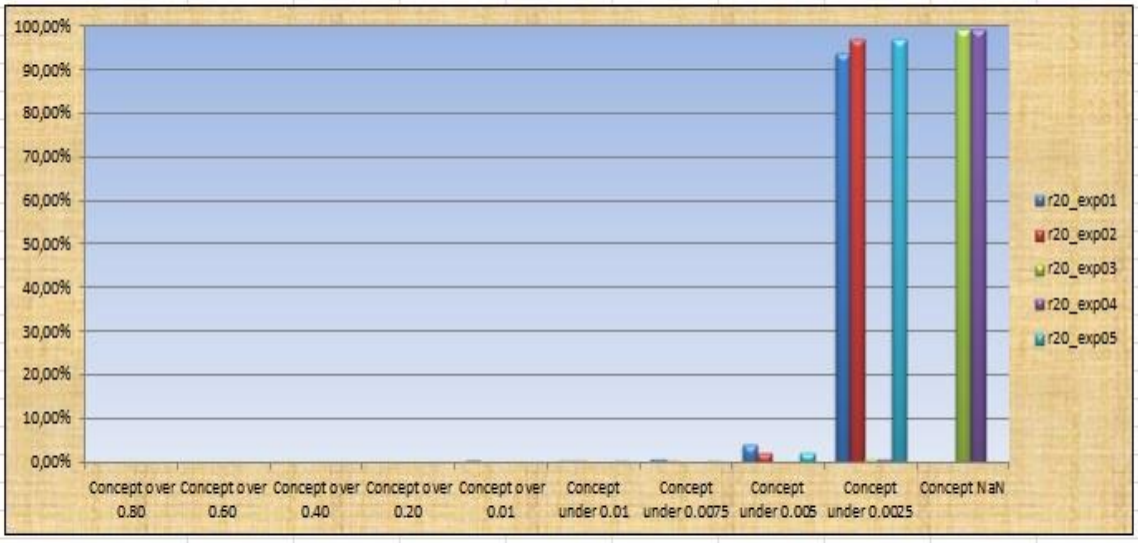
AMAZON SAME MOVIES REVIEWS (PER NUMBER OF REVIEWS)

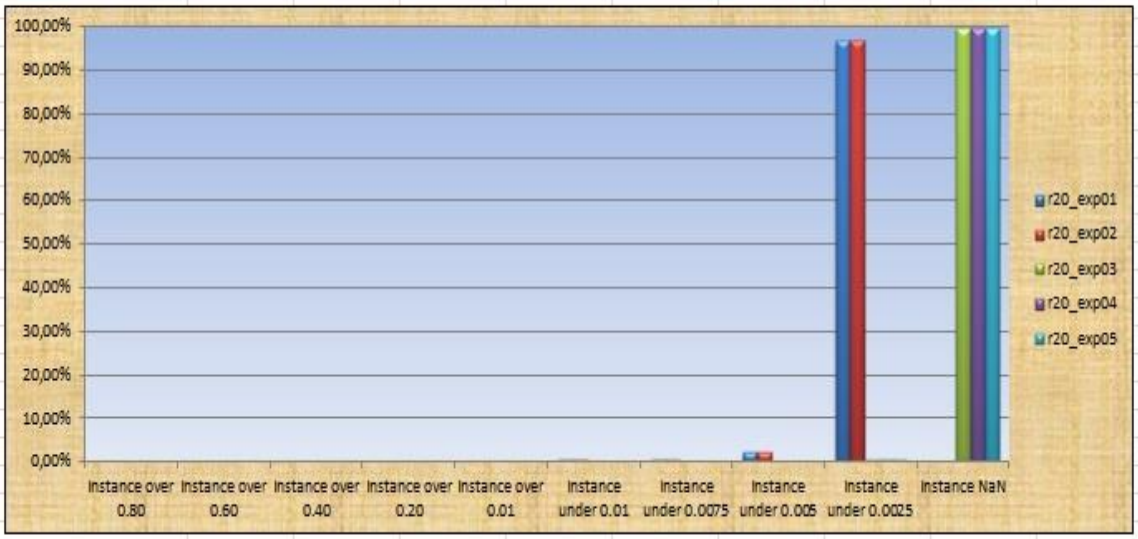


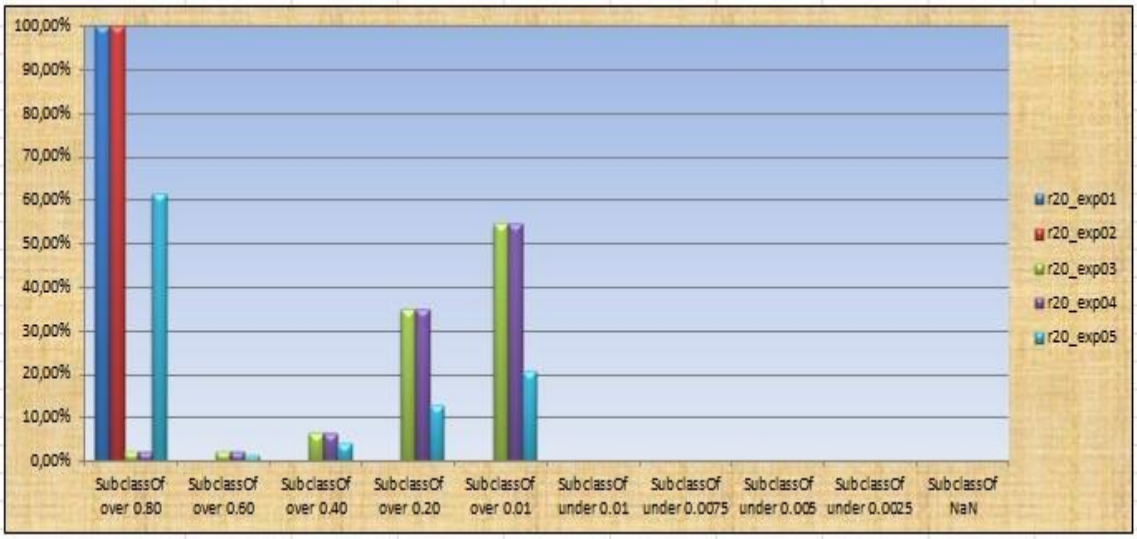


QUANTITATIVE GROUPING AT PROBABILITY LEVELS - PERCENTAGE RATIO

IMDB MOVIES REVIEWS (PER NUMBER OF REVIEWS)

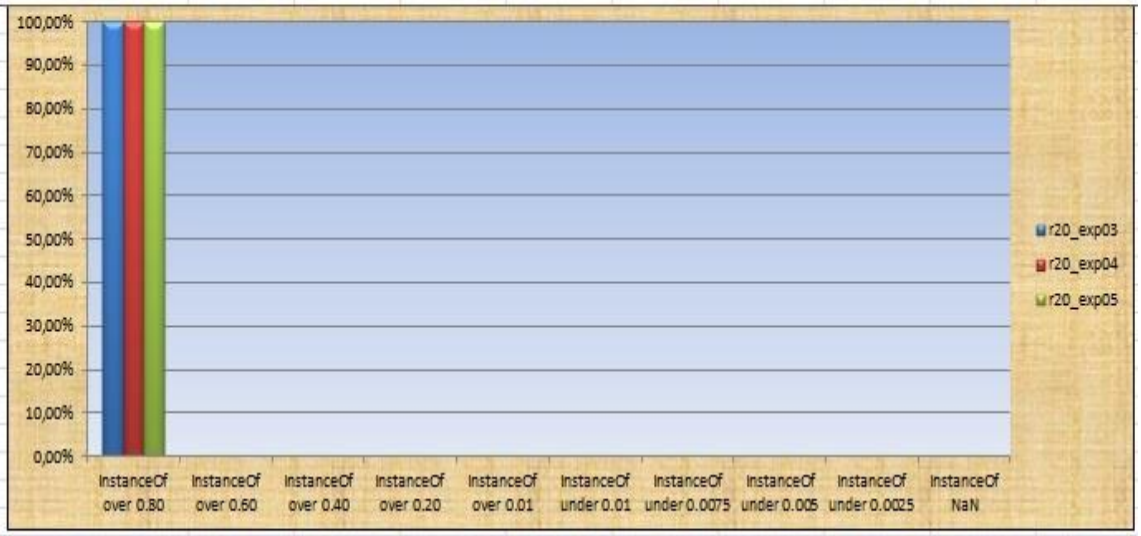


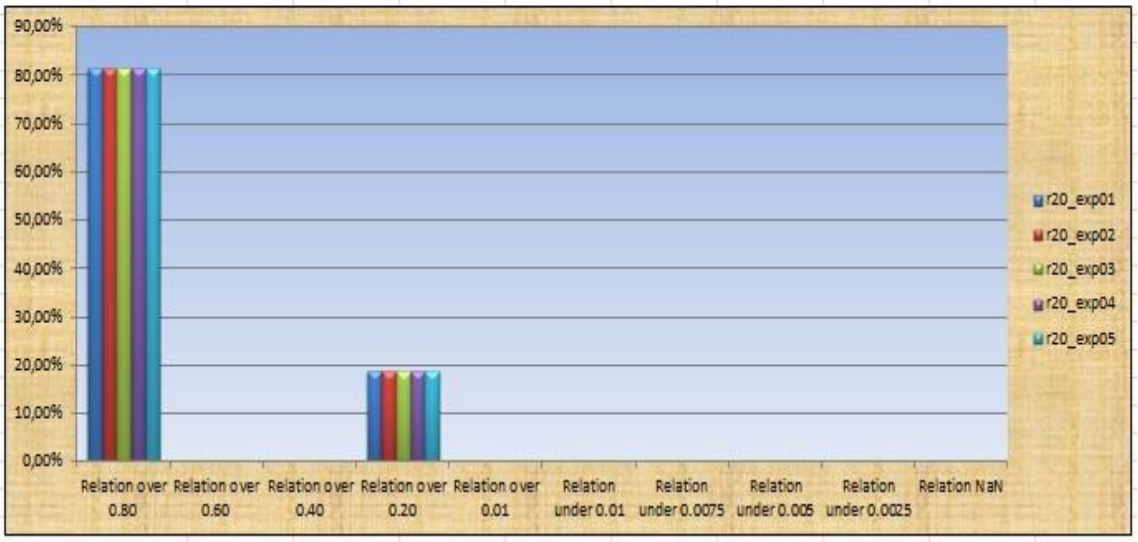


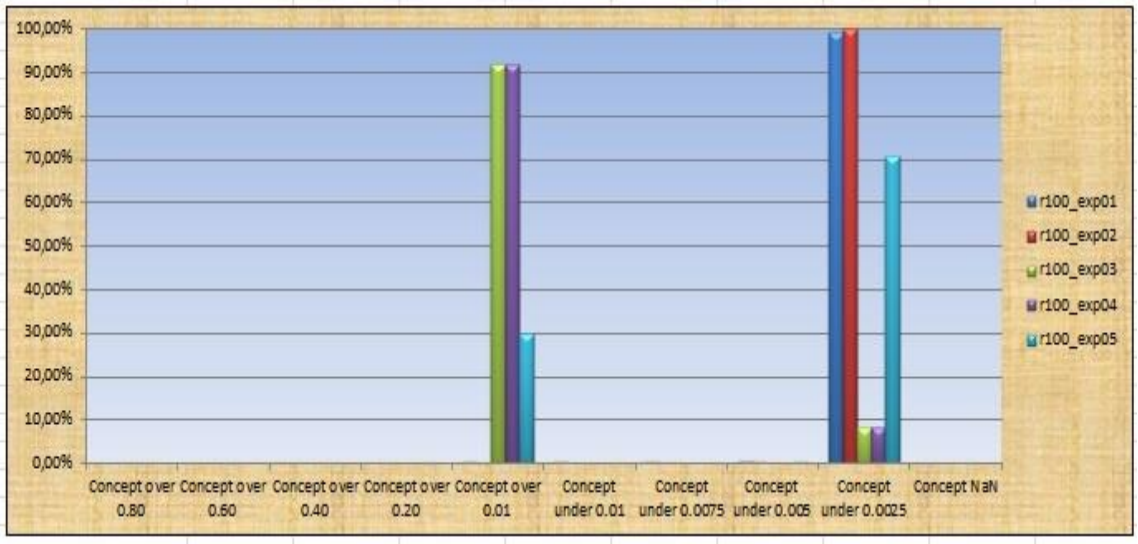


QUANTITATIVE GROUPING AT PROBABILITY LEVELS - PERCENTAGE RATIO

IMDB MOVIES REVIEWS (PER NUMBER OF REVIEWS)

