ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΤΜΗΜΑΤΟΣ ΕΦΑΡΜΟΣΜΕΝΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΕΞΟΡΥΞΗ ΠΟΛΙΤΙΚΩΝ ΤΑΣΕΩΝ ΑΠΟ ΑΝΑΡΤΗΣΕΙΣ ΣΕ ΚΟΙΝΩΝΙΚΑ ΔΙΚΤΥΑ ΚΑΙ ΦΟΡΟΥΜ

Διπλωματική Εργασία

του/της

Ιωάννη Ποτόγλου

Θεσσαλονίκη, 06/2019

ΕΞΟΡΥΞΗ ΠΟΛΙΤΙΚΩΝ ΤΑΣΕΩΝ ΑΠΟ ΑΝΑΡΤΗΣΕΙΣ ΣΕ ΚΟΙΝΩΝΙΚΑ ΔΙΚΤΥΑ ΚΑΙ ΦΟΡΟΥΜ

Ιωάννης Ποτόγλου

Πτυχίο Εφαρμοσμένης Πληροφορικής, ΠΑΜΑΚ, 2015

Διπλωματική Εργασία

υποβαλλόμενη για τη μερική εκπλήρωση των απαιτήσεων του

ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΤΙΤΛΟΥ ΣΠΟΥΔΩΝ ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΠΛΗΡΟΦΟΡΙΚΗ

Επιβλέπων Καθηγητής

Ρεφανίδης Ιωάννης

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την ηη/μμ/εεεε

Ονοματεπώνυμο 1	Ονοματεπώνυμο 2	Ονοματεπώνυμο 3
Ιωάννης Ρεφανίδης	Ηλίας Σακελλαρίου	Νικόλαος Σαμαράς

Ιωάννης Ποτόγλου

Περίληψη

Σκοπός της εργασίας είναι η ανάλυση δεδομένων από κοινωνικά δίκτυα και φόρουμ με σκοπό την εξόρυξη των πολιτικών τάσεων του κοινού. Αφού γίνει η ανάλυση ακολουθεί σύγκριση με τις γνωστές δημοσκοπήσεις για να φανεί αν μπορούν να συσχετιστούν. Κατασκευάστηκε ένα πρόγραμμα με τη χρήση της γλώσσας προγραμματισμού Java το οποίο πραγματοποιεί τις εξής λειτουργίες. Αρχικά αναλύει γνωστές ιστοσελίδες όπου οι χρήστες σχολιάζουν πολιτικά άρθρα και εξάγει τις τοποθετήσεις αυτών. Στη συνέχεια αφαιρεί όλα τα σημεία στίξης και τα τελικά σίγμα, πραγματοποιεί ανάλυση της πρότασης για να εξαχθεί αποτέλεσμα για το ποιο κόμμα αναφέρεται, ενώ τέλος ακολουθεί το τελευταίο στάδιο, αυτό της ανάλυσης συναισθήματος (sentiment analysis), όπου χρησιμοποιεί ένα εργαλείο του Azure που βαθμολογεί κάθε πρόταση με βάση το πόσο “καλή” ή “κακή” είναι ως προς το εκάστοτε κόμμα. Το αποτέλεσμα είναι για κάθε πρόταση να έχουμε μια ετικέτα, ένα label δηλαδή του κόμματος που αναφέρεται καθώς και μια βαθμολογία 0-1. Όσο πιο κοντά στο 0 τόσο πιο αρνητική η πρόταση ως προς το κόμμα. Τέλος παράγει ένα τελικό αποτέλεσμα σε καθημερινή βάση, το οποίο σε βάθος χρόνου δείχνει μια ορθή τάση ως προς το εκάστοτε κόμμα.

Ευχαριστίες

Ευχαριστώ τον επιβλέποντα καθηγητή μου κ. Ιωάννη Ρεφανίδη για τη συνεχή καθοδήγηση που μου παρείχε κατά το διάστημα εκπόνησης της εργασίας μου καθώς και για τις συμβουλές του σε διάφορα τεχνικά θέματα.

ΠΕΡΙΕΧΟΜΕΝΑ

Κατάλογος Εικόνων

1 Εισαγωγή

1.1 Πρόβλημα - Σημαντικότητα του θέματος

Η εργασία αφορά μια προσπάθεια εξόρυξης της πολιτικής τάσης από διάφορα κοινωνικά δίκτυα, ιστοσελίδες και φόρουμ. Πρόκειται για μια εφαρμογή που διαβάζει τα σχόλια των χρηστών σε διάφορες ιστοσελίδες και φόρουμ, ταξινομεί τα σχόλια αυτά και τα κατανέμει με βάση τις λέξεις κλειδιά ως προς το κόμμα ή τα κόμματα στα οποία αναφέρονται. Επίσης διεξάγει ανάλυση συναισθήματος για να δείξει εάν το σχόλιο είναι καλό ή κακό ως προς το κόμμα στο οποίο αναφέρεται. Τέλος πραγματοποιεί συσχέτιση με την εκάστοτε τάση που υπάρχει στα διάφορα μέσα (δημοσκοπήσεις).

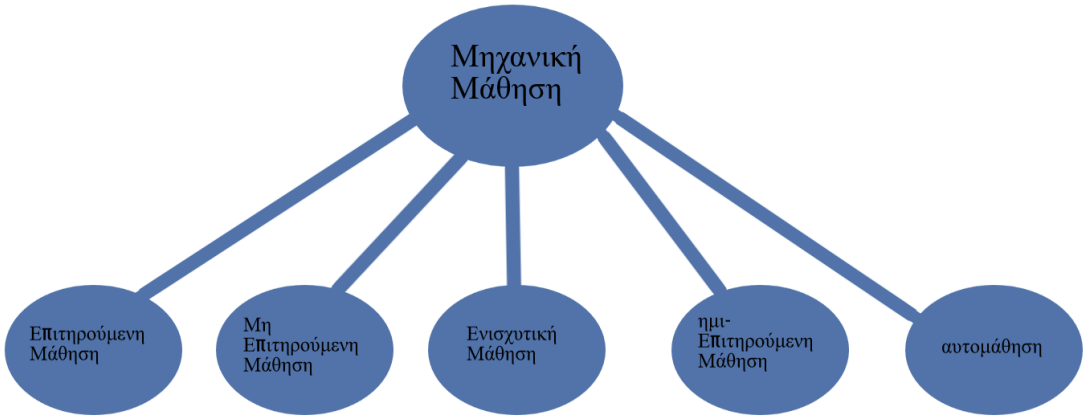
1.2 Τεχνητή Νοημοσύνη

Ο όρος Τεχνητή Νοημοσύνη αφορά τον κλάδο της πληροφορικής που ασχολείται με τη σχεδίαση προγραμμάτων - υπολογιστικών συστημάτων που επιδιώκουν να μιμηθούν την ανθρώπινη ευφυΐα. Υπάρχουν δύο είδη Τεχνητής Νοημοσύνης, η συμβολική, η οποία εξομοιώνει την ανθρώπινη ευφυΐα μέσω λογικών κανόνων, και η υποσυμβολική, η οποία πραγματοποιεί το ίδιο μέσω αριθμητικών μοντέλων. Πολλές φορές για την αντιμετώπιση ενός προβλήματος χρησιμοποιούνται ταυτόχρονα και οι δύο μέθοδοι. Αντικείμενα που αποτελούν υποκατηγορίες της Τεχνητής Νοημοσύνης ή χρησιμοποιούν μεθόδους της είναι η ρομποτική, τα νευρωνικά δίκτυα, τα λογικά προβλήματα περιορισμών, η μηχανική όραση, η Μηχανική Μάθηση κ.α. Σήμερα η Τεχνητή Νοημοσύνη θεωρείται από τα πιο εξελισσόμενα πεδία της πληροφορικής, ενώ χρησιμοποιεί περισσότερο υποσυμβολικές μεθόδους. Το αντικείμενο που μελετάται στην παρούσα εργασία είναι η ανάλυση συναισθημάτων (sentiment analysis).[1]

1.3 Μηχανική Μάθηση

Η Μηχανική Μάθηση είναι υποκατηγορία της Τεχνητής Νοημοσύνης και ουσιαστικά πρόκειται για αλγορίθμους που δίνουν τη δυνατότητα σε ένα υπολογιστικό σύστημα να μαθαίνει και να πραγματοποιεί προβλέψεις για αυτά. Οι αλγόριθμοι Μηχανικής Μάθησης αφού έχουν πρώτα εκπαιδευτεί σε ένα σύνολο δεδομένων κατασκευάζουν ένα μοντέλο στο οποίο αποδίδουν καινούριες τιμές με βάση αυτές στις οποίες εκπαιδεύτηκαν. Η Μηχανική Μάθηση είναι στενά συνδεδεμένη με την στατιστική καθώς οι μέθοδοι και οι αλγόριθμοι που χρησιμοποιεί είναι παρμένοι από τη στατιστική.

Στη Μηχανική Μάθηση υπάρχουν διάφορα είδη εκμάθησης: δέντρα αποφάσεων, κανόνες συσχέτισης, νευρωνικά δίκτυα, βαθιά μάθηση, επαγωγικός λογικός προγραμματισμός, μηχανές διανυσμάτων υποστήριξης, ομαδοποίηση, δίκτυα Bayes, ενισχυτική μάθηση, εκμάθηση με μέτρο ομοιότητας, γενετικοί αλγόριθμοι. [3]



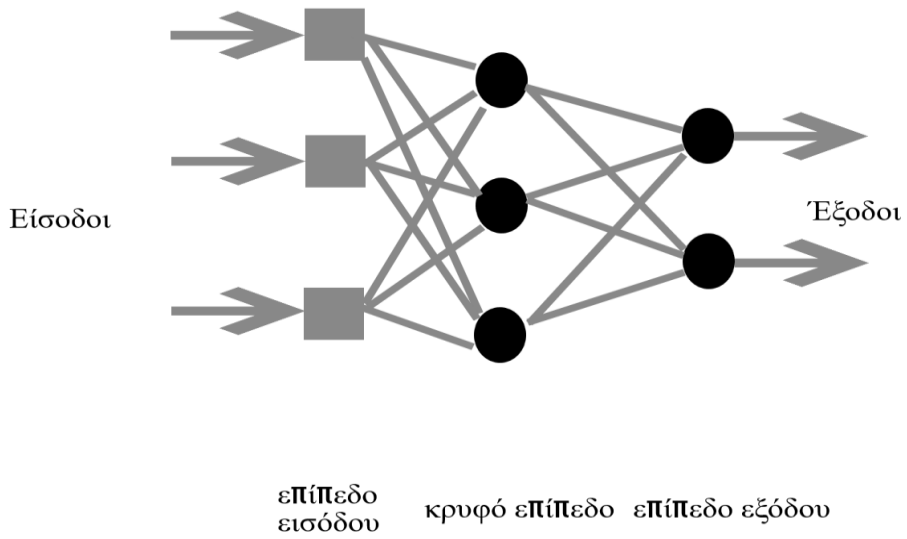
Εικόνα 1: Διάγραμμα Μηχανικής Μάθησης

1.4 Νευρωνικά Δίκτυα

Νευρωνικό δίκτυο ονομάζουμε ένα κύκλωμα - σύστημα από διασυνδεδεμένους νευρώνες. Ουσιαστικά πρόκειται για ένα σύνολο αλγορίθμων οι οποίοι προσομοιώνουν τους νευρώνες στον εγκέφαλο. Στόχος του είναι η επίλυση ενός προβλήματος μέσω της επαναλαμβανόμενης εισόδου δεδομένων στο νευρωνικό δίκτυο αφού έχει προηγηθεί η έκφανση κάποιου σφάλματος το οποίο προτρέπει την επανάληψη του υπολογισμού. Ουσιαστικά πρόκειται για πολλούς κόμβους οι οποίοι συνδέονται μεταξύ τους και ονομάζονται νευρώνες ο κάθε ένας από τους οποίους δέχεται για είσοδο δεδομένα είτε από κάποιον άλλο νευρώνα είτε από εξωτερικό παράγοντα. Με τη σειρά του ο νευρώνας αυτός πραγματοποιεί έναν υπολογισμό με τις εισόδους που δέχτηκε και παράγει μια έξοδο, συνέχεια η έξοδος αυτή μπορεί είτε να είναι είσοδος σε κάποιον άλλο νευρώνα ή να απελευθερώνεται στο περιβάλλον.

Υπάρχουν τρία είδη νευρώνων: οι νευρώνες εισόδου, οι νευρώνες εξόδου και οι υπολογιστικοί νευρώνες. Οι νευρώνες εισόδου βρίσκονται ανάμεσα στις εισόδους από το περιβάλλον και τους υπολογιστικούς νευρώνες και δεν πραγματοποιούν κανέναν υπολογισμό. Οι νευρώνες εξόδου παράγουν το αποτέλεσμα στο περιβάλλον ενώ οι υπολογιστικοί νευρώνες κάνουν τους πραγματικούς υπολογισμούς. Το αποτέλεσμα των υπολογιστικών νευρώνων χρησιμοποιείται σαν όρισμα στη συνάρτηση ενεργοποίησης η οποία υλοποιείται εσωτερικά σε κάθε κόμβο.