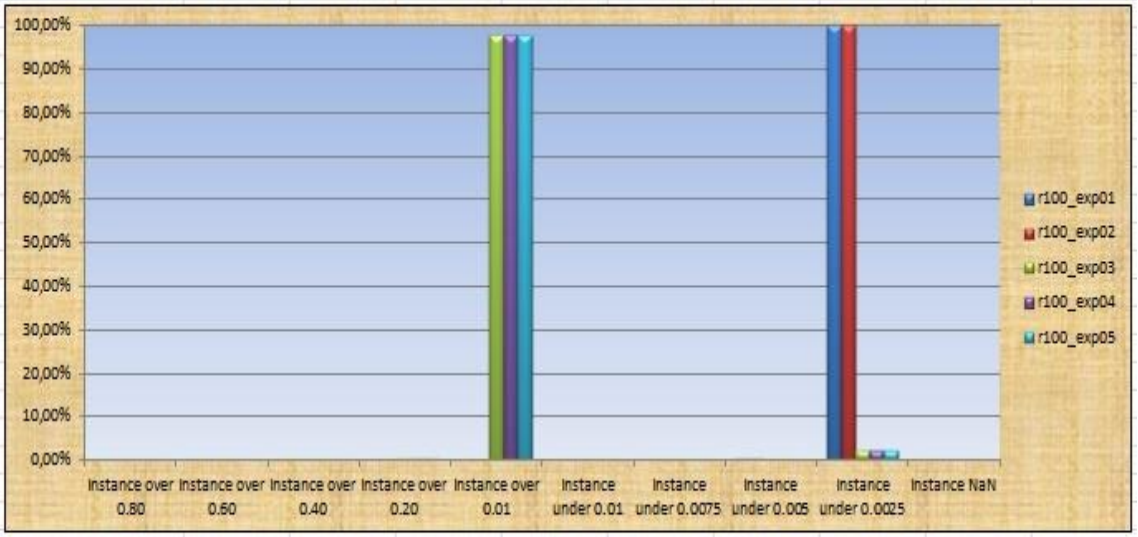


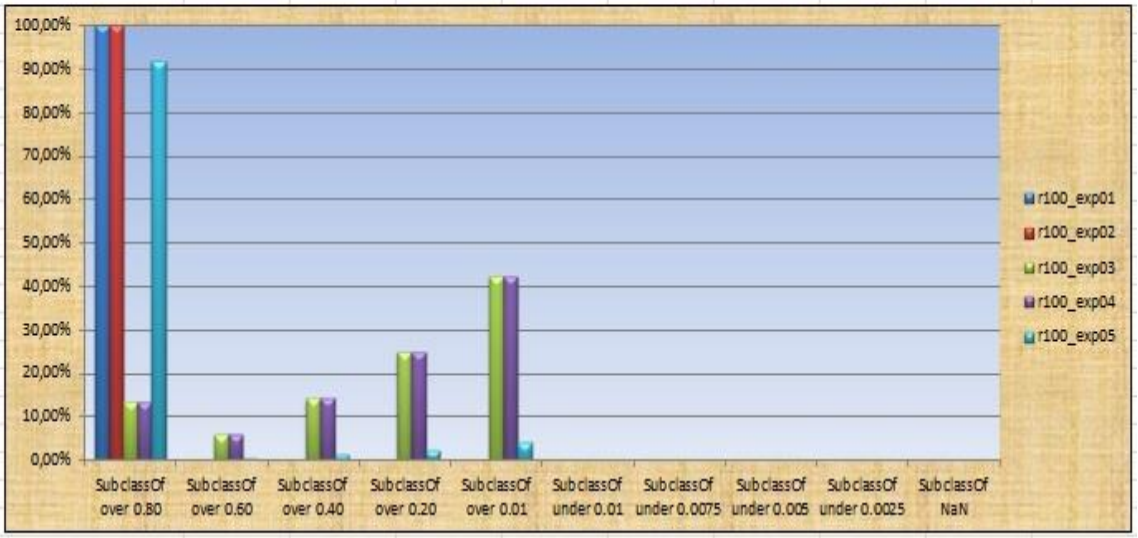


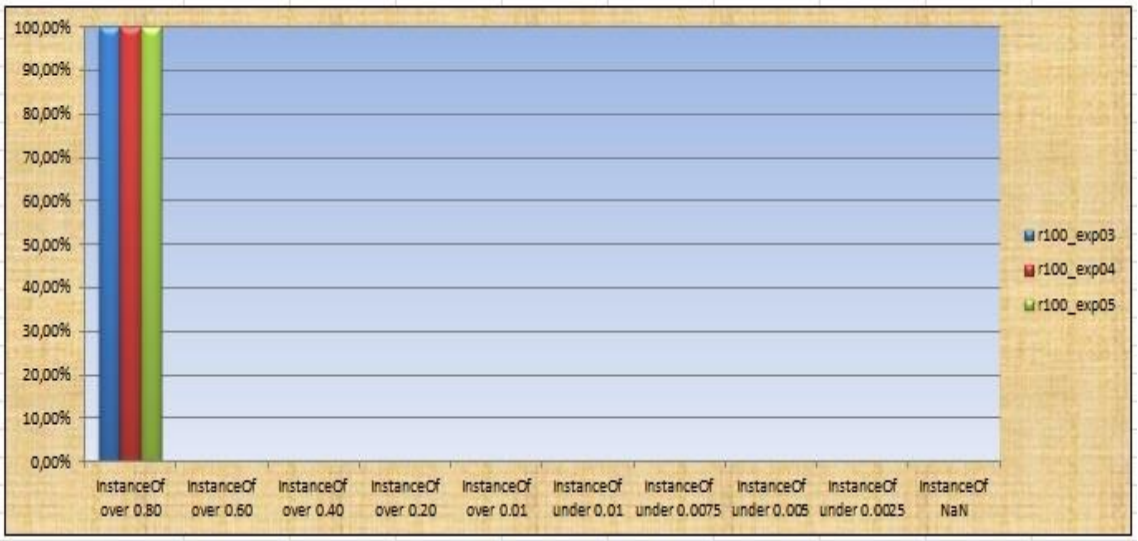


QUANTITATIVE GROUPING AT PROBABILITY LEVELS - PERCENTAGE RATIO

IMDB MOVIES REVIEWS (PER NUMBER OF REVIEWS)

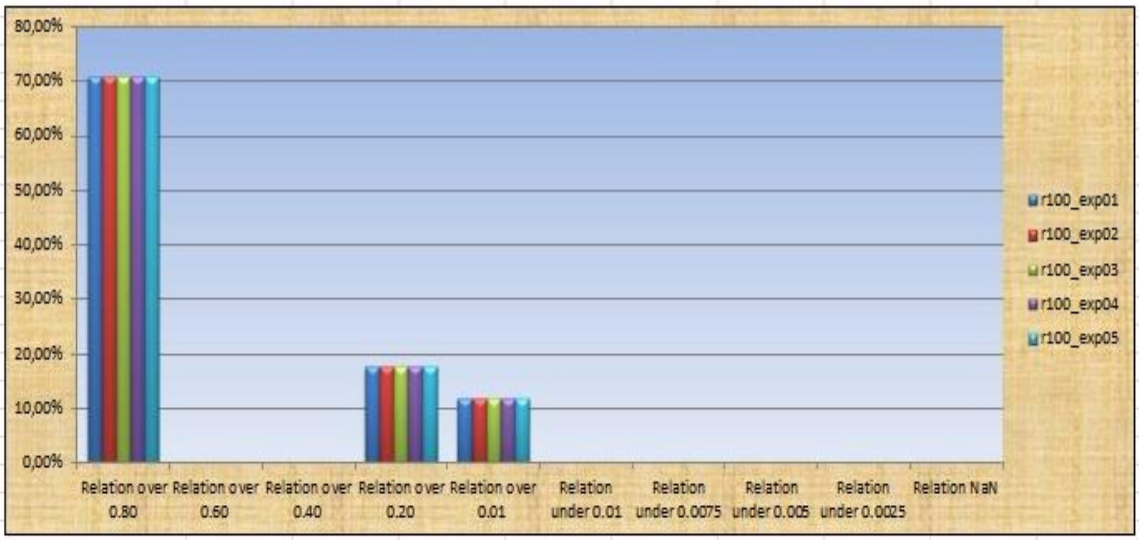


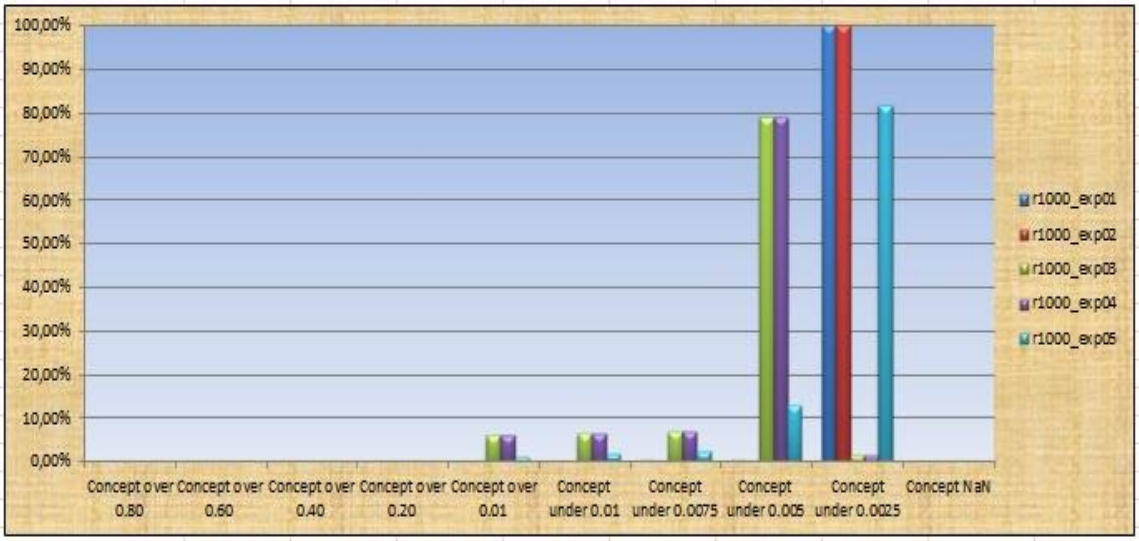


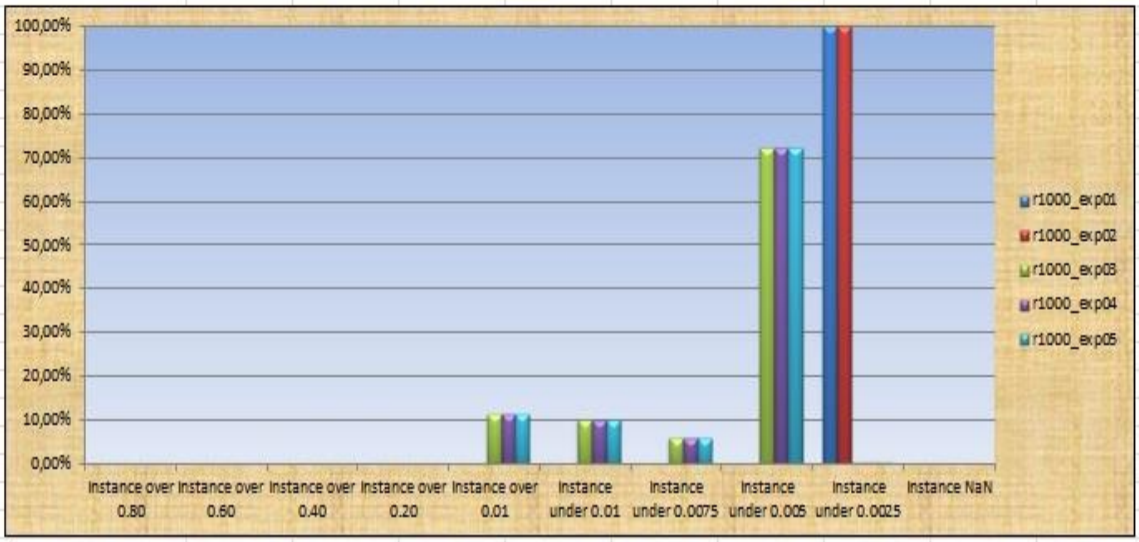


QUANTITATIVE GROUPING AT PROBABILITY LEVELS - PERCENTAGE RATIO

IMDB MOVIES REVIEWS (PER NUMBER OF REVIEWS)

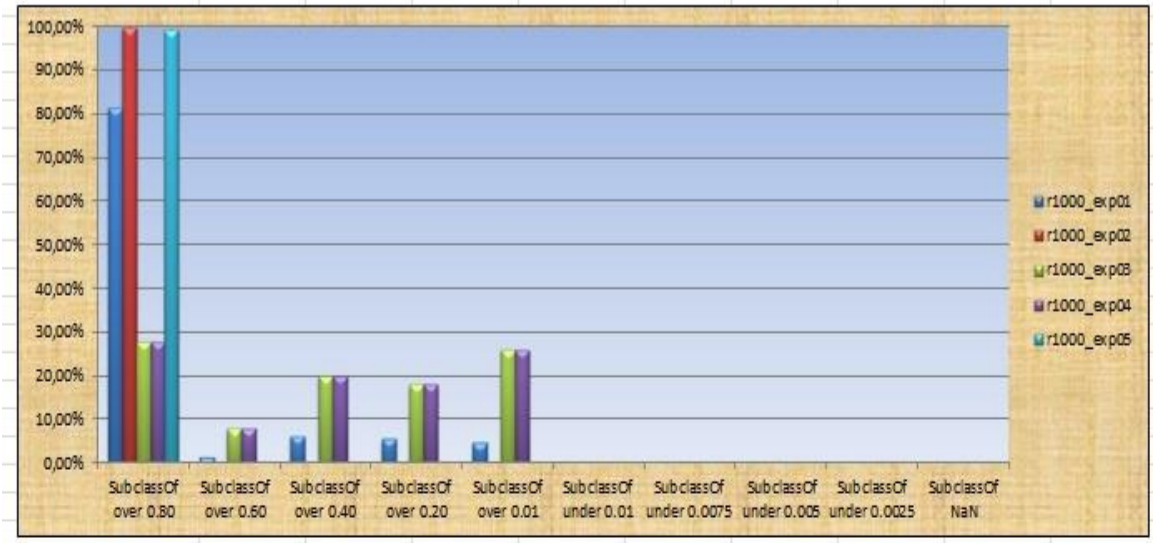


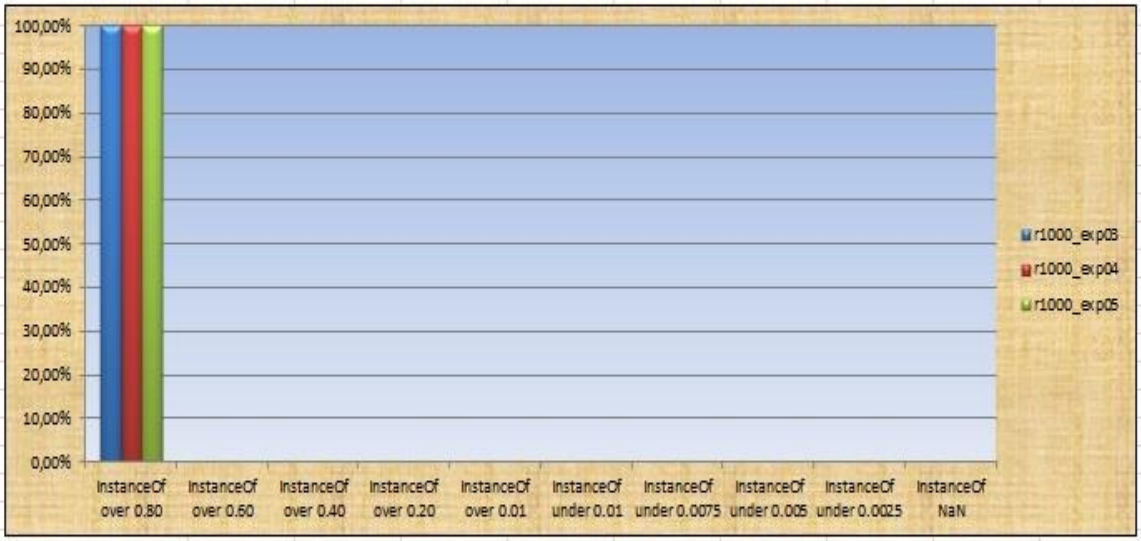


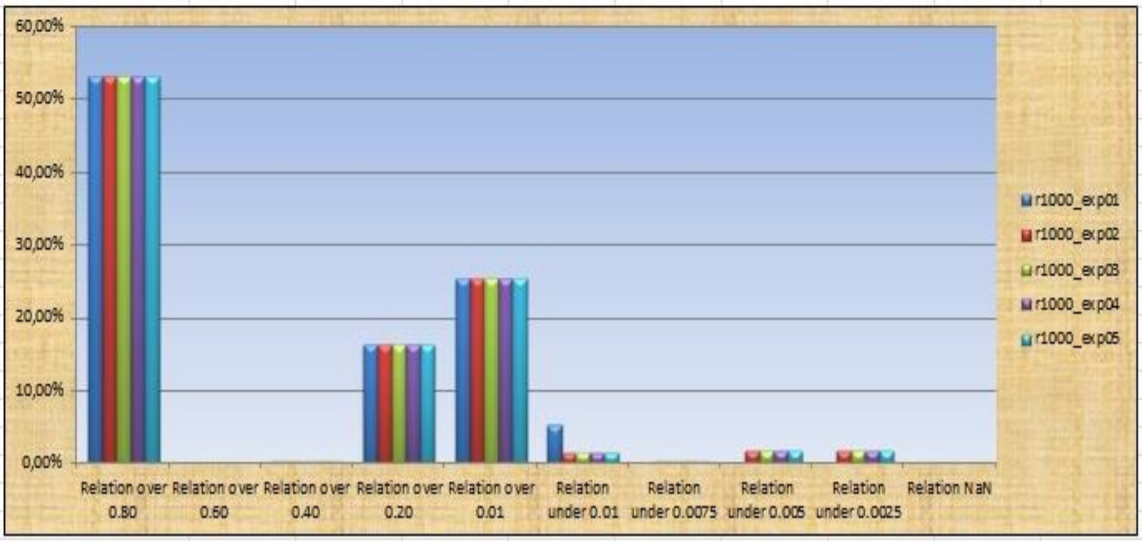


QUANTITATIVE GROUPING AT PROBABILITY LEVELS - PERCENTAGE RATIO

IMDB MOVIES REVIEWS (PER NUMBER OF REVIEWS)