Αυτό που έχει ιδιαίτερη σημασία σε ένα νευρωνικό δίκτυο είναι η ικανότητα του να μαθαίνει. Ο τρόπος που το πραγματοποιεί αυτό είναι μέσω της βελτίωσης των αποτελεσμάτων που παράγει. Μαζί με τα αποτελέσματα των εξόδων του νευρωνικού δικτύου παράγεται και ένα συνολικό σφάλμα το οποίο επιτρέπει την αλλαγή των δεδομένων εισόδου με σκοπό τη βελτίωσή του. [9]



Εικόνα 2: Νευρωνικό Δίκτυο

1.5 Ανάλυση Συναισθημάτων (Sentiment Analysis)

Η ανάλυση συναισθημάτων είναι υποκατηγορία της Τεχνητής Νοημοσύνης και αναφέρεται στην επεξεργασία φυσικής γλώσσας και ανάλυσης κειμένου με σκοπό την εξόρυξη και μελέτη πληροφορίας. Μια βασική αρχή στην ανάλυση συναισθημάτων είναι η ταξινόμηση των δεδομένων ανάλογα με το είδος έκφρασης του κειμένου, συνήθως των ξεχωριστών προτάσεων (θετικό, αρνητικό, ουδέτερο). Συνήθως αναλύονται ξεχωριστά οι λέξεις και τα συμφραζόμενα μιας πρότασης σαν θετικές ή αρνητικές για να δώσουν ένα πόρισμα για ολόκληρη την πρόταση.

Υπάρχουν τρεις κατηγορίας ανάλυσης συναισθημάτων: στατιστικές μέθοδοι, υβριδικές και βασισμένες σε γνώση (knowledge based).

Knowledge based μέθοδοι: Οι μέθοδοι βασισμένοι στη γνώση κατηγοριοποιούν το κείμενο μέσω συγκεκριμένων λέξεων που δηλώνουν συναίσθημα (χαρούμενος, λυπημένος, νευριασμένος).

Στατιστικές μέθοδοι: Οι στατιστικές μέθοδοι βγάζουν συμπέρασμα μέσω Μηχανικής Μάθησης π.χ. Λανθάνουσα σημασιολογική ανάλυση (latent semantic analysis), support-vector machines, bag of words, Pointwise Mutual Information. Κάποιες μέθοδοι προσπαθούν να διακρίνουν αυτόν που νιώθει το συναίσθημα από αυτόν που το δέχεται.

Υβριδικές μέθοδοι: Οι υβριδικές μέθοδοι χρησιμοποιούν και Μηχανική Μάθηση αλλά και βασισμένα σε γνώσεις μεθόδους.

Είναι δύσκολη η εκπλήρωση 100% του αποτελέσματος όπως θα έκρινε ένας άνθρωπος σχετικά με την ορθότητα των αποτελεσμάτων όμως συνήθως οι μετρήσεις των αλγορίθμων είναι πολύ καλές κοντά στο 70-80%. Συνήθεις προβλήματα αφορούν προτάσεις που περιέχουν σαρκασμό και Αστεϊσμούς με αποτελέσματα πολλά από αυτά να βγάζουν λανθασμένο αποτέλεσμα.[2]

1.6 Ανάλυση Παλλινδρόμησης (regression analysis)

Η ανάλυση παλινδρόμησης είναι ένα σύνολο από στατιστικά μοντέλα τα οποία υπολογίζουν τη συσχέτιση ανάμεσα σε μια εξαρτημένη μεταβλητή και σε μία ή περισσοτέρες ανεξάρτητες μεταβλητές. Υπάρχουν διάφορα είδη ανάλυσης παλινδρόμησης με την πιο κλασσική να είναι η γραμμική παλινδρόμηση όπου δημιουργείται μια γραμμή, πάνω στην οποία ταιριάζουν περισσότερο τα δεδομένα βάση μαθηματικών κριτηρίων. Η ανάλυσης παλινδρόμης χρησιμοποιείται κυρίως για δύο λόγους. Πρώτον για πρόβλεψη όπου το πεδίο ταιριάζει με τη Μηχανική Μάθηση και δεύτερον για σχέσεις αιτιότητας ανάμεσα σε εξαρτημένες και ανεξάρτητες μεταβλητές. [7]

1.7 Web Scraping

Η όλη διαδικασία ξεκινάει με μια έννοια που ονομάζεται web scraping. Ουσιαστικά πρόκειται για μια διαδικασία κατά την οποία αναλύεται ο html κώδικας μίας ιστοσελίδας έτσι ώστε να παραχθεί πληροφορία. Αν και μπορεί ο οποιοσδήποτε να πραγματοποιήσει web scraping μέσω εφαρμογής που μπορεί να φτιάξει, ο όρος αυτός συνήθως αναφέρεται σε συγκεκριμένα προγράμματα ή bots που πραγματοποιούν web crawling. Καθώς το html σαν πρότυπο είναι ανοιχτού κώδικα είναι πολύ εύκολο για ένα πρόγραμμα αφού επισκεφτεί μια ιστοσελίδα να μπορεί να "δει" εύκολα όλο το περιεχόμενό της και έτσι μέσω συγκεκριμένης μεθοδολογίας που θα ακολουθήσει να εξάγει την πληροφορία που θέλει. Θα έλεγε κανείς πως όλη η διαδικασία είναι σαν να εξάγουμε πληροφορία από ένα .txt αρχείο. Υπάρχουν περιπτώσεις όπου μια ιστοσελίδα αποτρέπει μια εφαρμογή από το να διαβάσει το περιεχόμενο της μέσω διάφορων μεθόδων. Σε αυτή την περίπτωση υπάρχουν ειδικοί scrapers οι οποίοι εφαρμόζουν τεχνικές computer visioning αλλά και επεξεργασίας φυσικής γλώσσας (natural language processing) για να προσομοιώσουν το χρήστη και να εξάγουν την πληροφορία που επιθυμούν. [5]

Στην περίπτωση της διπλωματικής αυτής εργασίας η πληροφορία αναφέρεται κυρίως στα σχόλια χρηστών σε ιστοσελίδες εφημερίδων και πιο συγκεκριμένα σε άρθρα που αναφέρονται σε κόμματα που ανήκουν στη Βουλή. Παίρνοντας σαν κείμενο τα σχόλια των χρηστών "χτίζουμε" μια βάση με πληροφορία σχετική με τα κόμματα σε καθημερινή βάση έτσι ώστε αργότερα να αναλυθούν τα σχόλια αυτά σχετικά με το κατά πόσο "καλή" ή "κακή" εικόνα έχει το εκάστοτε κόμμα από τους χρήστες στο διαδίκτυο. Η όλη διαδικασία στην εφαρμογή που πραγματοποιεί το scraping πραγματοποιείται μέσω της βιβλιοθήκης jsoup στη γλώσσα Java.

1.8 Σκοπός - Στόχοι

Ο σκοπός της εργασίας είναι η εξαγωγή συμπεράσματος μέσω scrapping ιστοσελίδων που περιέχουν σχόλια χρηστών έτσι ώστε να παραχθούν αρκετά σχόλια και αργότερα να χρησιμοποιηθεί (μέσω της χρήσης του Azure) χρήση αλγορίθμων ανάλυσης συναισθημάτων έτσι ώστε να παρατηρηθεί εάν τα σχόλια για το κάθε κόμμα είναι θετικά ή αρνητικά.

1.9 Ερωτήματα - Υποθέσεις

Ουσιαστικά πρόκειται για μια προσπάθεια, σε οποιοδήποτε βαθμό αυτό είναι εφικτό, εξαγωγής ενός συμπεράσματος για το εάν ταιριάζουν οι διάφορες δημοσκοπήσεις και οι πολιτικές τάσεις του κοινού που παρουσιάζονται στα μέσα σε σύγκριση με την εξόρυξη της τάσης αυτής από σχόλια σε διάφορες ιστοσελίδες (κυρίως εφημερίδων) αλλά και φόρουμ. Λαμβάνεται υπόψιν φυσικά η χρήση του διαδικτύου από τις μικρότερες κυρίως ηλικιακές ομάδες.

2 Μεθοδολογία

Στο κεφάλαιο αυτό περιγράφεται λεπτομερώς η διαδικασία που ακολουθείται για την απόκτηση των απαραίτητων δεδομένων, της ανάλυσης τους αλλά και της αποστολής τους με συγκεκριμένο τρόπο στο Azure για μελέτη και απόκτηση του απαραίτητου αποτελέσματος όσον αφορά το συναίσθημα που αυτά εκφράζουν. Πραγματοποιήθηκε η ίδια διαδικασία για όλα τα κόμματα της Βουλής μέχρι και τις αρχές Ιουλίου 2019 ενώ από τον Ιούλιο και μετά έγινε η μελέτη με τα καινούρια κόμματα που μπήκαν στη Βουλή μετά τις εκλογές.

Ακολουθεί αναλυτική αναφορά στα εργαλεία και τις βιβλιοθήκες που χρησιμοποιήθηκαν καθώς και παραδείγματα κώδικα.

2.1 Εργαλεία και βιβλιοθήκες

Στα πλαίσια της διπλωματικής εργασίας χρησιμοποιήθηκαν διάφορα εργαλεία-βιβλιοθήκες τα οποία βοήθησαν στην απόκτηση των απαραίτητων αποτελεσμάτων, στην ανάλυση τους και στην έκβαση του επιθυμητού αποτελέσματος.

2.1.1 Γλώσσα προγραμματισμού Java

Η χρήση της γλώσσας προγραμματισμού Java πραγματοποιήθηκε στα πλαίσια της αντικειμενοστρεφούς σχεδίασης του προγράμματος καθώς η γλώσσα αυτή ενδείκνυται για συγκεκριμένου τύπου εφαρμογές. Αρχικός σκοπός ήταν η χρήση της γλώσσας Python η οποία θεωρείται η ιδανικότερη για χρήση εφαρμογών για νευρωνικά δίκτυα και γενικά για Τεχνητή Νοημοσύνη αλλά λόγω της χρήσης του Azure για την ανάλυση και την παραλαβή των αποτελεσμάτων δεν κρίθηκε απαραίτητο.

2.1.2 Βιβλιοθήκη Jsoup

Η jsoup είναι μια βιβλιοθήκη της γλώσσας προγραμματισμού Java η οποία χρησιμοποιείται σε εφαρμογές που πραγματοποιούν εξόρυξη πληροφορίας από μια σελίδα HTML. Παρέχει API (διεπαφή προγραμματισμού εφαρμογών) για ανάλυση και εξόρυξη πληροφορίας από το HTML καθώς στην ουσία χρησιμοποιεί τις ήδη υπάρχουσες μεθόδους της Javascript για διαχείριση του DOM (Document Object Model). Οι συνήθεις εργασίες με τη βιβλιοθήκη αυτή είναι οι εξής:[4]

1. Ανάλυση του κώδικα HTML μέσω τύπου string ή τύπου URL.
2. Εύρεση και εξόρυξη πληροφορίας κυρίως μέσω CSS selectors
3. Διαχείριση των στοιχείων (elements) της ιστοσελίδας, των attributes των στοιχείων αυτών αλλά και του κειμένου μέσα στα στοιχεία.

Μπορεί και χρησιμοποιείται με όλα τα είδη HTML (ακόμη και εάν έχει γραφτεί κώδικας που αποτρέπει την εξόρυξη πληροφορίας από αυτό).

Είναι ανοιχτού κώδικα ενώ κάνει χρήση του MIT License. Ο κώδικας βρίσκεται στην ιστοσελίδα Github.

2.1.2.1 Παραδείγματα χρήσης της βιβλιοθήκης

Σύνδεση σε ιστοσελίδα :

```
String url = ‘www.XXXXXXX.gr’;
Document doc = Jsoup.connect(url);
```

Χρήση όλων των links σε μια ιστοσελίδα

```
Element body = doc.getElementByTag("body");
Elements links = body.getElementsByTag("a");
for (Element link : links) {
  String linkHref = link.attr("href");
  String linkText = link.text();
}
```

Τροποποίηση του κειμένου των στοιχείων

```
Element id= doc. getElementById("test"); // <div id=”test”></div>
id.html("<p>insert text</p>"); // <div id=”test”>insert text</div>
```

2.1.3 Microsoft Azure

Το Azure είναι μια cloud πλατφόρμα της εταιρείας Microsoft το οποίο περιέχει διαφόρων ειδών υπηρεσίες που εξυπηρετούν ιδιώτες ή άλλες εταιρείες. Διαθέτει πολλαπλούς διακομιστές σε διάφορες χώρες του κόσμου για καλύτερη χρήση των υπηρεσιών του. Η χρήση των cloud υπηρεσιών είναι όλο και πιο συχνή με την πάροδο του χρόνου με αποτέλεσμα προϊόντα όπως το Azure αλλά και διάφορα άλλα να γίνονται όλο και πιο σημαντικά. Κάποιες από τις κατηγορίες υπηρεσιών του Azure είναι συνοπτικά οι εξής: [6]

1. Τεχνητή Νοημοσύνη και Μηχανική Μάθηση
2. Analytics
3. Βάσεις Δεδομένων
4. Εργαλεία Προγραμματιστών
5. DevOps
6. Internet Of Things
7. Δίκτυα
8. Ασφάλεια

Στην εργασία αυτή γίνεται η χρήση του Azure και πιο συγκεκριμένα της ανάλυσης συναισθημάτων μέρος των Cognitive Services που ανήκουν στην κατηγορία της Τεχνητής Νοημοσύνης.