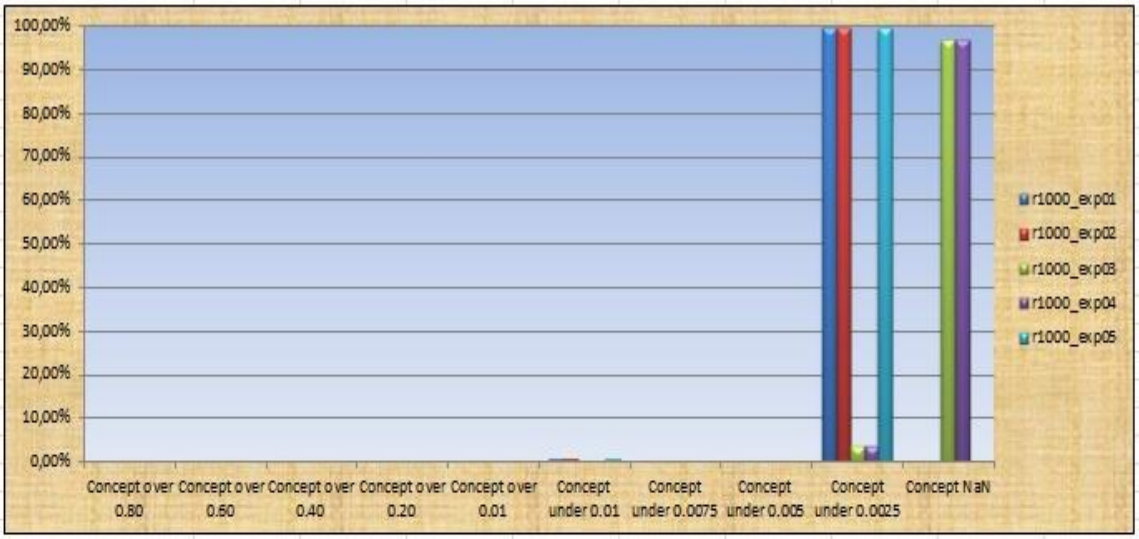


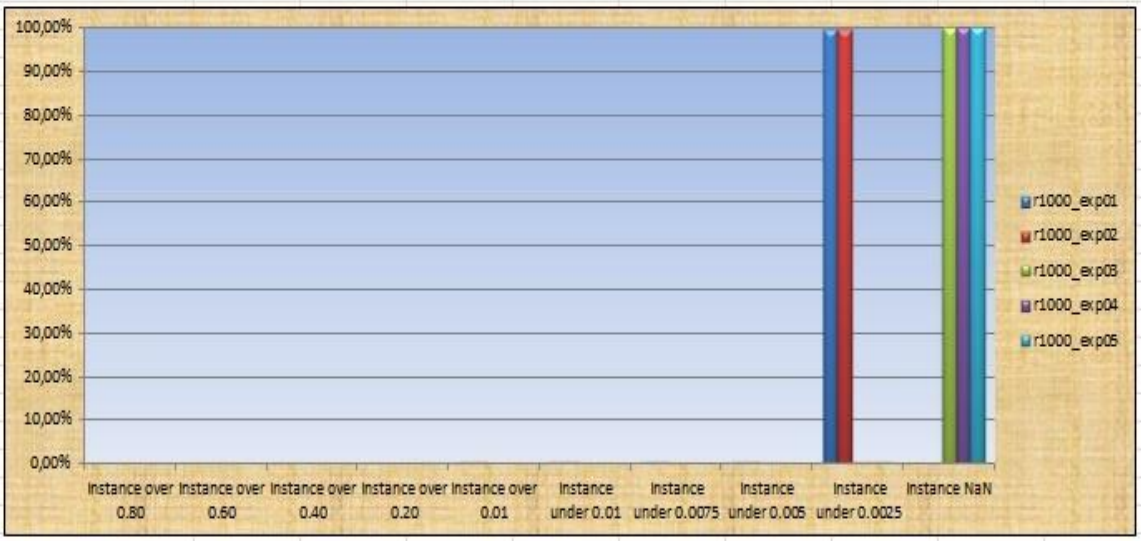


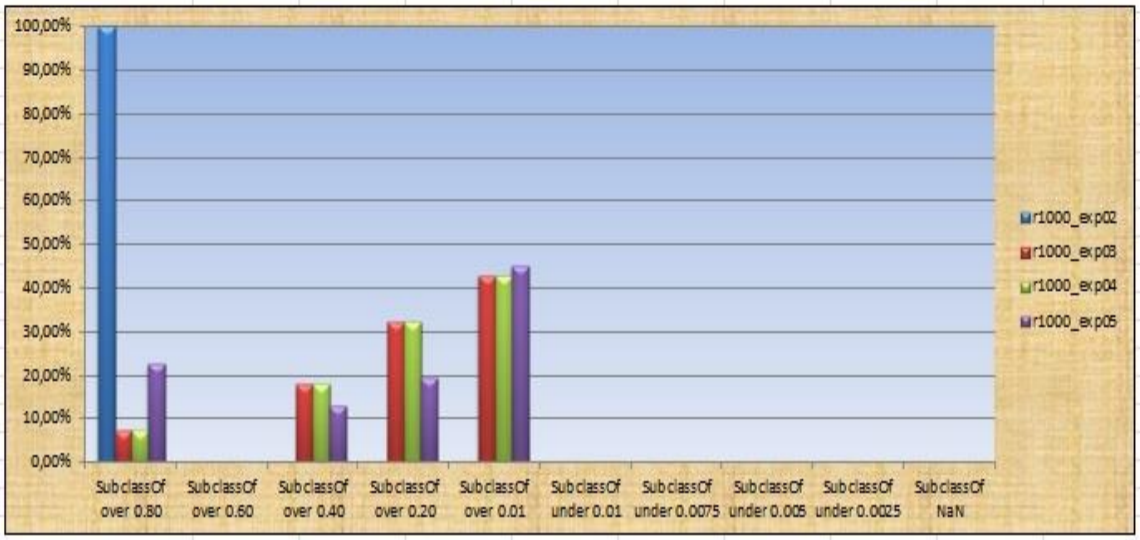


QUANTITATIVE GROUPING AT PROBABILITY LEVELS - PERCENTAGE RATIO

IMDB MOVIES DATA (PER NUMBER OF DATA RECORDS)

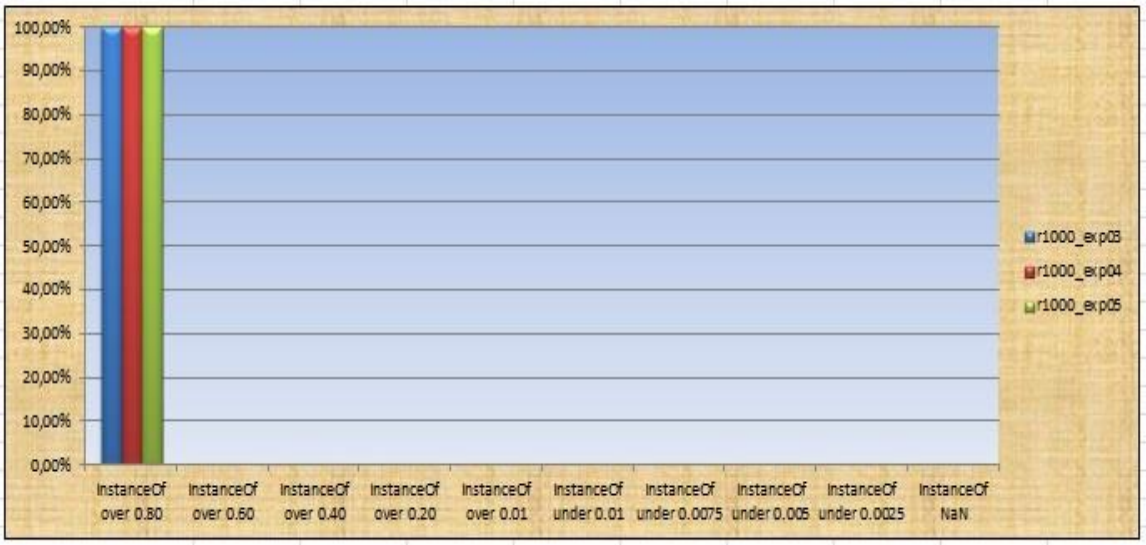


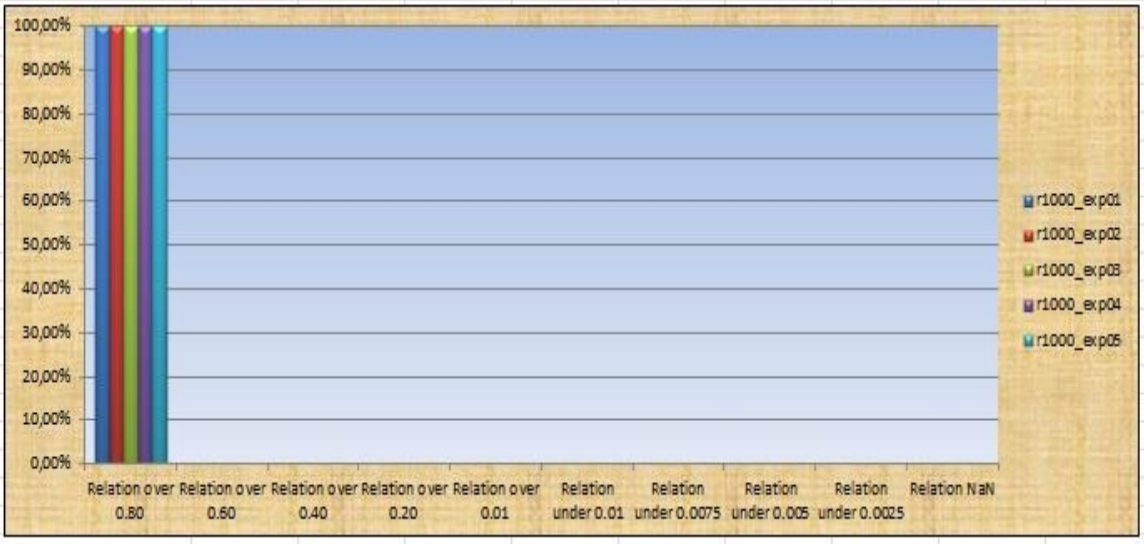




QUANTITATIVE GROUPING AT PROBABILITY LEVELS - PERCENTAGE RATIO

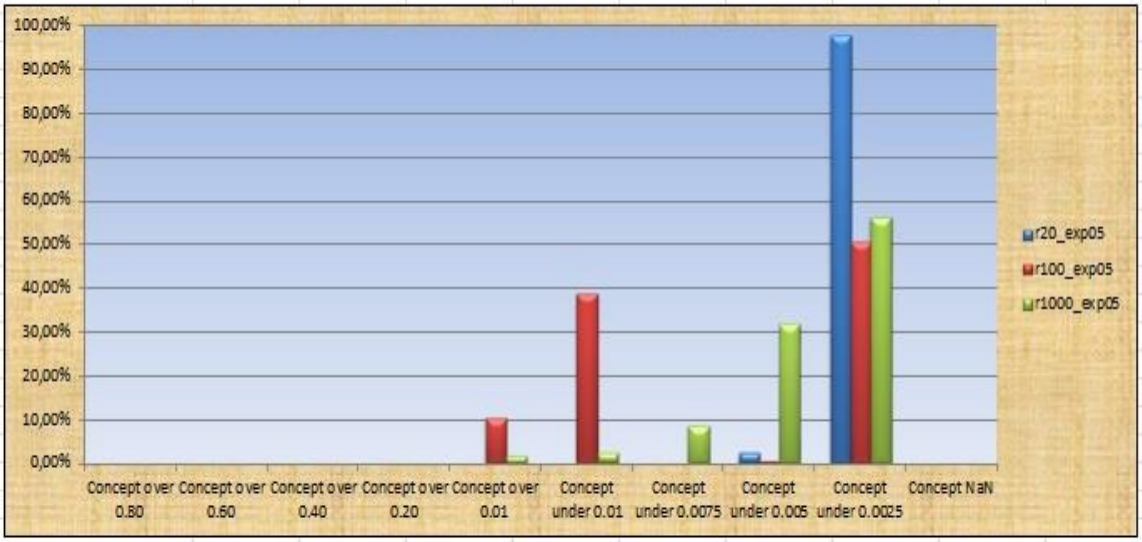
IMDB MOVIES DATA (PER NUMBER OF DATA RECORDS)

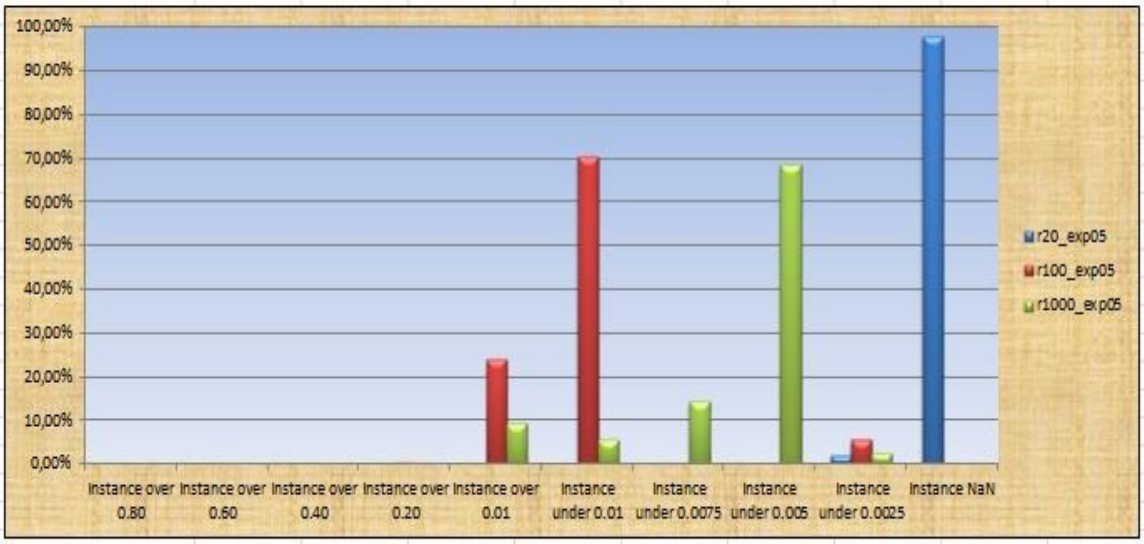


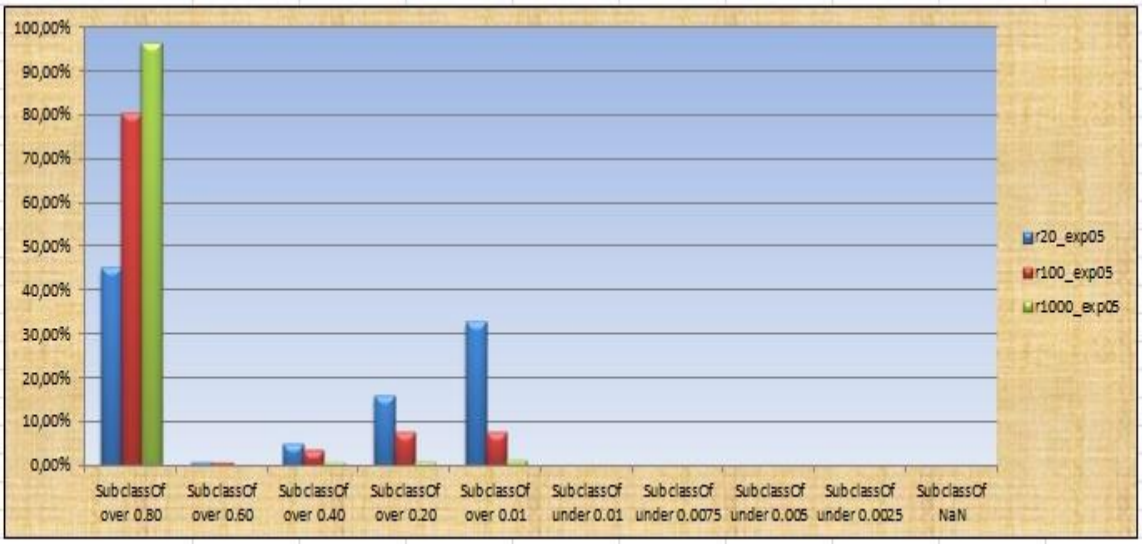


QUANTITATIVE GROUPING AT PROBABILITY LEVELS - PERCENTAGE RATIO

AMAZON RANDOM MOVIES REVIEWS (PER EXPERIMENT)

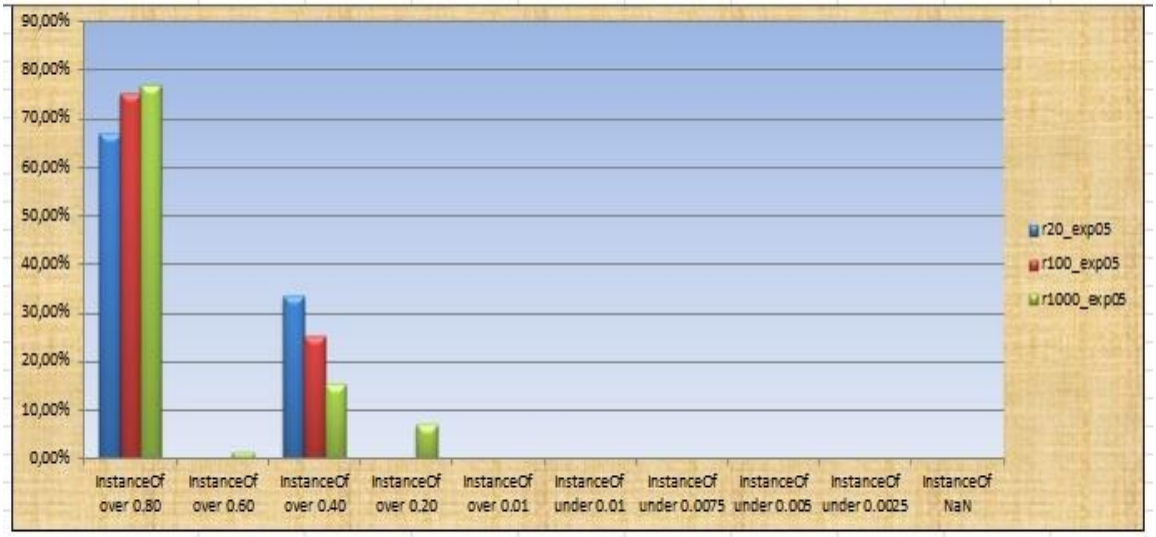


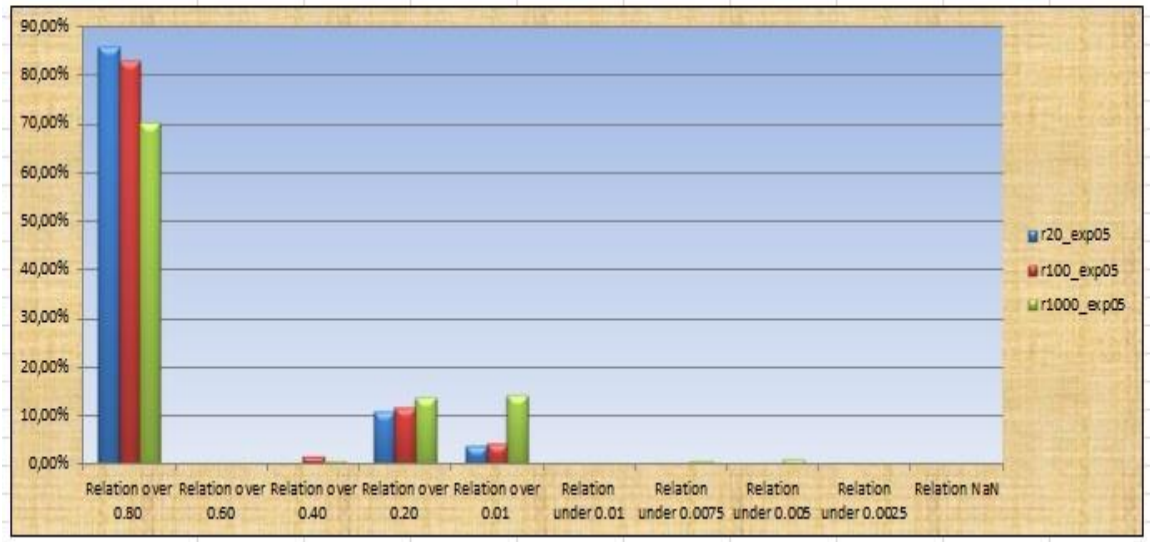




QUANTITATIVE GROUPING AT PROBABILITY LEVELS - PERCENTAGE RATIO

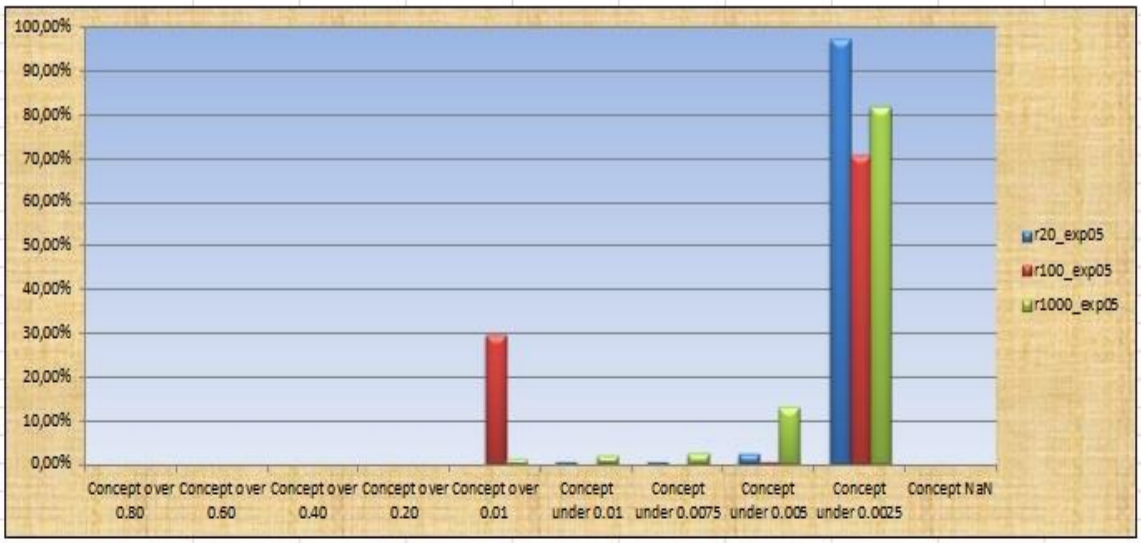
AMAZON RANDOM MOVIES REVIEWS (PER EXPERIMENT)

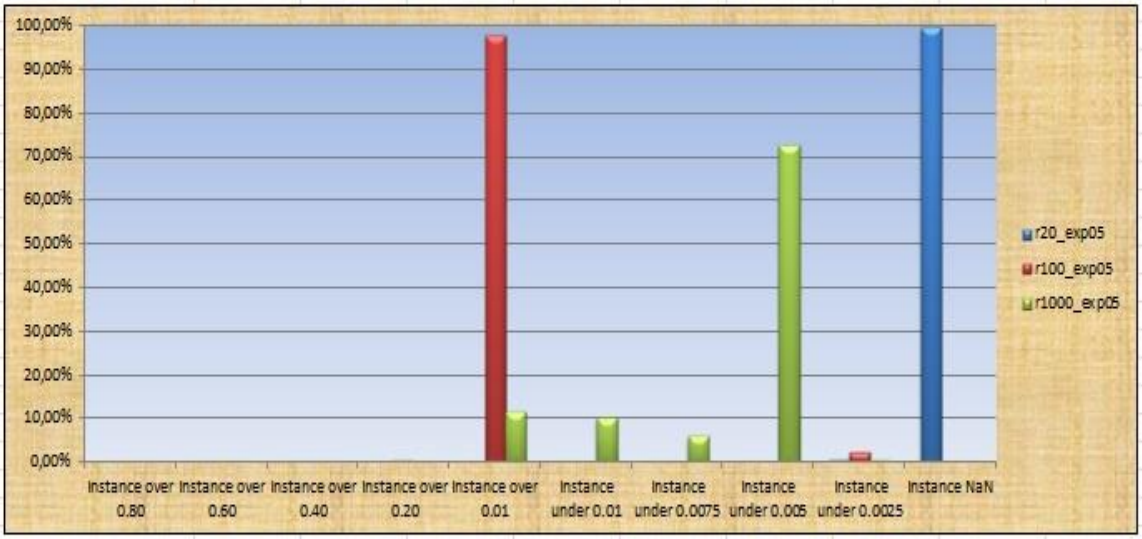


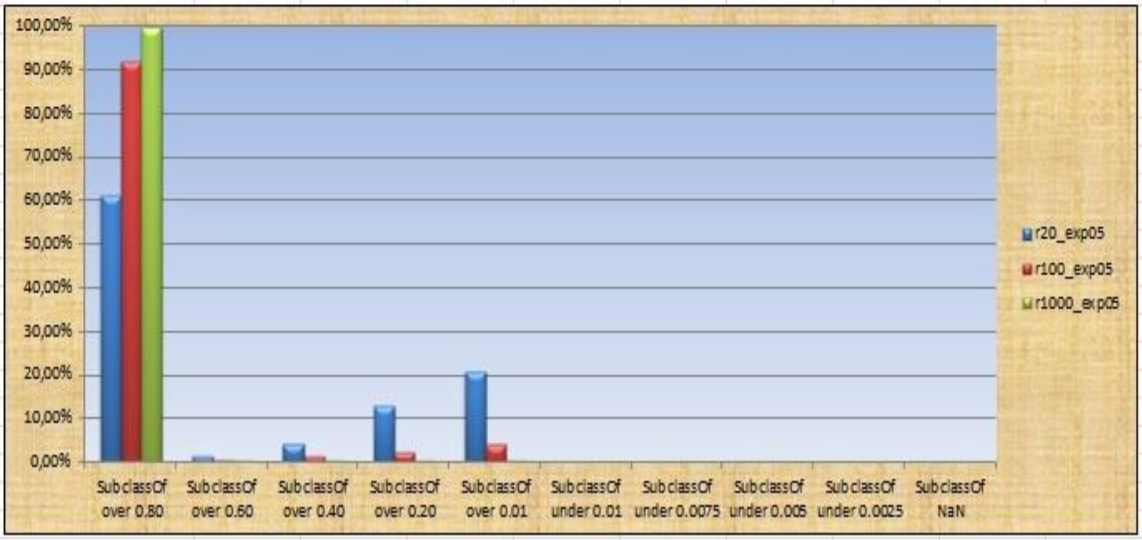


QUANTITATIVE GROUPING AT PROBABILITY LEVELS - PERCENTAGE RATIO

IMDB MOVIES REVIEWS (PER EXPERIMENT)