2.1.3.1 Cognitive Services - Ανάλυση Συναισθημάτων

Το κομμάτι των cognitive services το οποίο είναι μέρος των υπηρεσιών Τεχνητής Νοημοσύνης καταφέρνει να δημιουργεί “έξυπνες” εφαρμογές με χρήση απλώς ενός API. Οι Υπηρεσίες χωρίζονται σε πέντε κατηγορίες :

1. Αποφάσεις
2. Γλώσσα
3. Ομιλία
4. Δικτυακή αναζήτηση
5. “Όραμα”

Το τμήμα που αφορά την ανάλυση συναισθημάτων ανήκει στη Γλώσσα και ουσιαστικά πρόκειται για χρήση ενός API το οποίο επιτρέπει τη βαθμολογία των προτάσεων σε “καλές” ή “κακές” αναλόγως το είδος λόγου που περιέχουν (0-1). Χρησιμοποιούν έναν αλγόριθμο Μηχανικής Μάθησης ο οποίος πραγματοποιεί τη βαθμολογία αυτή. Βαθμολογίες κοντά στο 1 δείχνουν θετική έκφανση λόγου ενώ κοντά στο 0 το αντίθετο. Το μοντέλο χρησιμοποιεί ένα μεγάλο κομμάτι δεδομένων που αφορούν συναισθήματα για να βγάλει συμπεράσματα για τις προτάσεις. Χρησιμοποιούνται διάφορες τεχνικές κατά τη διαδικασία όπως επεξεργασία κειμένου, ανάλυση λόγου, καθώς και τοποθέτηση και συσχέτιση λέξεων.

2.2 Ανάλυση και Σχεδίαση

Ο σχεδιασμός του προγράμματος όπως προαναφέρθηκε χωρίζεται σε τρία στάδια

a. Χρήση της Jsoup για την ανάλυση του HTML κώδικα των ιστοσελίδων.
b. Συλλογή των προτάσεων από τους χρήστες και πραγματοποίηση διαδικασιών όπως (αφαίρεση διπλοτύπων και αφαίρεση σημείων στίξης).
c. Συγκέντρωση του σε μορφοποίηση json και αποστολή στο Azure.
d. Παραλαβή αποτελεσμάτων με τις βαθμολογίες και συσχέτιση των αποτελεσμάτων με βάση τις δημοσκοπήσεις.

2.2.1 Ανάλυση του κώδικα των ιστοσελίδων

Η διαδικασία χωρίζεται σε 3 στάδια.

1. Σύνδεση στην Ιστοσελίδα και ανάλυση όλων των συνδέσμων που οδηγούν σε άρθρα που αφορούν την πολιτική.

```
Document doc = null;
      try {
          doc = Jsoup
                .connect(URL)
                .userAgent(
                  "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)")
                .timeout(5000).get();
      } catch (IOException e) {
         System.out.println("Couldn't connect to the site");
      }

      Elements links = doc.select("a[href]");
      for (Element link : links) {
         String temp = link.attr("href");
         if(temp.contains("article") && temp.contains("politics")){
            if(!temp.contains("http") && !temp.contains("https") && !temp.contains("www"))
               articles.add(URL + temp);
            else
               articles.add(temp);
         }
}
```

2. Ανάλυση του εκάστοτε συνδέσμου και εξόρυξη των σχόλιων από αυτό.

```
Document doc = null;

 Elements comment = null;
      System.out.println("site has: " + articles.size() + " articles");
      for(int i=0; i<articles.size(); i++) {
         try {
            doc = Jsoup.connect(articles.get(i)).get();
            comment = doc.select("div > p");
               for(Element commentss : comment)
                  comments.add(commentss.toString());

         } catch (IOException e) {
            System.out.println("Couldn't connect to " + comments.get(i));
            System.out.println();System.out.println();
         }
 }
```

3 Αποθήκευση των σχόλιων σε αρχείο.

```
try {
        writer = new PrintWriter("C://temp/ProtoThema/ProtoThema comments-" + day + "-" + month + "-" + year
+".txt","UTF-8");
      } catch (FileNotFoundException | UnsupportedEncodingException e) {
         System.out.println("Couldn't create the comments' file");
      }
      for(int i=0; i<comments.size(); i++) {
         writer.append(comments.get(i));
         writer.append("\n");
 }
```

Και οι τρεις διαδικασίες πραγματοποιούνται με τη βιβλιοθήκη Jsoup.

2.2.2 Συλλογή και επεξεργασία

Η διαδικασία αυτή αφορά τη συλλογή των δεδομένων σε αρχείο και την εκάστοτε επεξεργασία τους για την τελική αποστολή στο Azure. Η συλλογή γίνεται καθημερινά σε ξεχωριστά .txt αρχεία έτσι ώστε η πληροφορία να αφορά την κάθε μέρα ξεχωριστά. Χρησιμοποιείται η βιβλιοθήκη Ucharacter η οποία επεξεργάζεται Unicode χαρακτήρες, μ'αυτό τον τρόπο πραγματοποιείται η επεξεργασία της κάθε πρότασης ξεχωριστά έτσι ώστε να αφαιρεθούν όλα τα σημεία στίξης π.χ.

```
while((ch = reader1.read()) != -1) {
 if(Character.isUpperCase(ch))
   ch = Character.toLowerCase(ch);
```

```
if(Character.getName(ch).equals("EXCLAMATION MARK") || Character.getName(ch).equals("DIAERESIS") ||
Character.getName(ch).equals("QUESTION MARK") || Character.getName(ch).equals("SEMICOLON") ||
Character.getName(ch).equals("COMMA") || Character.getName(ch).equals("COLON") ||
Character.getName(ch).equals("QUOTATION MARK") || Character.getName(ch).equals("LEFT SQUARE BRACKET") ||
Character.getName(ch).equals("RIGHT SQUARE BRACKET") || Character.getName(ch).equals("LEFT PARENTHESIS") ||
Character.getName(ch).equals("RIGHT PARENTHESIS") || Character.getName(ch).equals("RIGHT CURLY BRACKET") ||
Character.getName(ch).equals("LEFT CURLY BRACKET") || Character.getName(ch).equals("GRAVE ACCENT") ||
Character.getName(ch).equals("ACUTE ACCENT") ||  Character.getName(ch).equals("APOSTROPHE") ||
Character.getName(ch).equals("SOLIDUS") || Character.getName(ch).equals("VERTICAL LINE") ||
Character.getName(ch).equals("REVERSE SOLIDUS") || Character.getName(ch).equals("LEFT-POINTING DOUBLE ANGLE
QUOTATION MARK") || Character.getName(ch).equals("RIGHT-POINTING DOUBLE ANGLE QUOTATION MARK") ||
Character.getName(ch).equals("RIGHT DOUBLE QUOTATION MARK") || Character.getName(ch).equals("LEFT DOUBLE
QUOTATION MARK") || Character.getName(ch).equals("RIGHT SINGLE QUOTATION MARK") ||
Character.getName(ch).equals("LEFT SINGLE QUOTATION MARK") )
 ch = 32;
if(Character.getName(ch).contains("WITH TONOS"))
 {
 ch = UCharacter.getCharFromName(Character.getName(ch).substring(0 , Character.getName(ch).length() - 11));
  }
   if(ch == 962)
     ch = 963;
     writer1.append((char)(ch));
```

Έτσι αποστέλουμε “καθαρές” προτάσεις χωρίς σημεία στίξης στο Azure για ανάλυση.

2.2.3 Συγκέντρωση σε json και αποστολή στο Azure

Στη φάση αυτή πραγματοποιείται η συγκέντρωση των δεδομένων σε json αρχείο και στη συνέχεια η αποστολή στο Microsoft Azure.

1. Διαδικασία συγκέντρωσης σε Json: Στη διαδικασία αυτή εξάγονται όλες οι προτάσεις που αφορούν το κάθε κόμμα. Κατασκευάζεται ένα jsonArray το οποίο παίρνει 20 json objects κάθε φορά.

```
int size = 0;
      if(finalResultList.size() >= jsonSize)
         size = jsonSize;
      else
         size = finalResultList.size();

      int timesSend = 0;
      for(int j=0; j<finalResultList.size(); j++)
         timesSend++;
      timesSend = timesSend /= jsonSize;

      for(int k=0; k<=timesSend;k++) {//πόσες φορές να τα στείλει σε 20άδες
         //if(k != 2)
            //continue;
         JSONArray jsonArray = new JSONArray();//τελικό json που περιέχει ότι data χρειαζόμαστε πριν τα στείλουμε για
ανάλυση στο azure
         for(int i=1; i<=size; i++)
         {
            int temp = i +(k * size);
            if(k == timesSend)
```

```
            {
               if(temp >= finalResultList.size())
                  break;
            }
            //System.out.println(finalResultList.get(i).substring(0,finalResultList.get(i).length() - 4));
            //System.out.println(finalResultList.get(i).substring(finalResultList.get(i).length() - 2));

JSONObject json = new JSONObject();
            json.put("language", language);
            json.put("id",String.valueOf(temp));
            json.put("text", finalResultList.get(temp - 1).substring(0,finalResultList.get(temp - 1).length() - 4));
            jsonArray.put(json);
            requestList.add(json.toString());
            tagList.add(finalResultList.get(temp - 1).substring(finalResultList.get(temp - 1).length() - 2));//ξεχωριστή λίστα με
τα ταγκς της κάθε πρότασης

         }
         JSONObject jsonFinal = new JSONObject();
         jsonFinal.put("Documents",jsonArray);
```

Τα αντικείμενα που αποστέλονται είναι της μορφής

{"language":"el","id":"1","text":"xxxxxxxxxxxxx."}

Όπου language είναι η γλώσσα του κειμένου που θα ακολουθήσει, δηλαδή τα ελληνικά, id είναι ένα διακριτικό κάθε αντικειμένου που αποστέλλεται και text είναι το κείμενο που θα πρέπει να βαθμολογηθεί με βάση το συναίσθημα.

2. Για τη διαδικασία της αποστολής στο Microsoft Azure ανοίγει μια σύνδεση (http request) με ένα συγκεκριμένο url το οποίο θα δεχτεί το json. Κάθε φορά δημιουργεί ξεχωριστές ομάδες των 20 αντικειμένων json τα οποία αποστέλλει μέχρι να φτάσει στο όριο των προτάσεων για τη συγκεκριμένη ημέρα (20 είναι το όριο που δέχεται κάθε φορά το κάθε request).

```
HttpClient httpclient = HttpClients.createDefault();
         URIBuilder builder;
         try {

         writer2 = new BufferedWriter(new OutputStreamWriter(new FileOutputStream("C://temp/ProtoThema/final-
"+textFile+" comments-" + day + "-" + month + "-" + year +".txt"),"UTF-8"));
         builder = new URIBuilder("https://westeurope.api.cognitive.microsoft.com/text/analytics/v2.0/sentiment");



         URI uri = builder.build();
         HttpPost request = new HttpPost(uri);
         request.setHeader("Content-Type", "application/json");
         request.setHeader("Ocp-Apim-Subscription-Key", key);

         // Request body
         StringEntity reqEntity = new StringEntity(jsonFinal.toString());
         request.setEntity(reqEntity);

         HttpResponse response = httpclient.execute(request);//αποστολή του request στο azure
```

```
HttpEntity entity = response.getEntity();
String azureResults = EntityUtils.toString(entity);


if(response.toString().contains("Bad Request"))
{
   System.out.println("Bad Request");
   continue;
}
if (entity != null)
{
   azureResults = azureResults.substring(14, azureResults.length());
   System.out.println(azureResults);//τα αποτελέσματα απο το response
   String[] finalAzureResults = azureResults.split("}\\,");
   for(int i=0; i<jsonArray.length(); i++)
   {
      responseList.add(finalAzureResults[i]);
   }
}
int min = tagList.size();
if(requestList.size() < min)
   min = requestList.size();
if(responseList.size() < min)
   min = responseList.size();
```

Τα αποτελέσματα του Azure είναι της μορφής {"id":"1","score":0.32758620381355286}

Όπου id είναι το διακριτικό του κειμένου που στάλθηκε προηγουμένως και score είναι η βαθμολογία που δίνει το Azure για την πρόταση αυτή (πιο κοντά στο 0 αρνητικό, πιο κοντά στο 1 θετικό).

2.2.4 Παραλαβή και συσχέτιση των αποτελεσμάτων

Σε αυτή τη διαδικασία πραγματοποιείται αρχικά η παραλαβή των αποτελεσμάτων σε μορφή json από το Azure. Τα αποτελέσματα έρχονται με τη σειρά που στάλθηκαν με την κάθε πρόταση να ξεχωρίζει από το id, το διακριτικό πεδίο δηλαδή της καθεμιάς. Η αποθήκευση της κάθε πρότασης αποθηκεύεται σε αρχείο .txt όπου μηνιαία πραγματοποιείται η καταγραφή του μέσου όρου βαθμολογιών του κάθε κόμματος.

Π.χ Average Values Σύριζα 0.41 Νέα Δημοκρατία 0.40 Χρυσή Αυγή 0.35 Κίνημα Αλλαγής 0.42 ΚΚΕ 0.41 Ένωση Κεντρώων 0.40 Ποτάμι 0.35 Ανεξάρτητοι Έλληνες 0.39

Τα νούμερα δίπλα από κάθε κόμμα είναι οι μέσοι όροι των αποτελεσμάτων κάθε πρότασης που αφορούν το συγκεκριμένο κόμμα για ολόκληρο το μήνα (στο συγκεκριμένο παράδειγμα του Απριλίου).

Η συσχέτιση των αποτελεσμάτων πραγματοποιείται ανάμεσα στη μηνιαία βαθμολογία ενός κόμματος με μια αντίστοιχη δημοσκόπηση που πραγματοποιήθηκε μέσα στο μήνα αυτό και το είδος συσχέτισης που πραγματοποιείται είναι ανάλυση παλινδρόμησης (regression analysis).

Η ανάλυση παλινδρόμησης πραγματοποιήθηκε μέσω του προγράμματος excel και τα αποτελέσματα φαίνονται στην εικόνα παρακάτω.