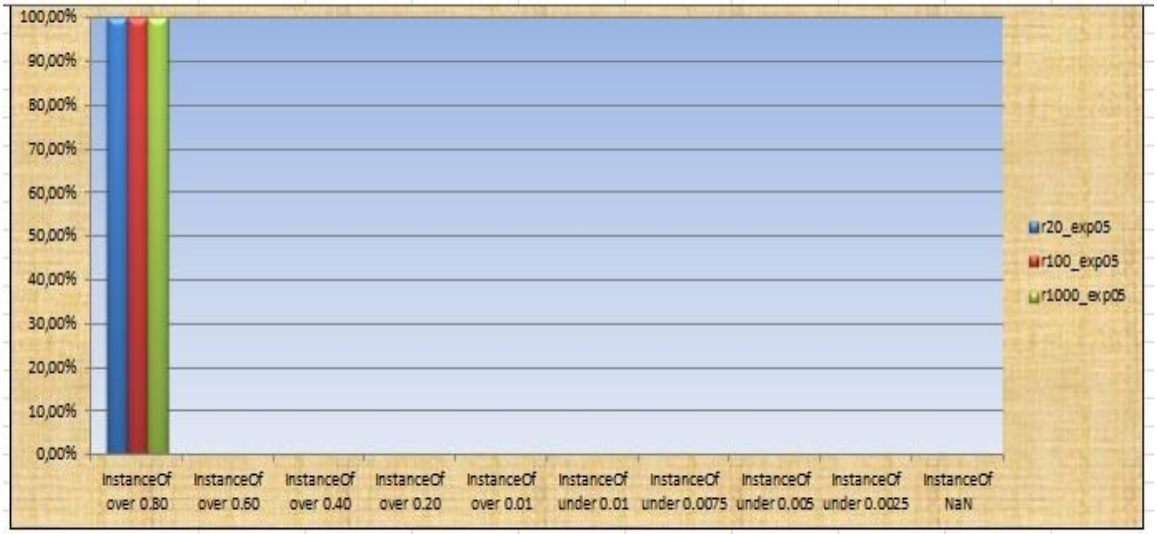


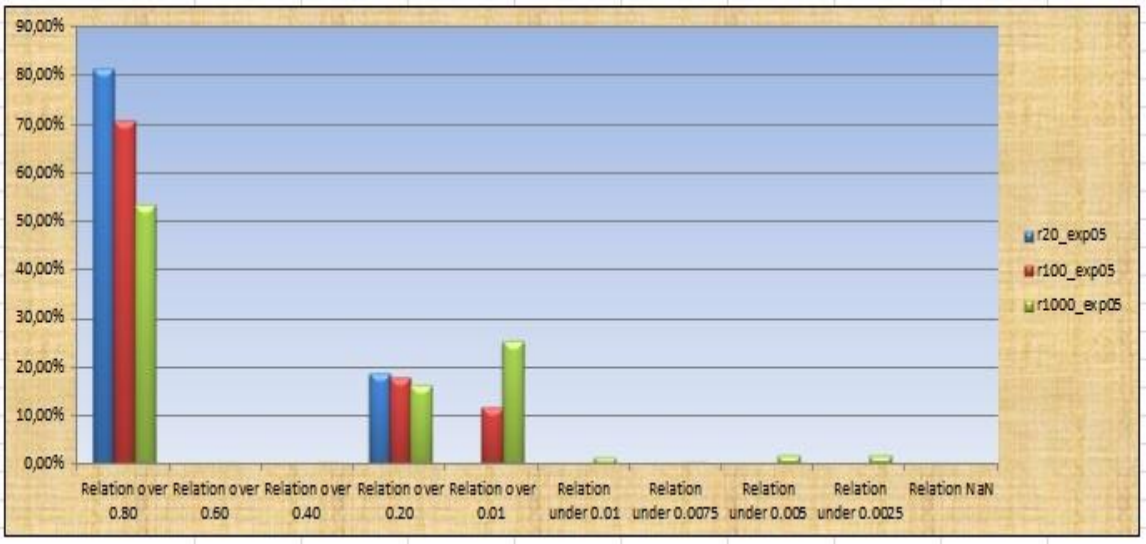




QUANTITATIVE GROUPING AT PROBABILITY LEVELS - PERCENTAGE RATIO

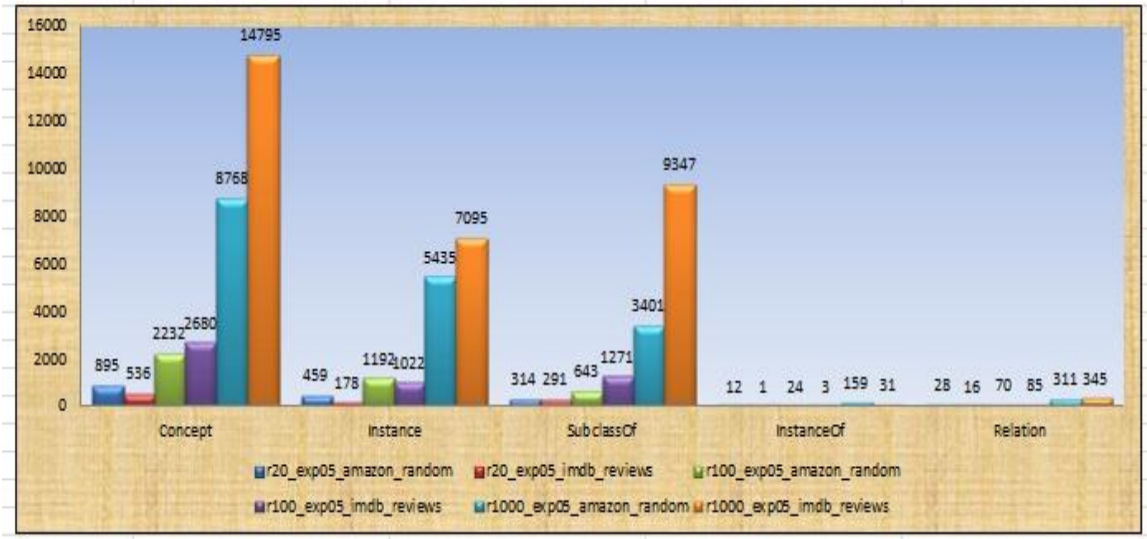
IMDB MOVIES REVIEWS (PER EXPERIMENT)





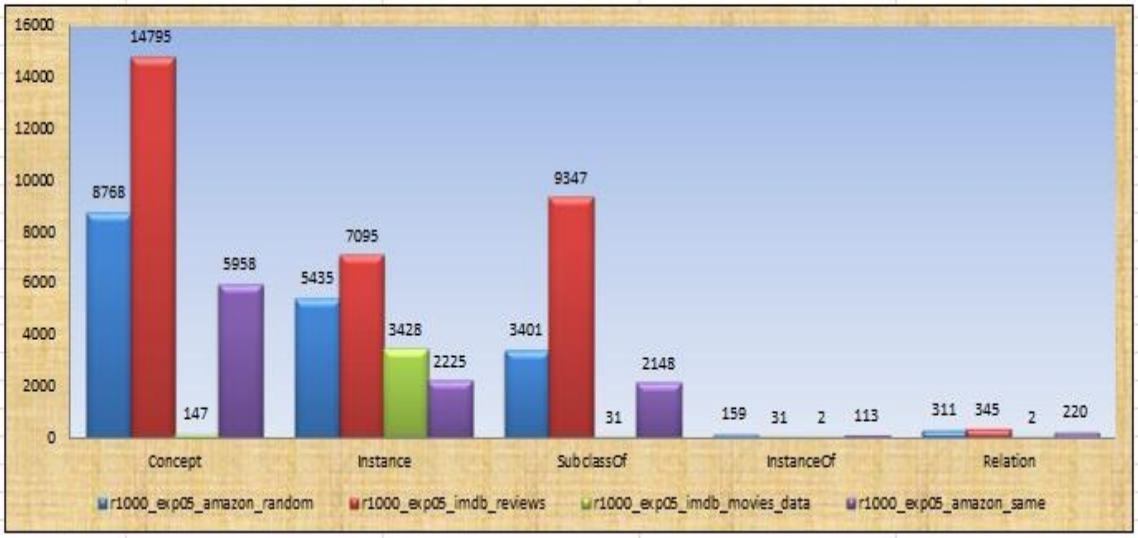
ENUMERATION

COMPARISON - AMAZON RANDOM MOVIES REVIEWS, IMDB MOVIES REVIEWS (PER NUMBER OF REVIEWS)



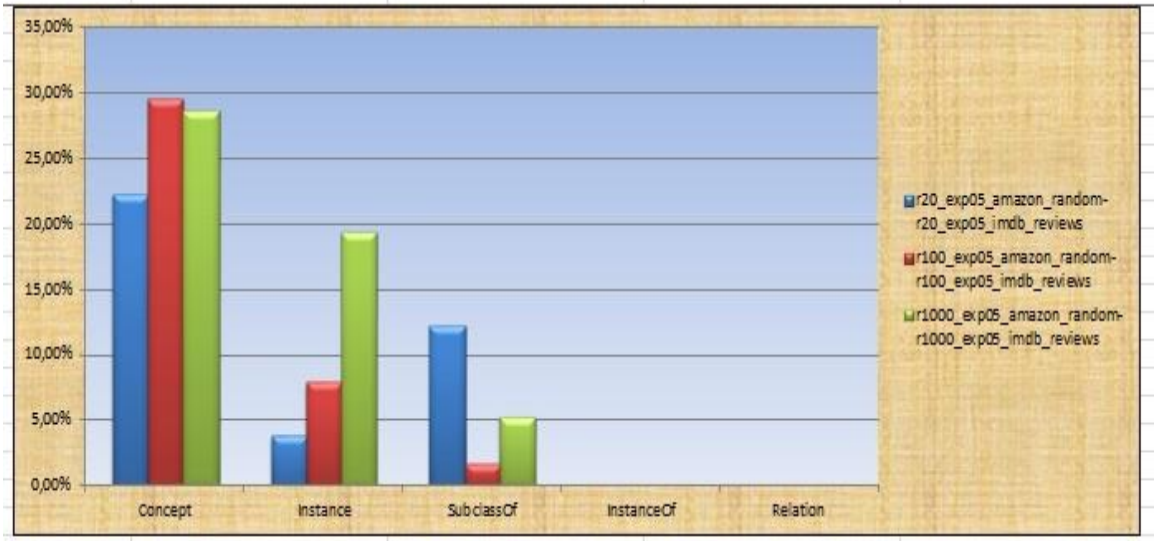
ENUMERATION

COMPARISON - AMAZON RANDOM MOVIES REVIEWS, IMDB MOVIES REVIEWS, IMDB MOVIES DATA, AMAZON SAME MOVIES REVIEWS



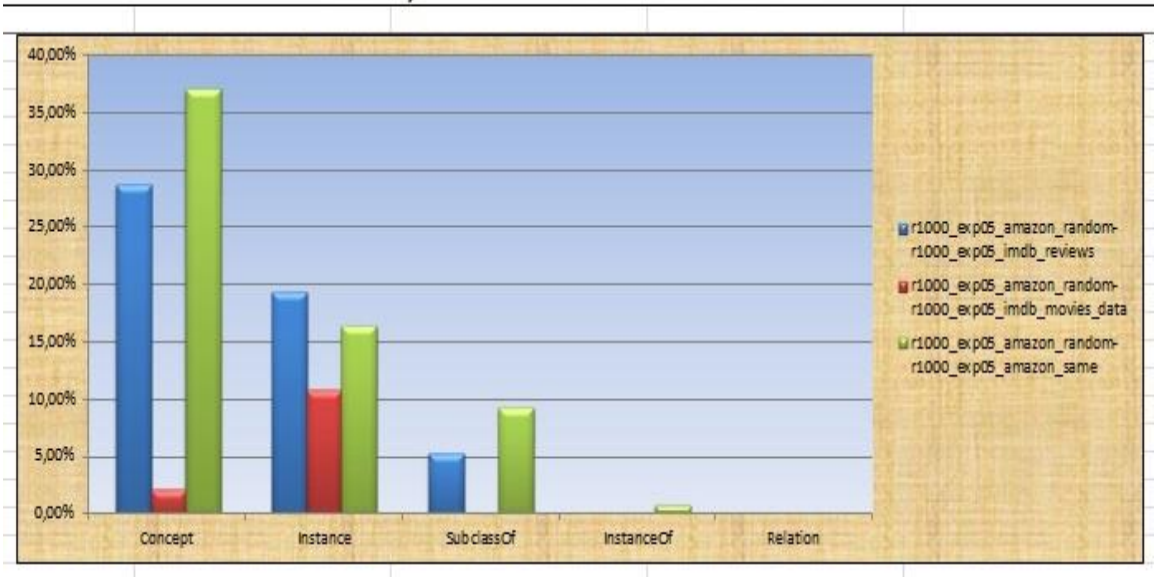
SAME PRIMITIVES COMPARISON - PERCENTAGE RATIO

COMPARISON - AMAZON RANDOM MOVIES REVIEWS, IMDB MOVIES REVIEWS (PER NUMBER OF REVIEWS)

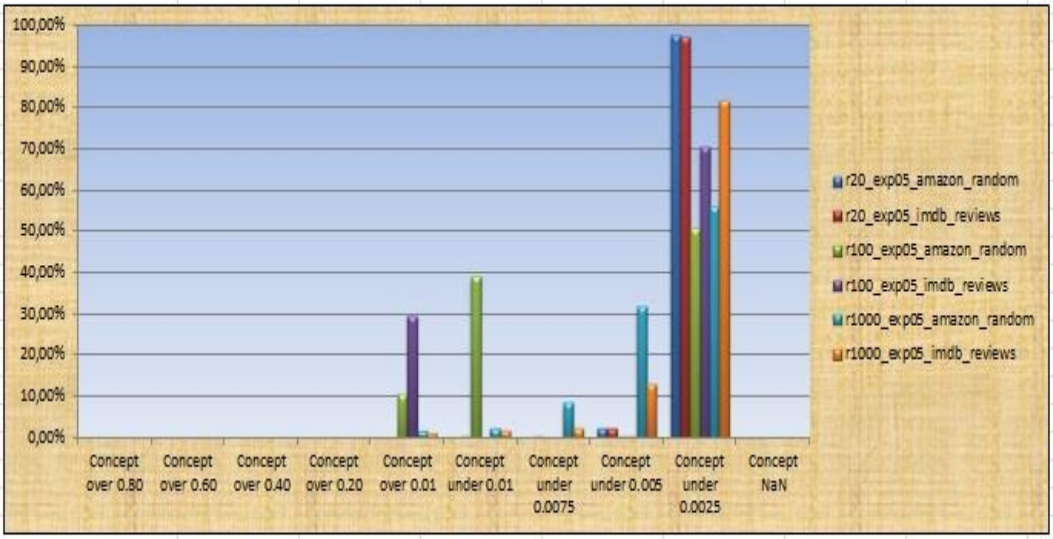


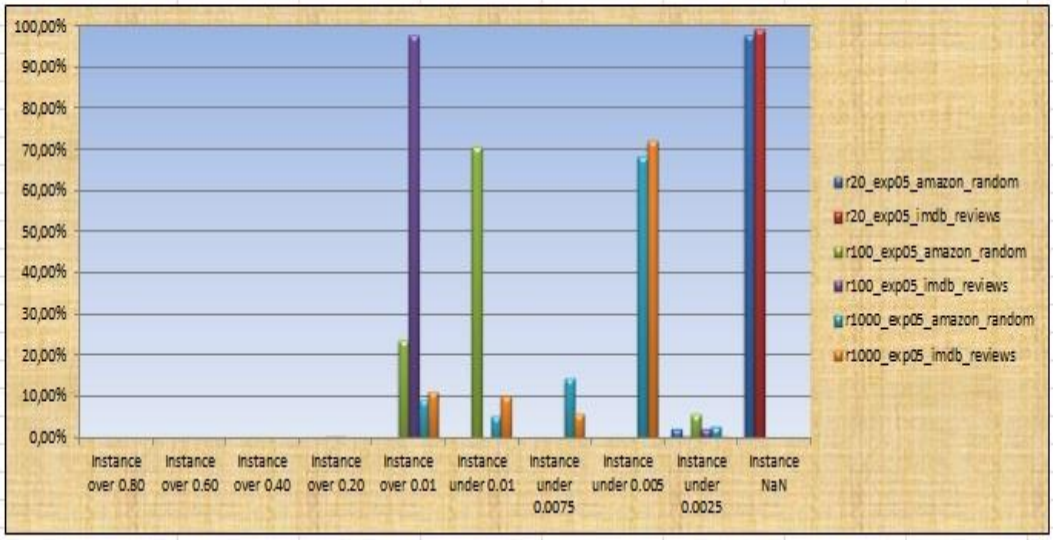
SAME PRIMITIVES COMPARISON - PERCENTAGE RATIO

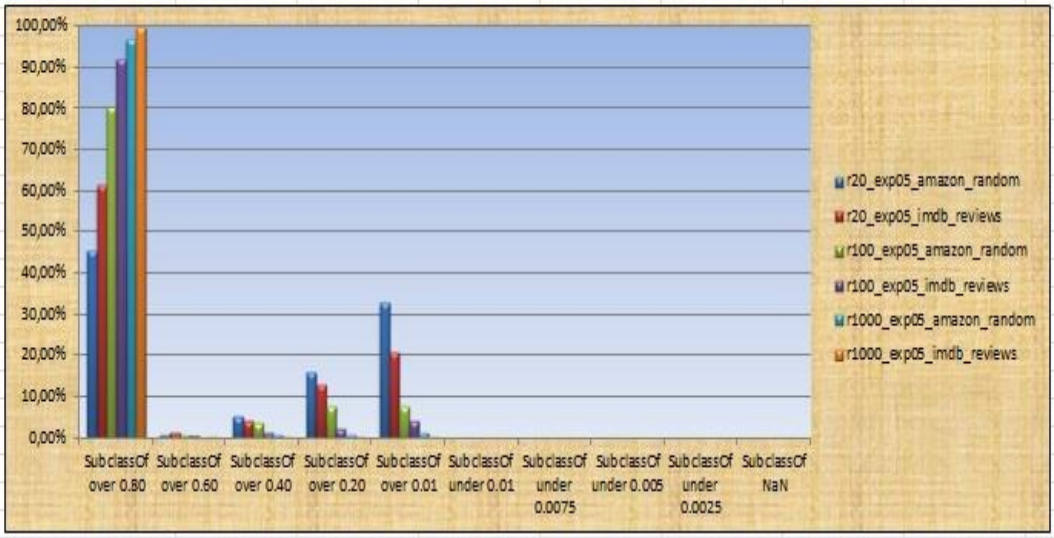
COMPARISON - AMAZON RANDOM MOVIES REVIEWS, IMDB MOVIES REVIEWS, IMDB MOVIES DATA, AMAZON SAME MOVIES REVIEWS



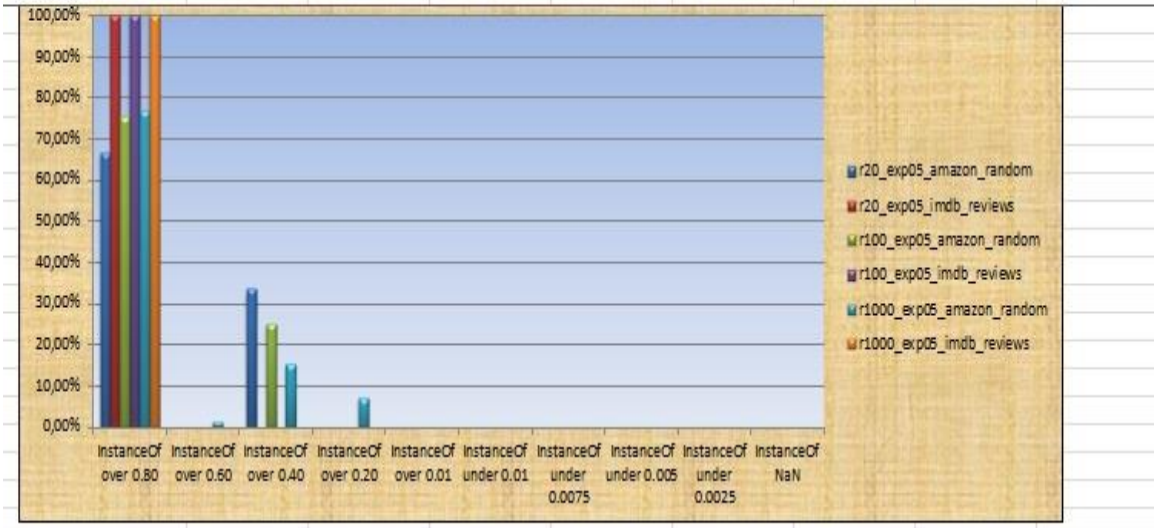
QUANTITATIVE GROUPING AT PROBABILITY LEVELS - PERCENTAGE RATIO COMPARISON - AMAZON RANDOM MOVIES REVIEWS, IMDB MOVIES REVIEWS (PER NUMBER OF REVIEWS)