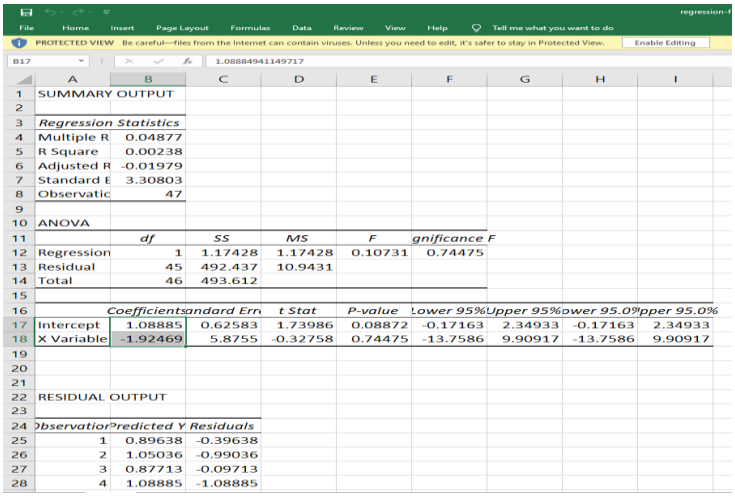
1) Το multiple R μετράει τη γραμμική συσχέτιση ανάμεσα σε δύο μεταβλητές (στην περίπτωση μας στην ένδειξη της δημοσκόπησης με τον αντίστοιχο βαθμό από το σχόλιο). Οι τιμές του είναι ανάμεσα σε -1 και 1 και όσο πιο κοντά στο 1 τόσο ισχυρότερη η συσχέτιση.

2) Ο συντελεστής R squared δείχνει πόσες τιμές βρίσκονται πάνω στην ευθεία της ανάλυσης παλινδρόμησης, δηλαδή πάνω στην ευθεία που δημιουργείται ανάλογα με τα αποτελέσματα της παλινδρόμησης.

3) Ο συντελεστής adjusted R squared είναι ο R Squared αλλά προσαρμοσμένος στον αριθμό των ανεξάρτητων μεταβλητών στην ανάλυση. Χρησιμοποιείται κυρίως όταν πραγματοποιούνται πολλαπλές αναλύσεις παλινδρόμησης.

4) Το standard error δείχνει την ακρίβεια της ανάλυσης παλινδρόμησης. Όσο χαμηλότερο το νούμερο τόσο καλύτερη η εξίσωση της ανάλυσης παλινδρόμησης. Πρόκειται ουσιαστικά για το κατά πόσο οι τιμές ξεφεύγουν από την ευθεία της παλινδρόμησης. [8]

5) Το observations δέιχνει το νούμερο των παρατηρήσεων στην ανάλυση.



	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.04877							
5	R Square	0.00238							
6	Adjusted R	-0.01979							
7	Standard E	3.30803							
8	Observatic	47							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*gnificance F*			
12	Regression	1	1.17428	1.17428	0.10731	0.74475			
13	Residual	45	492.437	10.9431					
14	Total	46	493.612						
15									
16		*Coefficients*	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
17	Intercept	1.08885	0.62583	1.73986	0.08872	-0.17163	2.34933	-0.17163	2.34933
18	X Variable	-1.92469	5.8755	-0.32758	0.74475	-13.7586	9.90917	-13.7586	9.90917
19									
20									
21									
22	RESIDUAL OUTPUT								
23									
24	*bservatior*	*redicted Y*	*Residuals*						
25	1	0.89638	-0.39638						
26	2	1.05036	-0.99036						
27	3	0.87713	-0.09713						
28	4	1.08885	-1.08885						

Εικόνα 3: Ανάλυση Παλλινδρόμη

3 Περιπτώσεις Χρήσης

Η ανάλυση των περιπτώσεων χρήσης πραγματοποιείται σε δύο στάδια. Στο πρώτο έχουμε τα αποτελέσματα με τις δημοσκοπήσεις να πραγματοποιούνται στο ίδιο χρονικό στάδιο ενω στο δεύτερο οι δημοσκοπήσεις συσχετίζονται με τα αποτελέσματα ενός μήνα πριν. Για παράδειγμα η δημοσκόπηση του Φεβρουαρίου θα συσχετιστεί με τα αποτελέσματα του Ιανουαρίου κ.ο.κ. Έτσι έχουμε μια πιο ρεαλιστική εικόνα καθώς τα αποτελέσματα αργούν να εμφανιστούν στις δημοσκοπήσεις.

3.1 Συσχέτιση δημοσκοπήσεων – αποτελεσμάτων στο ίδιο χρονικό στάδιο

3.1.1 Αναλυτική περίπτωση χρήσης για τη Νέα Δημοκρατία

Ας δούμε ένα παράδειγμα για να κατανοήσουμε τον τρόπο με τον οποίο λειτουργεί η εφαρμογή. Παίρνοντας για παράδειγμα τα δεδομένα από τα σχόλια από γνωστή ιστοσελίδα για το μήνα Ιούνιο.

Λαμβάνουμε λοιπόν ένα αρχείο το οποίο περιέχει όλα τα σχόλια των χρηστών για όλες τις αναρτήσεις που αφορούν την πολιτική για τη μέρα 16-6-2019. Το κείμενο αυτό μετά από την επεξεργασία χωρίζεται σε ξεχωριστές προτάσεις και από το περιεχόμενο της πρότασης δημιουργείται ένα διακριτικό για κάθε μία, το οποίο δείχνει σε πιο κόμμα αναφέρεται. Ο τρόπος με τον οποίο καθορίζεται το διακριτικό αυτό εξαρτάται από το περιεχόμενο της πρότασης και τις λέξεις που χρησιμοποιούνται για να φανεί σε ποιο κόμμα αναφέρεται η πρόταση. Έχει δημιουργηθεί ένα λεξικό το οποίο περιέχει όλες τις λέξεις που μας ενδιαφέρουν (ονόματα βουλευτών, ονόματα κομμάτων, χαρακτηριστικά κομμάτων) έτσι ώστε οποιαδήποτε λέξη ταιριάξει με κάποια από το λεξικό τότε αυτόματα η πρόταση αυτή αναφέρεται στο αντίστοιχο κόμμα. Έτσι λοιπόν χωρίζονται όλες οι προτάσεις σε προτάσεις που αφορούν κάποιο συγκεκριμένο κόμμα, αποστέλλονται στο azure σε json και το Azure αναλαμβάνει να πραγματοποιήσει την ανάλυση για το “’ύφος” της πρότασης και το πόσο “καλή” ή “κακή” είναι. Αφού το πραγματοποιήσει μας επιστρέφει για κάθε πρόταση την αντίστοιχη βαθμολογία και τα αποτελέσματα καταγράφονται σε αρχείο όπως φαίνεται παρακάτω.

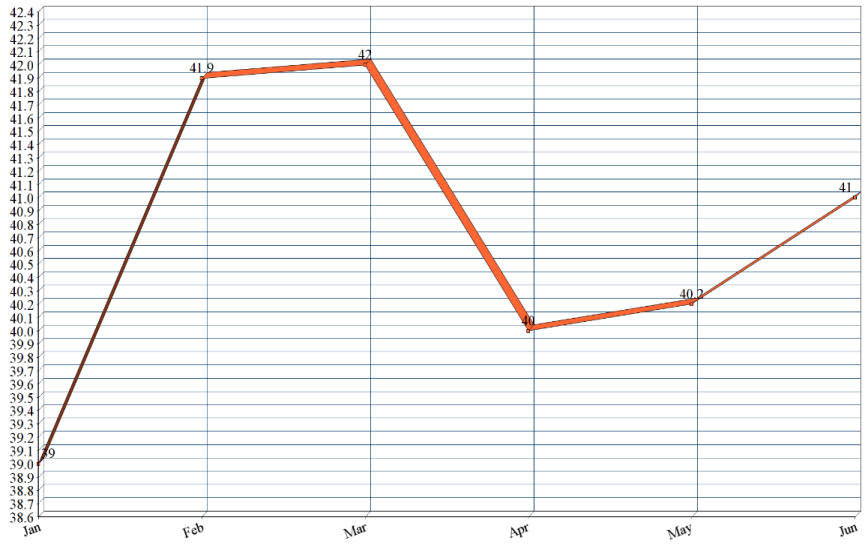
Σύριζα 0.4352001786231995 Νέα Δημοκρατία 0.43356581528981525 Χρυσή Αυγή 0.4626845568418503 Κίνημα Αλλαγής 0.43852607409159344 ΚΚΕ 0.0 Έν
Σύριζα 0.43145269714295864 Νέα Δημοκρατία 0.4102985238035552 Χρυσή Αυγή 0.3484848439693451 Κίνημα Αλλαγής 0.45424439013004303 ΚΚΕ 0.0 Ένω
Σύριζα 0.4390179365873337 Νέα Δημοκρατία 0.4024159349501133 Χρυσή Αυγή 0.40043161312739056 Κίνημα Αλλαγής 0.425618718067805 ΚΚΕ 0.0 Ένωσ
Σύριζα 0.4125159888163857 Νέα Δημοκρατία 0.4074076198041439 Χρυσή Αυγή 0.40957997739315033 Κίνημα Αλλαγής 0.38932594869818005 ΚΚΕ 0.0 Έν
Σύριζα 0.4471944421529777 Νέα Δημοκρατία 0.4192607245947185 Χρυσή Αυγή 0.43064201623201137 Κίνημα Αλλαγής 0.45590680340925854 ΚΚΕ 0.443878
Σύριζα 0.39334842562675476 Νέα Δημοκρατία 0.3643877257903417 Χρυσή Αυγή 0.404945174853007 Κίνημα Αλλαγής 0.46294060349464417 ΚΚΕ 0.17073
Σύριζα 0.4088848130777478 Νέα Δημοκρατία 0.43234909772872926 Χρυσή Αυγή 0.4047619104385376 Κίνημα Αλλαγής 0.391050564746062 ΚΚΕ 0.439393
Σύριζα 0.4110239770101464 Νέα Δημοκρατία 0.41789659976959226 Χρυσή Αυγή 0.3947368562221527 Κίνημα Αλλαγής 0.42414682110150653 ΚΚΕ 0.3275
Σύριζα 0.40176548063755035 Νέα Δημοκρατία 0.4244074327605111 Χρυσή Αυγή 0.39784422516822815 Κίνημα Αλλαγής 0.4523809552192688 ΚΚΕ 0.4374
Σύριζα 0.4452496975660324 Νέα Δημοκρατία 0.39902294391677495 Χρυσή Αυγή 0.0 Κίνημα Αλλαγής 0.40983993428559493 ΚΚΕ 0.4285714328289032 Ένω
Σύριζα 0.41301178802614624 Νέα Δημοκρατία 0.42207641350595576 Χρυσή Αυγή 0.4267967939376831 Κίνημα Αλλαγής 0.3726534843444824 ΚΚΕ 0.3903
Σύριζα 0.44039289832115175 Νέα Δημοκρατία 0.42426297594519224 Χρυσή Αυγή 0.4191713531812032 Κίνημα Αλλαγής 0.0 ΚΚΕ 0.44349993268648785 Ε
Σύριζα 0.4110124036669731 Νέα Δημοκρατία 0.442933699663947763 Χρυσή Αυγή 0.4908915162086487 Κίνημα Αλλαγής 0.38835416237513226 ΚΚΕ 0.4350
Σύριζα 0.4321545537780313 Νέα Δημοκρατία 0.4266594762985523 Χρυσή Αυγή 0.42648619413375854 Κίνημα Αλλαγής 0.3970174007117748 ΚΚΕ 0.0 Ένω
Σύριζα 0.39723254504956695 Νέα Δημοκρατία 0.44031662245591148 Χρυσή Αυγή 0.42960883677005777 Κίνημα Αλλαγής 0.0 ΚΚΕ 0.0 Ένωση Κεντρώων 0.4
Σύριζα 0.4122725874185562 Νέα Δημοκρατία 0.42029342693941935 Χρυσή Αυγή 0.17073380947113037 Κίνημα Αλλαγής 0.39263278245925903 ΚΚΕ 0.407
Σύριζα 0.4105482433111437 Νέα Δημοκρατία 0.4144122861325741 Χρυσή Αυγή 0.45485053459803265 Κίνημα Αλλαγής 0.39378511905670166 ΚΚΕ 0.4481
Σύριζα 0.40682049335971954 Νέα Δημοκρατία 0.392143942415714 26 Χρυσή Αυγή 0.41092549264431 Κίνημα Αλλαγής 0.4832775890827179 ΚΚΕ 0.0 Ένωσ
Σύριζα 0.41058384502927464 Νέα Δημοκρατία 0.4013578280806541 Χρυσή Αυγή 0.4106847941875458 Κίνημα Αλλαγής 0.41658882583890644 ΚΚΕ 0.3529
Σύριζα 0.4243145630675919 Νέα Δημοκρατία 0.4290017922376764 Χρυσή Αυγή 0.4341947615146637 Κίνημα Αλλαγής 0.41864770236942506 ΚΚΕ 0.42763
Σύριζα 0.42972131073474884 Νέα Δημοκρατία 0.4216438037427989 Χρυσή Αυγή 0.40740740299224854 Κίνημα Αλλαγής 0.0 ΚΚΕ 0.4212598502635956 Έν

Σύριζα 0.41 Νέα Δημοκρατία 0.41 Χρυσή Αυγή 0.40 Κίνημα Αλλαγής 0.41 ΚΚΕ 0.39 Ένωση Κεντρώων 0.40 Ποτάμι 0.42 Ανεξάρτητοι Έλληνες 0.43

Εικόνα 4: Βαθμολογίες μετά την ανάλυση

Ο βαθμός (0-1) δείχνει το συναίσθημα και εξάγεται ο μέσος όρος για το μήνα αυτό στο τέλος.

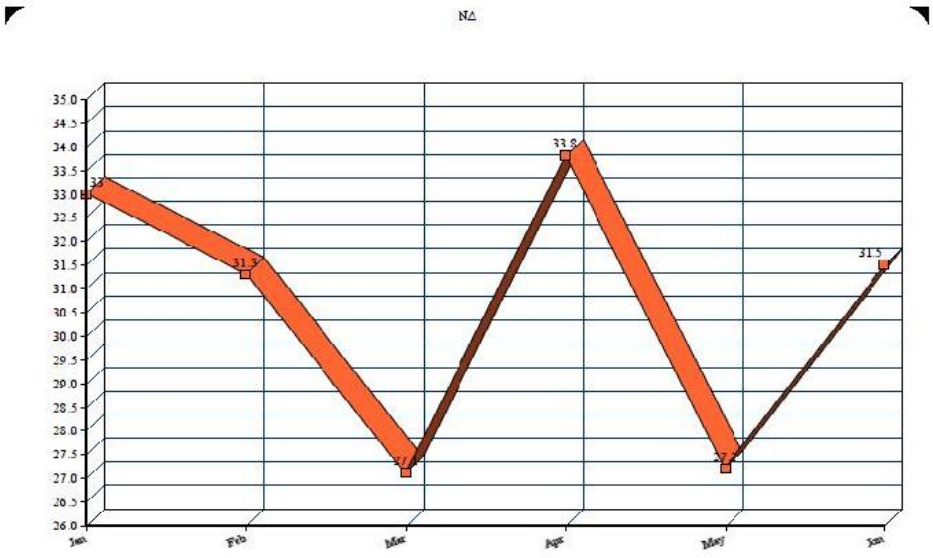
Εάν πάρουμε για το κάθε κόμμα, για κάθε μήνα που πραγματοποιούμε την ανάλυση αυτή, τον αντίστοιχο μέσο όρο τότε μπορούμε να παράξουμε ένα διάγραμμα όπως φαίνεται παρακάτω.



Εικόνα 5: ΝΔ - αποτελέσματα ανάλυσης

Το συγκεκριμένο διάγραμμα είναι για το κόμμα της Νέας Δημοκρατίας για το διάστημα Ιανουάριος - Ιούνιος.

Λαμβάνοντας αντίστοιχα τα αποτελέσματα των δημοσκοπήσεων που δημοσιεύονται για κάθε μήνα δημιουργούμε ένα αντίστοιχο διάγραμμα όπως φαίνεται παρακάτω.



Εικόνα 6: ΝΔ - αποτελέσματα δημοσκοπήσεων

Παρατηρούμε αμέσως ότι δεν υπάρχει άμεση σύγκριση της τάσης που έχει το κάθε διάγραμμα και άρα φαίνεται σαν να μην ταιριάζουν. Αυτό που έχουμε να κάνουμε στη συνέχεια είναι να πραγματοποιήσουμε ανάλυση παλινδρόμησης στα δύο αυτά διαγράμματα για να βγάλουμε τα συμπεράσματά μας.

Πραγματοποιώντας λοιπόν ανάλυση παλινδρόμησης μέσω του προγράμματος excel παρατηρούμε τις εξής τιμές.

Regression Statistics					
Multiple F	0.661513				
R Square	0.437599				
Adjusted I	0.296999				
Standard E	0.981263				
Observati	6				
ANOVA					
	df	*SS*	*MS*	*F*	*gnificance F*
Regressio	1	2.996827	2.996827	3.112369	0.152469
Residual	4	3.851506	0.962877		
Total	5	6.848333			

Εικόνα 7: ΝΔ - ανάλυσης παλινδρόμησης

Όσον αφορά τα αποτελέσματα, παίρνοντας με τη σειρά τους δείκτες:

1) Multiple R 0.661513: Όσο πιο κοντά στο 1 τόσο ισχυρότερη η συσχέτιση ανάμεσα στην ανεξάρτητη και την εξαρτημένη μεταβλητή. Στη συγκεκριμένη περίπτωση υπάρχει μια σχετική εξάρτηση.

2) R Square 0.4375: Δείχνει πόσες τιμές βρίσκονται κοντά στην ευθεία της παλινδρόμησης. Στο συγκεκριμένο παράδειγμα το 0.4375 δείχνει ότι το 43% των τιμών πέφτουν πάνω στην ευθεία παλινδρόμησης.

3) Adjusted R Square 0.2969: Χρησιμοποιείται για πολλαπλές αναλύσεις παλινδρόμησης άρα δεν έχει τόση σημασία στην περίπτωσή μας.

4) Standrad error 0.9812: Είναι μια απόλυτη τιμή που δείχνει πόσο απέχουν τα δεδομένα από τη γραμμή παλινδρόμησης.

5) Ανάλυση διακύμανσης (ANOVA) : Η ανάλυση διακύμανσης (ANOVA) δείχνει στοιχεία μεταβλητότητας στο μοντέλο παλινδρόμησης.

Df: οι βαθμοί ελευθερίας του μοντέλου

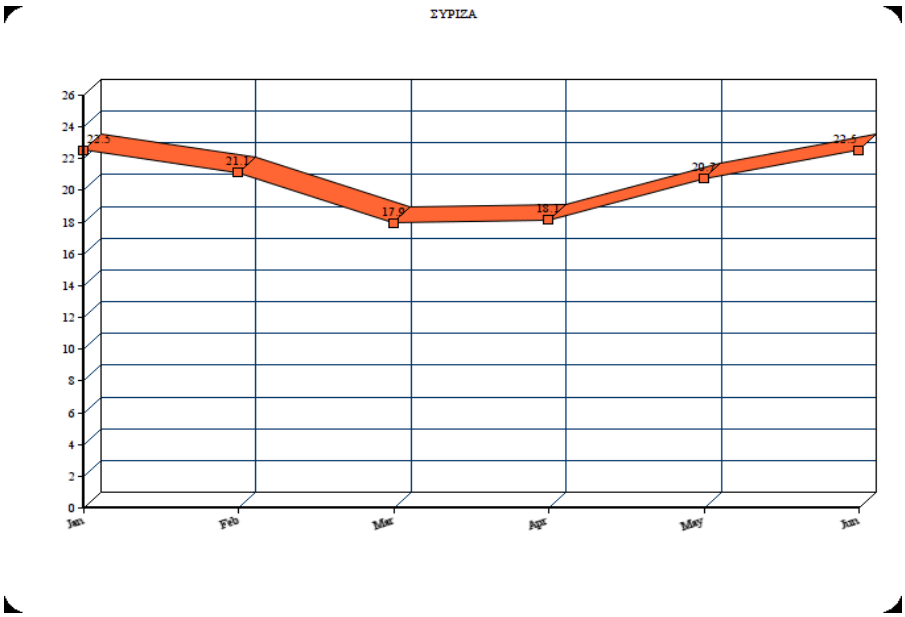
Ss: το άθροισμα των τετραγώνων, όσο μικρότερο το κατάλοιπο άθροισμα τετραγώνων (residual ss) σε σύγκριση με το συνολικό άθροισμα τόσο καλύτερο το μοντέλο.

Η χρήση γενικώς της ανάλυσης παλινδρόμησης (ANOVA) είναι σπάνια για μια απλή γραμμική παλινδρόμησης αλλά κάποιες φορές αξίζει να μελετάται. Το στοιχείο που έχει ιδιαίτερη σημασία είναι το significance F το οποίο δείχνει πόσο στατιστικά σημαντικά είναι τα αποτελέσματα. Εάν είναι κάτω από 5% τότε τα αποτελέσματα είναι πολύ καλά. Στην περίπτωσή μας που είναι στο 15% αυτό σημαίνει ότι τα αποτελέσματα δεν είναι απόλυτα ικανοποιητικά.

3.1.2 Παραδείγματα περιπτώσεων χρήσης των υπόλοιπων κομμάτων

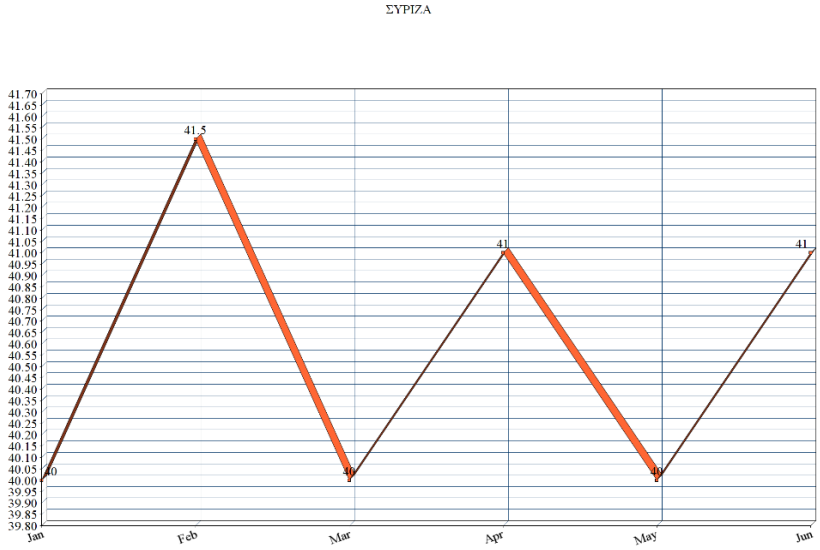
Ας δούμε και τα υπόλοιπα κόμματα για να βγάλουμε κάποιο συμπέρασμα για την εγκυρότητα των αποτελεσμάτων μας. Ας ξεκινήσουμε από το Σύριζα.

Οι δημοσκοπήσεις για το Σύριζα δείχνουν τα εξής αποτελέσματα.



Εικόνα 8: Συριζα- αποτελέσματα δημοσκοπήσεων

Η ανάλυση των σχολίων και η επιστροφή των αποτελεσμάτων από το Azure φέρνουν τα εξής αποτελέσματα



Εικόνα 9: Συριζα- αποτελέσματα ανάλυσης

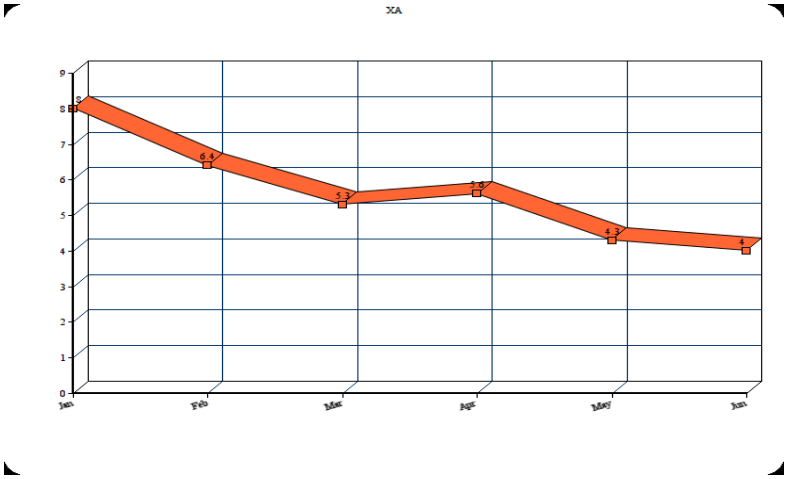
Παρατηρούμε ότι δεν υπάρχει άμεση συσχέτιση των διαγραμμάτων οπότε πραγματοποιούμαι την ανάλυση παλινδρόμησης για να δούμε τα αποτελέσματα

SUMMARY OUTPUT					
Regression Statistics					
Multiple F	0.145262				
R Square	0.021101				
Adjusted I	-0.22362				
Standard E	0.735142				
Observati	6				
ANOVA					
	df	*SS*	*MS*	*F*	*gnificance F*
Regressio	1	0.046598	0.046598	0.086223	0.78364
Residual	4	2.161735	0.540434		
Total	5	2.208333			

Εικόνα 10: Συριζα- ανάλυση παλινδρόμησης

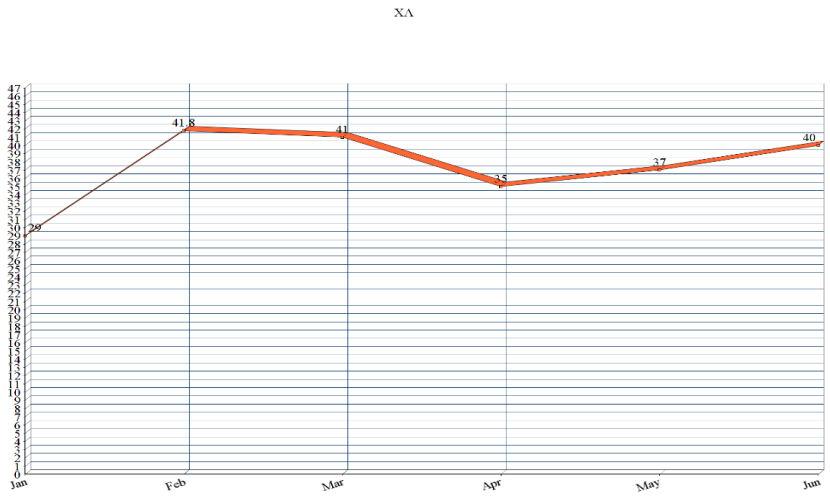
Βλέποντας και τα αποτελέσματα από την ανάλυση παλινδρόμησης παρατηρούμε ότι δεν υπάρχει ιδιαίτερη συσχέτιση των αποτελεσμάτων. Η τιμή του R είναι κοντά στο 0 και η τιμή του significance F είναι αρκετά μεγαλύτερη από 5%, στοιχεία που δείχνουν ότι τα αποτελέσματα δεν συσχετίζονται ιδιαίτερα μεταξύ τους.