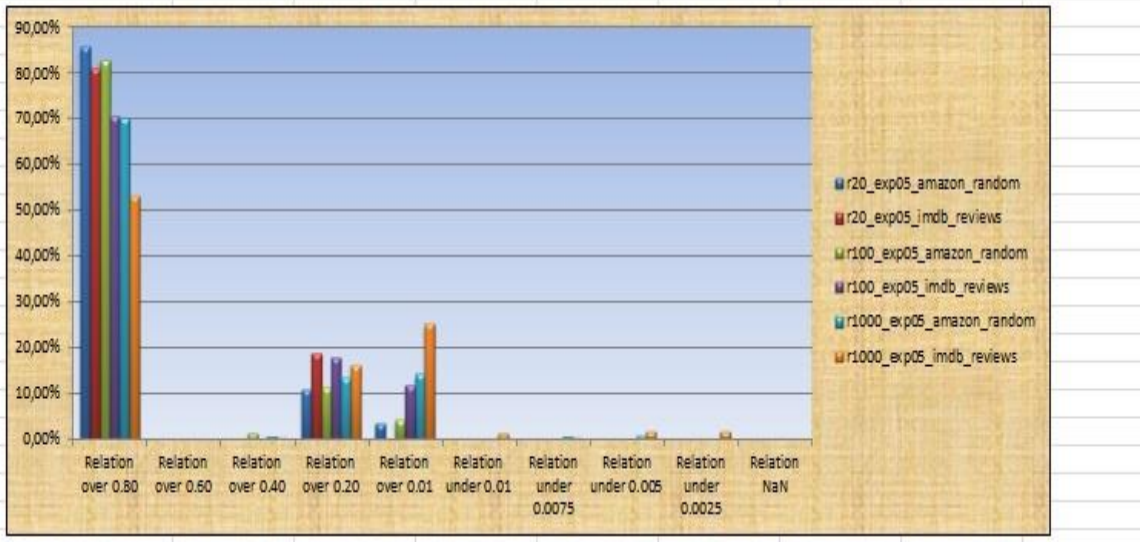




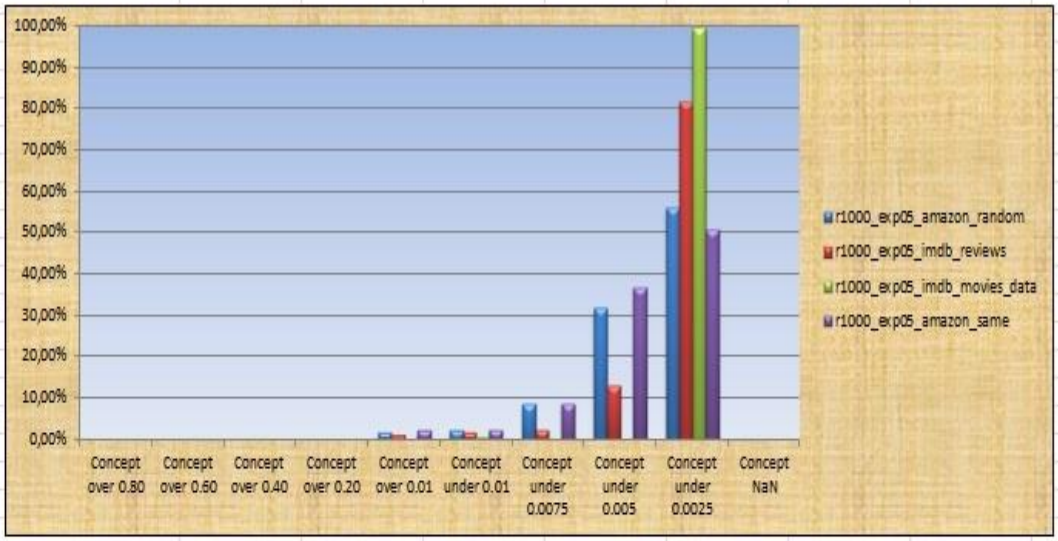


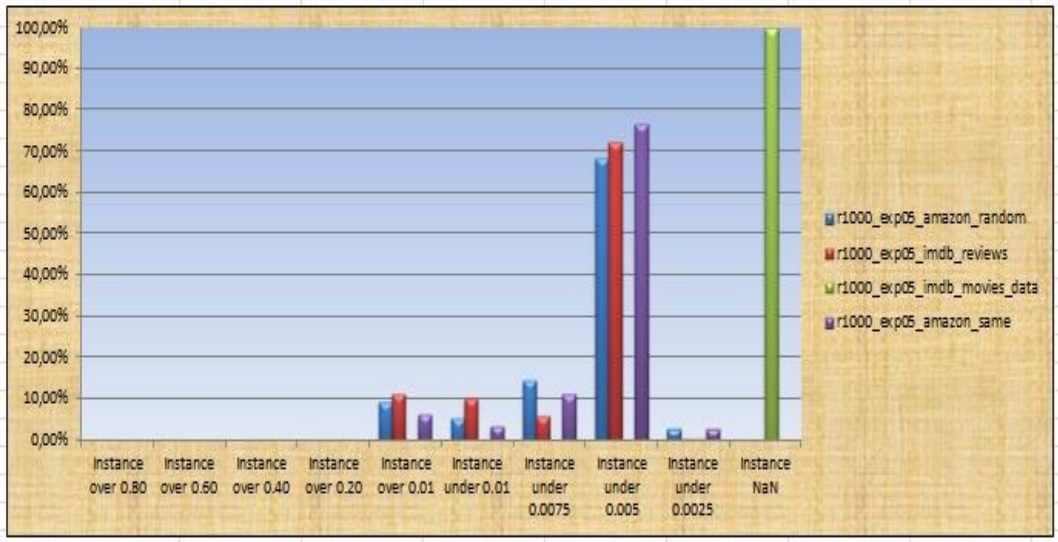
QUANTITATIVE GROUPING AT PROBABILITY LEVELS - PERCENTAGE RATIO COMPARISON - AMAZON RANDOM MOVIES REVIEWS, IMDB MOVIES REVIEWS (PER NUMBER OF REVIEWS)

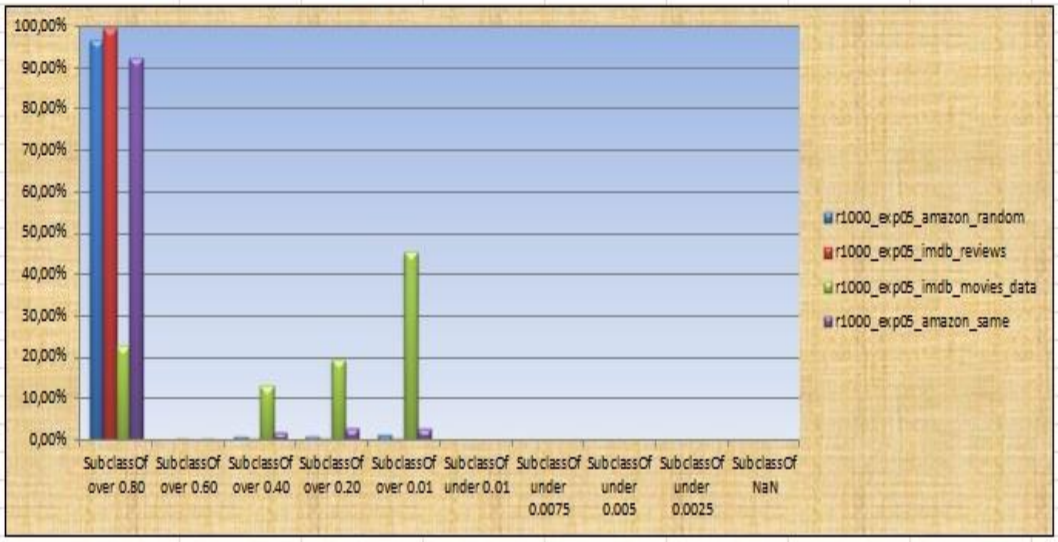




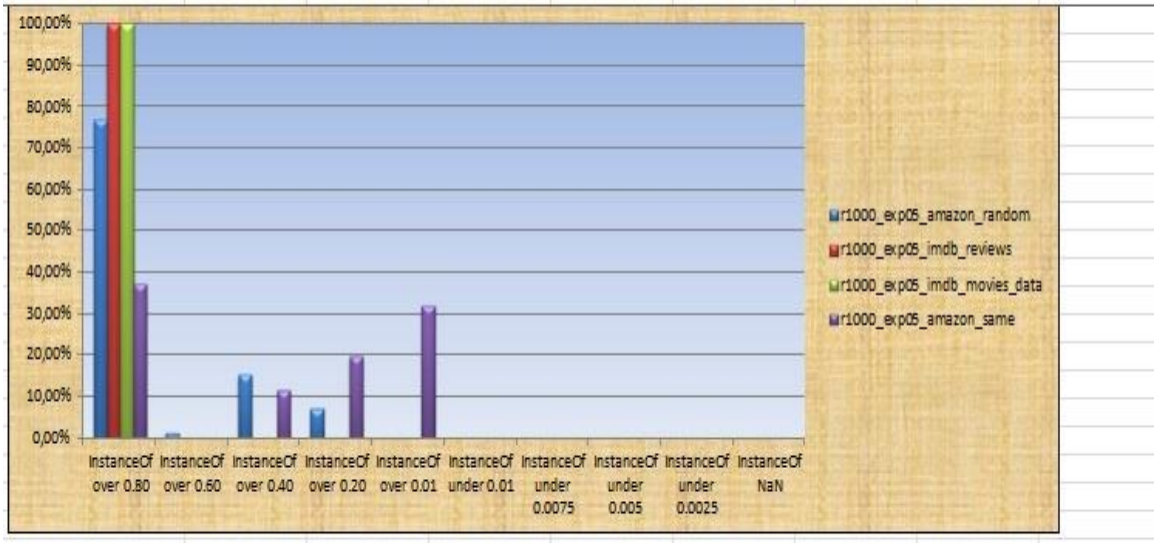
QUANTITATIVE GROUPING AT PROBABILITY LEVELS - PERCENTAGE RATIO COMPARISON - AMAZON RANDOM MOVIES REVIEWS, IMDB MOVIES REVIEWS, IMDB MOVIES DATA, AMAZON SAME MOVIES REVIEWS

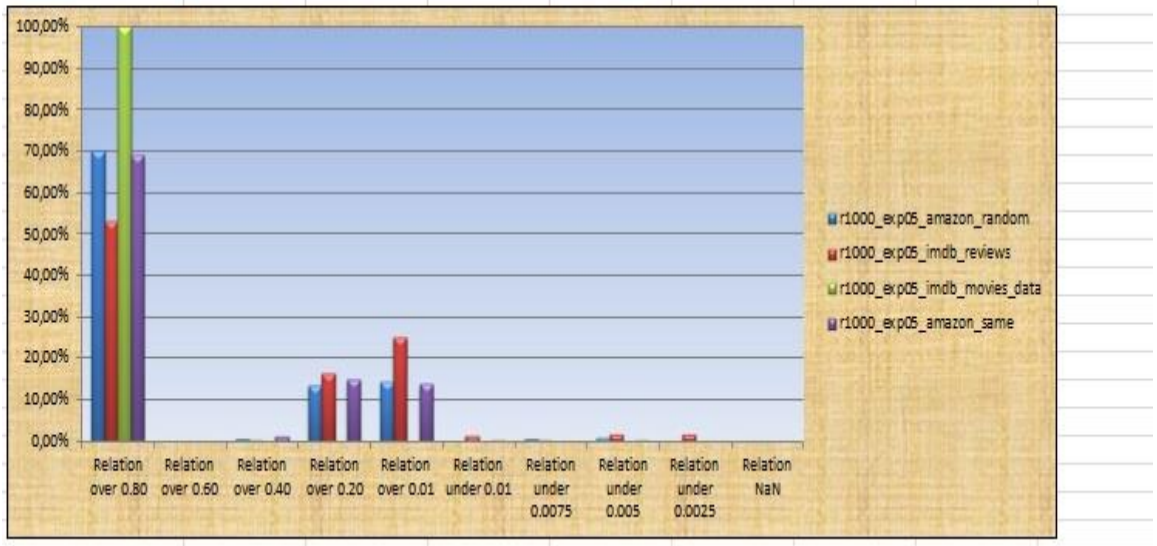






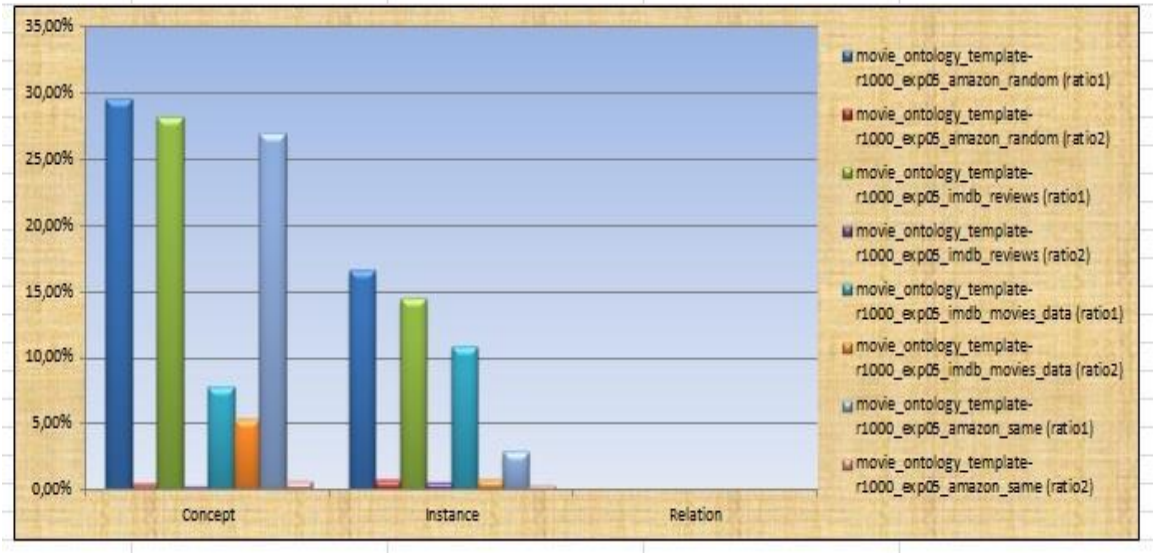
QUANTITATIVE GROUPING AT PROBABILITY LEVELS - PERCENTAGE RATIO COMPARISON - AMAZON RANDOM MOVIES REVIEWS, IMDB MOVIES REVIEWS, IMDB MOVIES DATA, AMAZON SAME MOVIES REVIEWS





SAME PRIMITIVES COMPARISON - PERCENTAGE RATIO

COMPARISON WITH MOVIE ONTOLOGY TEMPLATE - AMAZON RANDOM MOVIES REVIEWS, IMDB MOVIES REVIEWS, IMDB MOVIES DATA, AMAZON SAME MOVIES REVIEWS



SAME PRIMITIVES COMPARISON - PERCENTAGE RATIO

COMPARISON WITH MOVIE ONTOLOGY TEMPLATE - AMAZON RANDOM MOVIES REVIEWS, IMDB MOVIES REVIEWS, IMDB MOVIES DATA, AMAZON SAME MOVIES REVIEWS - DELETED PRIMITIVES

Concept, Instance with rating under 0.01 deleted.