Οι δημοσκοπήσεις για τη Χρυσή Αυγή δείχνουν τα εξής αποτελέσματα



Εικόνα 11: Χρυση Αυγή - αποτελέσματα δημοσκοπήσεων

Η ανάλυση των σχολίων και η επιστροφή από το Azure φέρνουν τα εξής αποτελέσματα



Εικόνα 12: Χρυση Αυγή - αποτελέσματα ανάλυσης

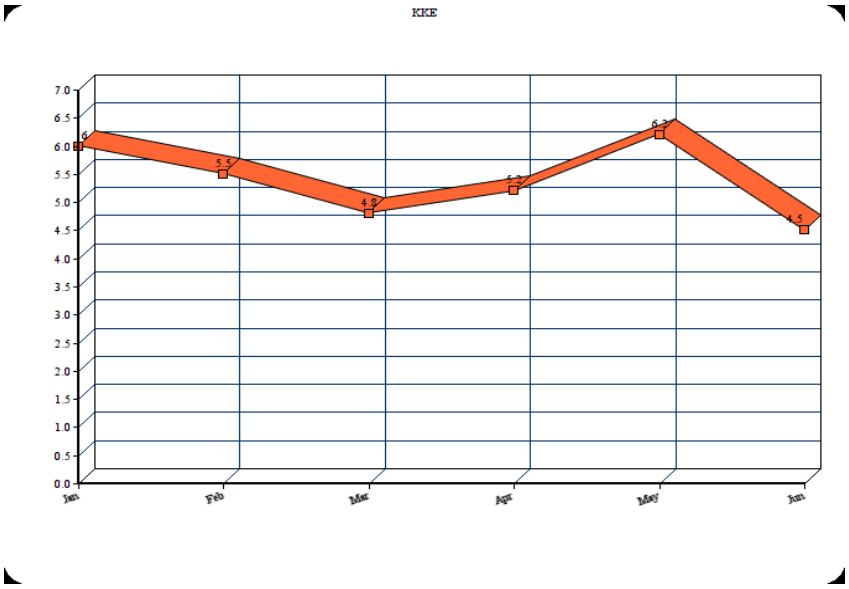
Βλέπουμε ότι υπάρχουν κάποια κοινά αλλά όχι όπως τα επιθυμούμε. Ας δούμε και την ανάλυση παλινδρόμησης.

SUMMARY OUTPUT					
Regression Statistics					
Multiple F	0.606468				
R Square	0.367804				
Adjusted I	0.209755				
Standard E	4.272548				
Observati	6				
ANOVA					
	df	*SS*	*MS*	*F*	*gnificance F*
Regressio	1	42.48134	42.48134	2.32715	0.201828
Residual	4	73.01866	18.25466		
Total	5	115.5			

Εικόνα 13: Χρυση Αυγή - ανάλυση παλινδρόμησης

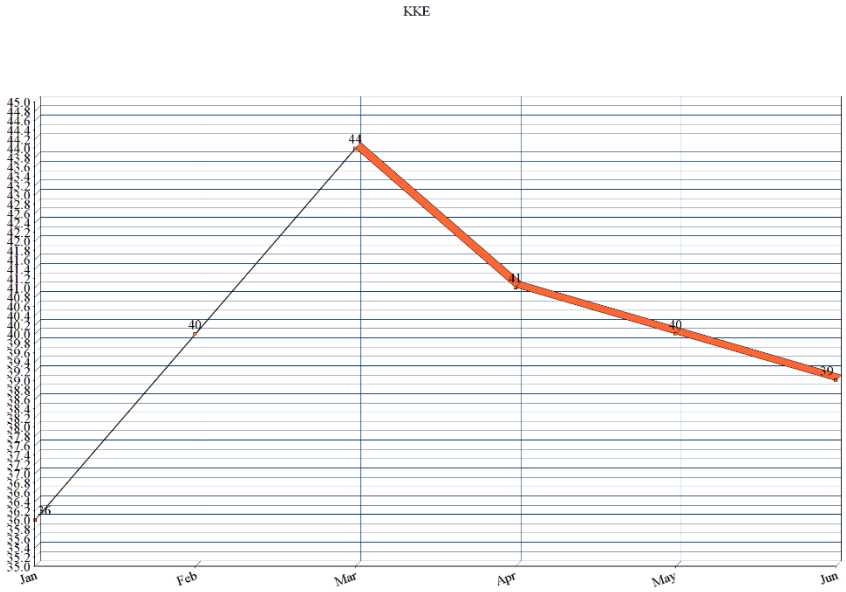
Το multiple R είναι στο 0.60 το οποίο δείχνει μια σχετικά καλή συσχέτιση ενώ η τιμή significance F βρίσκεται στο 0.2 σχετικά μακριά από το 5 % το οποίο δεν είναι τόσο καλό.

Συνεχίζοντας με το ΚΚΕ. Ας δούμε αρχικά τις δημοσκοπήσεις.



Εικόνα 14: ΚΚΕ - αποτελέσματα δημοσκοπήσεων

Η ανάλυση των σχολίων και η επιστροφή από το Azure φέρνουν τα εξής αποτελέσματα



Εικόνα 15: Χρυση Αυγή - αποτελέσματα ανάλυσης

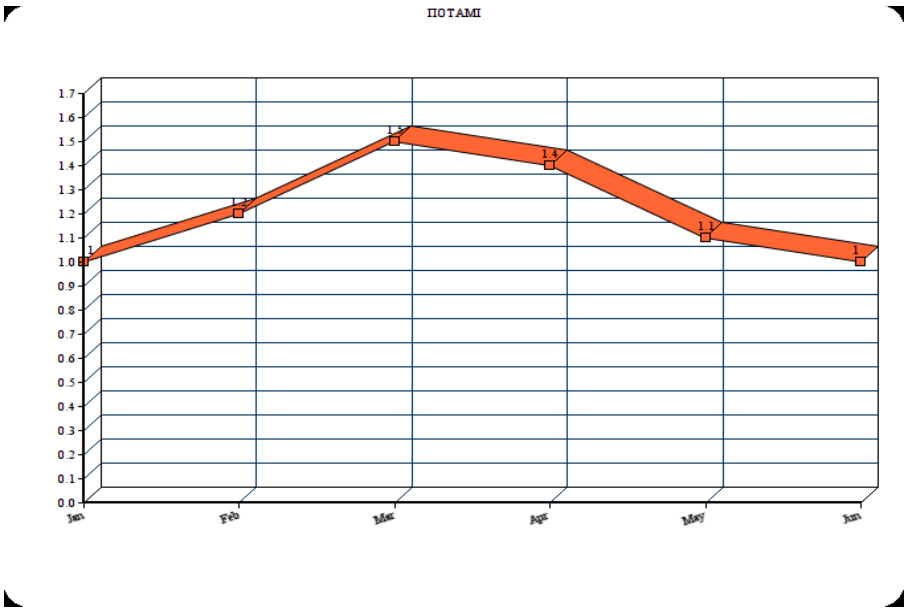
Συνεχίζοντας με τα αποτελέσματα από την ανάλυση παλινδρόμησης

SUMMARY OUTPUT					
Regression Statistics					
Multiple F	0.47263				
R Square	0.223379				
Adjusted I	0.029224				
Standard E	2.569296				
Observati	6				
ANOVA					
	df	*SS*	*MS*	*F*	*gnificance F*
Regressio	1	7.59488	7.59488	1.150516	0.343843
Residual	4	26.40512	6.60128		
Total	5	34			

Εικόνα 16: Χρυση Αυγή - ανάλυση παλινδρόμησης

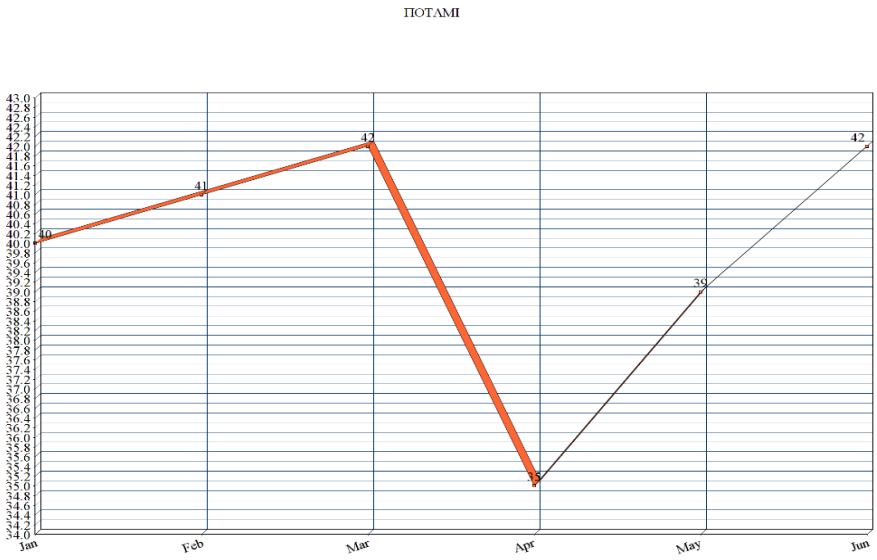
Τα στοιχεία 0.47 από το multiple R και 0.34 από το significance F δεν μας δείχνουν και ιδιαίτερα καλή συσχέτιση ανάμεσα στα δεδομένα.

Συνεχίζοντας με τα αποτελέσματα για το Ποτάμι ξεκινώντας από τις δημοσκοπήσεις.



Εικόνα 17: Ποτάμι - αποτελέσματα δημοσκοπήσεων

Η ανάλυση των σχολίων και η επιστροφή από το Azure φέρνουν τα εξής αποτελέσματα



Εικόνα 18: Ποτάμι - αποτελέσματα ανάλυσης

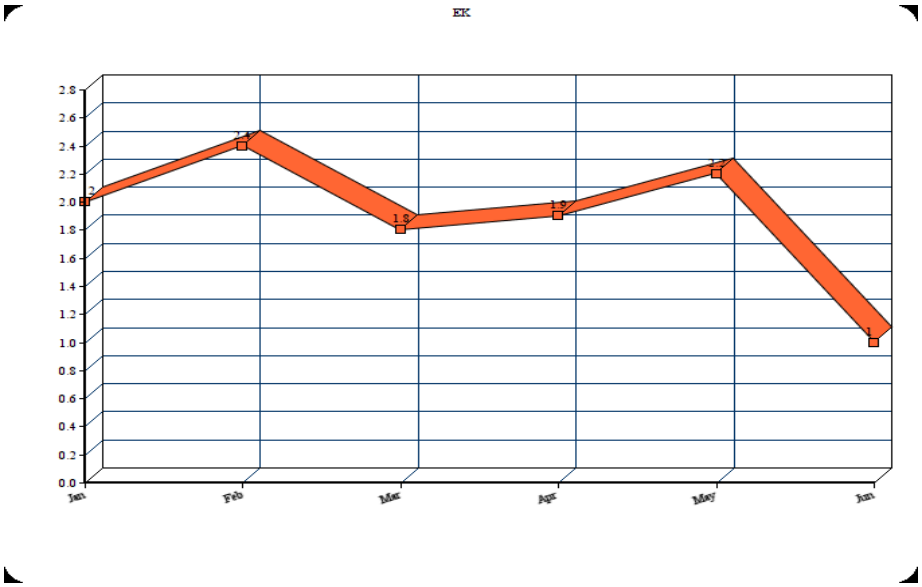
Βλέπουμε κάποια κοινά στοιχεία. Συνεχίζουμε με την ανάλυση παλινδρόμησης.

SUMMARY OUTPUT					
Regression Statistics					
Multiple F	0.252865				
R Square	0.063941				
Adjusted I	-0.17007				
Standard E	2.855086				
Observati	6				
ANOVA					
	df	*SS*	*MS*	*F*	*gnificance F*
Regressio	1	2.227273	2.227273	0.273234	0.628786
Residual	4	32.60606	8.151515		
Total	5	34.83333			

Εικόνα 19: Ποτάμι - ανάλυση παλινδρόμησης

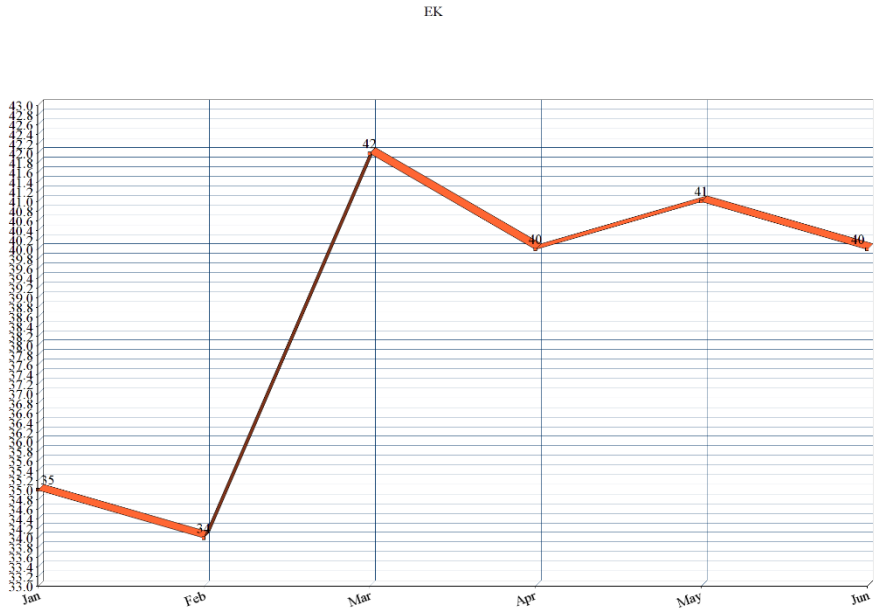
Το multiple R είναι στο 0.25 και το significance F στο 0.62 στοιχεία που δείχνουν χαμηλή συσχέτιση των δεδομένων.

Συνεχίζουμε με τα αποτελέσματα για την Ένωση Κεντρώων ξεκινώντας από τις δημοσκοπήσεις.



Εικόνα 20: ΕΚ - αποτελέσματα δημοσκοπήσεων

Η ανάλυση των σχολίων και η επιστροφή από το Azure φέρνουν τα εξής αποτελέσματα



Εικόνα 21: ΕΚ - αποτελέσματα ανάλυσης

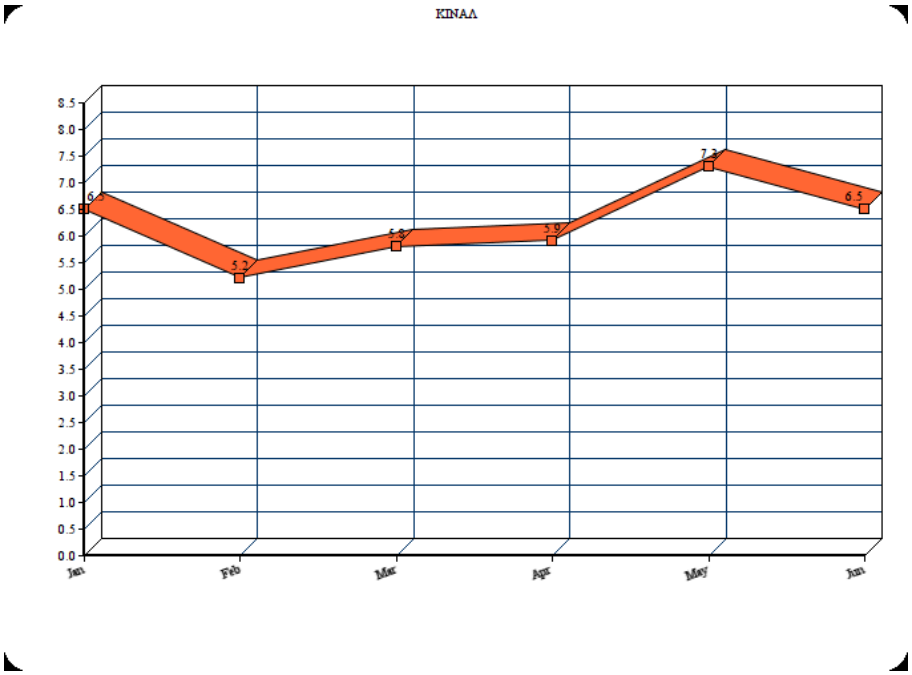
Συνεχίζουμε με τα αποτελέσματα των δημοσκοπήσεων.

SUMMARY OUTPUT					
Regression Statistics					
Multiple F	0.439448				
R Square	0.193115				
Adjusted I	-0.00861				
Standard E	3.340945				
Observati	6				
ANOVA					
	df	*SS*	*MS*	*F*	*gnificance F*
Regressio	1	10.68569	10.68569	0.957335	0.38326
Residual	4	44.64765	11.16191		
Total	5	55.33333			

Εικόνα 22: ΕΚ - ανάλυση παλινδρόμησης

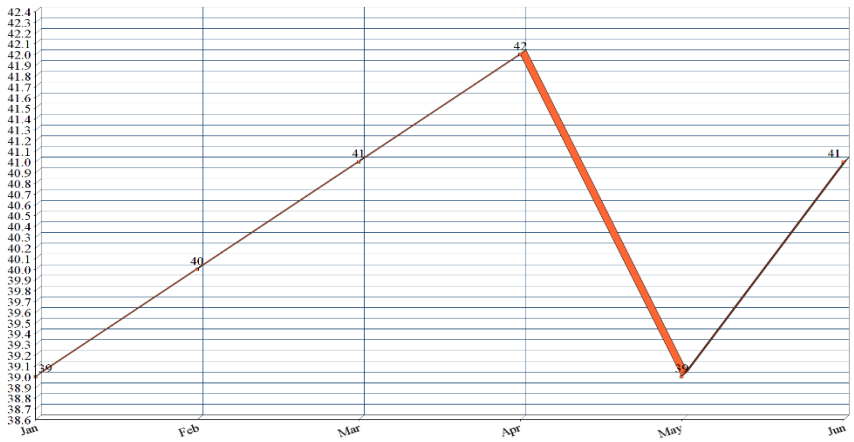
Και πάλι παρατηρούμε ότι οι συντελεστές Multiple R και significance F (0.43 και 0.38 αντίστοιχα) δεν είναι καλοί δείκτες συσχέτισης των δεδομένων.

Συνεχίζουμε με το Κίνημα Αλλαγής ξεκινώντας από τα στοιχεία δημοσκοπήσεων.



Εικόνα 23: ΚΙΝΑΛ - αποτελέσματα δημοσκοπήσεων

Η ανάλυση των σχολίων και η επιστροφή από το Azure φέρνουν τα εξής αποτελέσματα



Εικόνα 24: ΕΚ - αποτελέσματα ανάλυσης

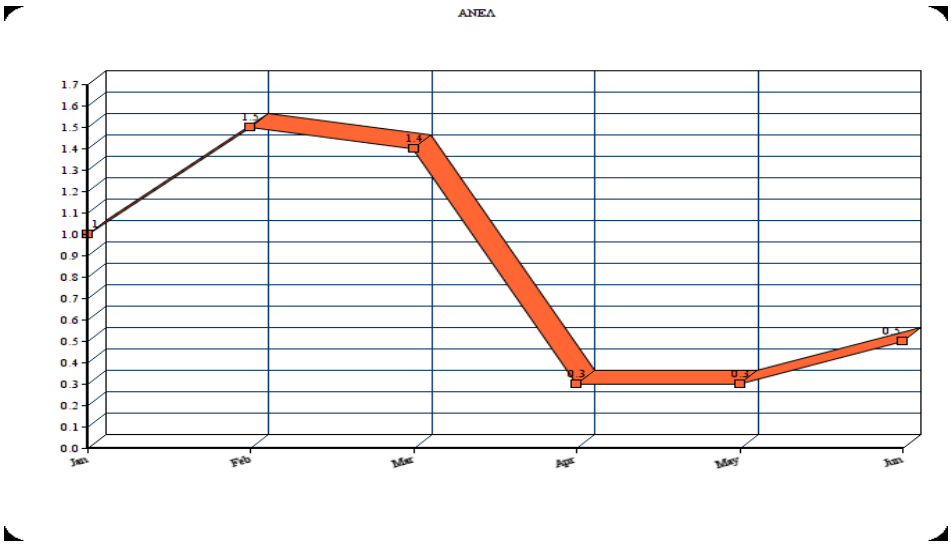
Συνεχίζουμε με τα αποτελέσματα από την ανάλυση παλινδρόμησης.

SUMMARY OUTPUT					
Regression Statistics					
Multiple F	0.466043				
R Square	0.217196				
Adjusted I	0.021495				
Standard E	1.197974				
Observati	6				
ANOVA					
	df	*SS*	*MS*	*F*	*gnificance F*
Regressio	1	1.592771	1.592771	1.109836	0.351547
Residual	4	5.740563	1.435141		
Total	5	7.333333			

Εικόνα 25: ΚΙΝΑΛ - ανάλυση παλινδρόμησης

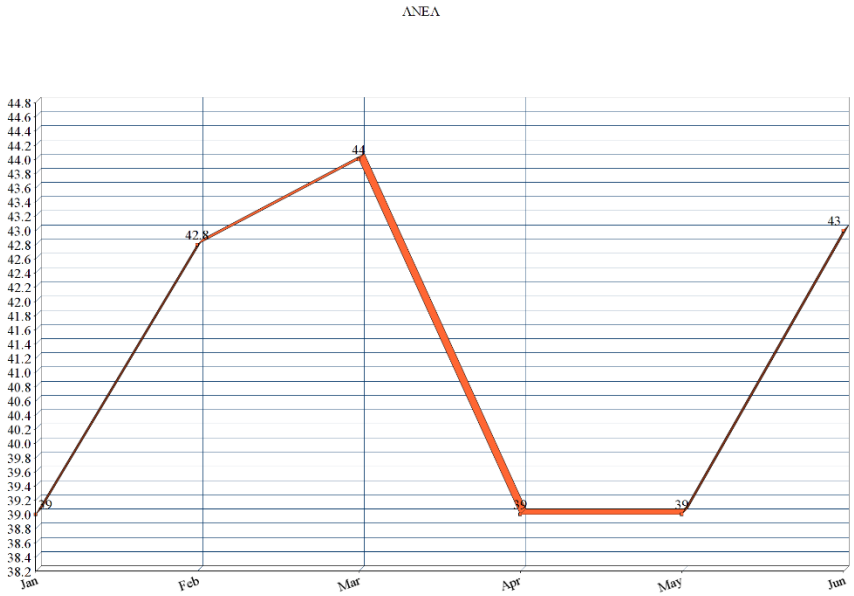
Για ακόμη μία φορά τα στοιχεία Multiple R (0.46) και significance F (0.35) δεν δείχνουν καλή συσχέτιση των δεδομένων.

Τελευταίο κόμμα, οι Ανεξάρτητοι Έλληνες ξεκινώντας από τα αποτελέσματα των δημοσκοπήσεων.



Εικόνα 26: ΑΝΕΛ - αποτελέσματα δημοσκοπήσεων

Η ανάλυση των σχολίων και η επιστροφή από το Azure φέρνουν τα εξής αποτελέσματα



Εικόνα 27: ΑΝΕΛ - αποτελέσματα ανάλυσης

Παρατηρούμε ότι υπάρχει μια αρκετά καλή συσχέτιση σύμφωνα με τα διαγράμματα. Συνεχίζουμε με την ανάλυση παλινδρόμησης.

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.626474				
R Square	0.39247				
Adjusted R Sq	0.240587				
Standard Erro	2.067115				
Observations	6				
ANOVA					
	df	*SS*	*MS*	*F*	*gnificance F*
Regression	1	11.04148	11.04148	2.584033	0.183225
Residual	4	17.09186	4.272964		
Total	5	28.13333			

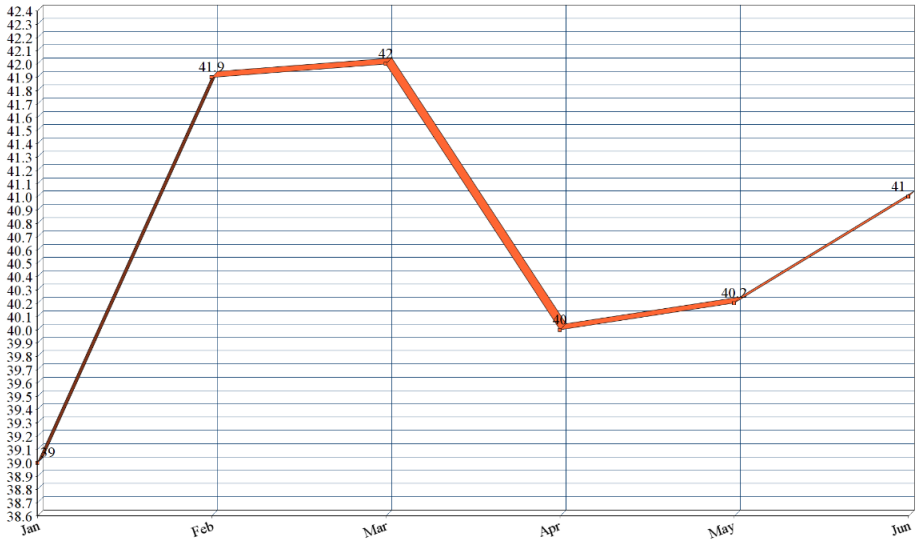
Εικόνα 28: ΑΝΕΛ - ανάλυση παλινδρόμησης

Παρατηρούμε ότι τα στοιχεία είναι σχετικά καλά και δείχνουν ότι συμφωνούν με τα προηγούμενα διαγράμματα. Οι τιμές 0.62 και 0.18 για το Multiple R και το significance F αντίστοιχα δείχνουν μια σχετικά ικανοποιητική συσχέτιση.

3.2 Συσχέτιση δημοσκοπήσεων – αποτελεσμάτων σε διαφορετικό χρονικό στάδιο

3.2.1 Αναλυτική περίπτωση χρήσης για τη Νέα Δημοκρατία

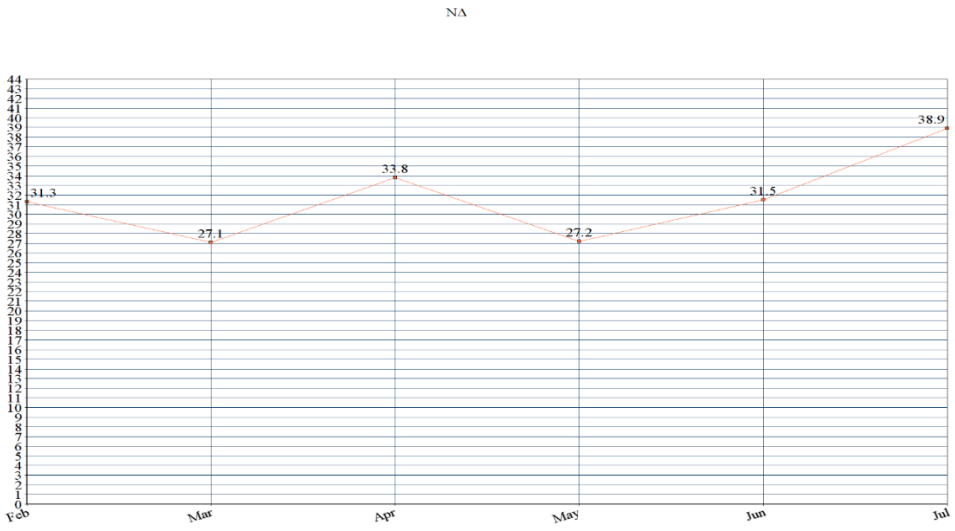
Άμα πάρουμε για το κάθε κόμμα, για κάθε μήνα που πραγματοποιείται η ανάλυση αυτή, τον αντίστοιχο μέσο όρο τότε μπορούμε να παράξουμε ένα διάγραμμα όπως φαίνεται παρακάτω.



Εικόνα 29: ΝΔ - αποτελέσματα ανάλυσης

Το συγκεκριμένο διάγραμμα είναι για το κόμμα της Νέας Δημοκρατίας για το διάστημα Ιανουάριος - Ιούνιος.

Παίρνοντας αντίστοιχα τα αποτελέσματα των δημοσκοπήσεων που δημοσιεύονται για κάθε μήνα δημιουργούμε ένα αντίστοιχο διάγραμμα όπως φαίνεται παρακάτω. Οι μήνες με τους οποίους θα συσχετίσουμε είναι Φεβρουάριος – Ιούλιος.



Εικόνα 32: ΝΔ- αποτελέσματα δημοσκοπήσεων

Παρατηρούμε αμέσως ότι υπάρχει μια σχετική συσχέτιση ανάμεσα στα δύο γραφήματα κυρίως για τους τελευταίους τέσσερις μήνες. Στη συνέχεια πραγματοποιούμε ανάλυση παλινδρόμησης στα δύο αυτά διαγράμματα για να βγάλουμε τα συμπεράσματά μας.

Πραγματοποιώντας ανάλυση παλινδρόμησης μέσω του προγράμματος excel παρατηρούμε τις εξής τιμές.

SUMMARY OUTPUT					
Regression Statistics					
Multiple F	0.435899				
R Square	0.190008				
Adjusted I	-0.07999				
Standard E	0.964858				
Observati	5				
ANOVA					
	df	*SS*	*MS*	*F*	*gnificance I*
Regressio	1	0.655147	0.655147	0.70374	0.463111
Residual	3	2.792853	0.930951		
Total	4	3.448			

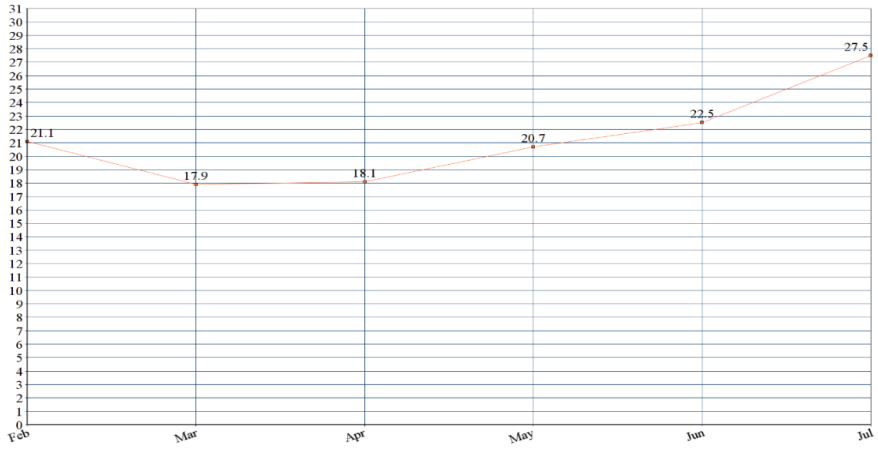
Εικόνα 30: ΝΔ - ανάλυσης παλινδρόμησης

Βλέποντας και τα αποτελέσματα από την ανάλυση παλινδρόμησης, παρατηρούμε ότι υπάρχει σχετική συσχέτιση των αποτελεσμάτων αλλά όχι ιδανική. Η τιμή του R είναι κοντά στο 0.5 και η τιμή του significance F είναι αρκετά μεγαλύτερη από 5%, στοιχεία που δεν είναι ιδανικά αλλά δείχνουν εν τέλει μια σχετική συσχέτιση.

3.2.2 Παραδείγματα περιπτώσεων χρήσης των υπόλοιπων κομμάτων

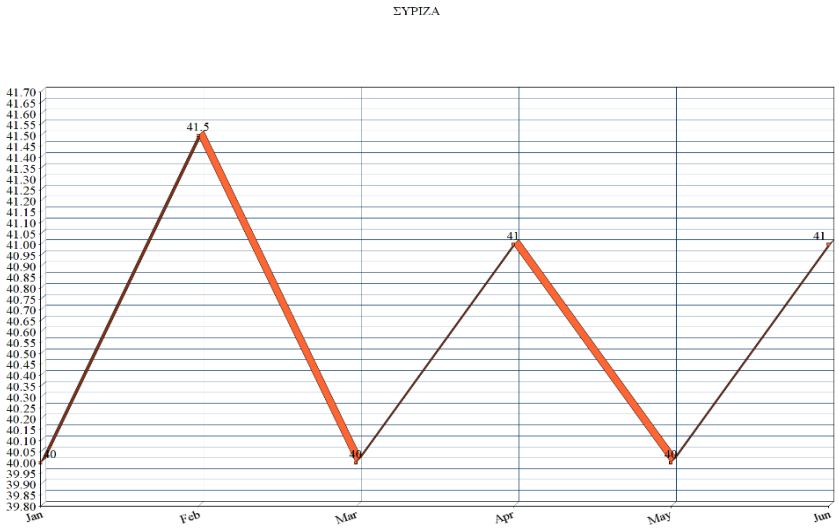
Συνεχίζουμε με τα υπόλοιπα κόμματα για να εξάγουμε κάποιο συμπέρασμα για την εγκυρότητα των αποτελεσμάτων μας, ξεκινώντας από το Σύριζα.

Οι δημοσκοπήσεις για το Σύριζα δείχνουν τα εξής αποτελέσματα.



Εικόνα 31: Συριζα- αποτελέσματα δημοσκοπήσεων

Η ανάλυση των σχολίων και η επιστροφή από το Azure φέρνουν τα εξής αποτελέσματα



Εικόνα 32: Συριζα- αποτελέσματα ανάλυσης

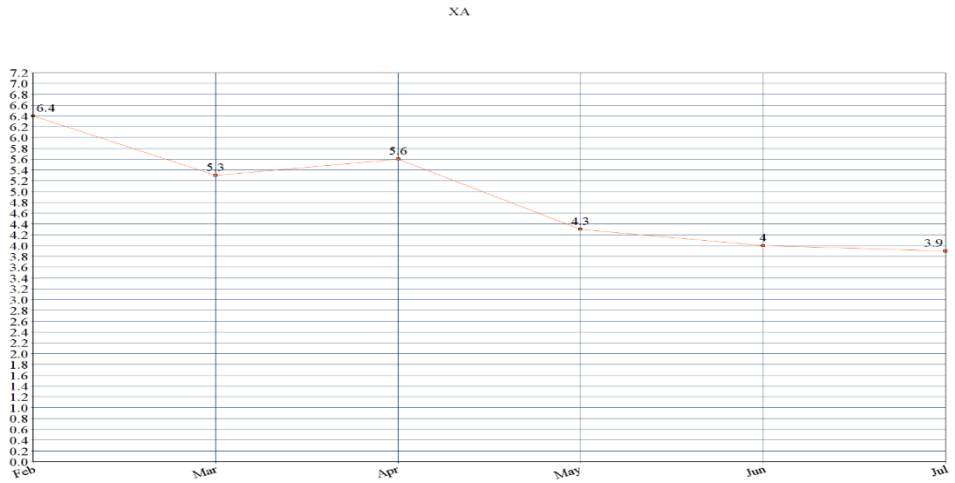
Παρατηρούμε ότι δεν υπάρχει άμεση συσχέτιση των διαγραμμάτων οπότε εκτελούμε την ανάλυση παλινδρόμησης για να δούμε τα αποτελέσματα

SUMMARY OUTPUT					
Regression Statistics					
Multiple F	0.022596				
R Square	0.000511				
Adjusted I	-0.33265				
Standard E	0.774399				
Observati	5				
ANOVA					
	df	*SS*	*MS*	*F*	*gnificance I*
Regressio	1	0.000919	0.000919	0.001533	0.971232
Residual	3	1.799081	0.599694		
Total	4	1.8			

Εικόνα 33: Συριζα- ανάλυση παλινδρόμησης

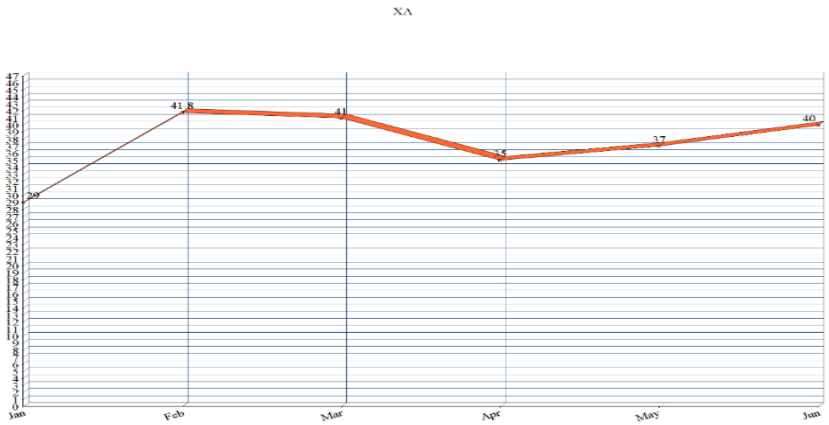
Βλέποντας και τα αποτελέσματα από την ανάλυση παλινδρόμησης παρατηρούμε ότι δεν υπάρχει καμία ουσιαστική συσχέτιση των αποτελεσμάτων. Η τιμή του R είναι πολύ κοντά στο 0 και η τιμή του significance F είναι αρκετά μεγαλύτερη από 5%, στοιχεία που δείχνουν ότι τα αποτελέσματα δεν συσχετίζονται καθόλου μεταξύ τους.

Οι δημοσκοπήσεις για τη Χρυσή Αυγή δείχνουν τα εξής αποτελέσματα



Εικόνα 34: Χρυση Αυγή - αποτελέσματα δημοσκοπήσεων

Η ανάλυση των σχολίων και η επιστροφή από το Azure φέρνουν τα εξής αποτελέσματα



Εικόνα 35: Χρυση Αυγή - αποτελέσματα ανάλυσης

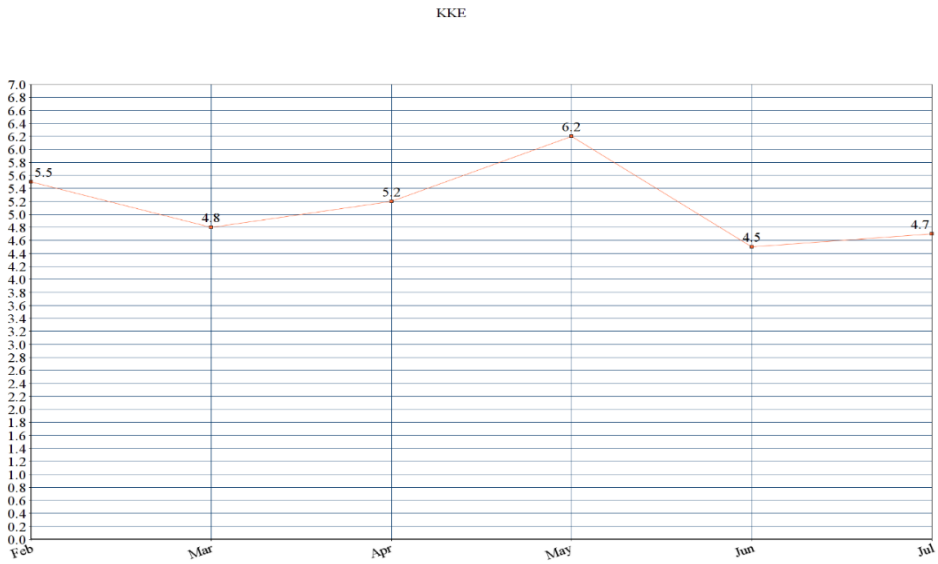
Βλέπουμε ότι υπάρχουν κάποια κοινά στοιχεί. Ας δούμε και την ανάλυση παλινδρόμησης.

SUMMARY OUTPUT					
Regression Statistics					
Multiple F	0.634381				
R Square	0.402439				
Adjusted I	0.203253				
Standard E	2.557284				
Observati	5				
ANOVA					
	df	*SS*	*MS*	*F*	*gnificance I*
Regressio	1	13.21289	13.21289	2.020411	0.250318
Residual	3	19.61911	6.539703		
Total	4	32.832			

Εικόνα 36: Χρυση Αυγή - ανάλυση παλινδρόμησης

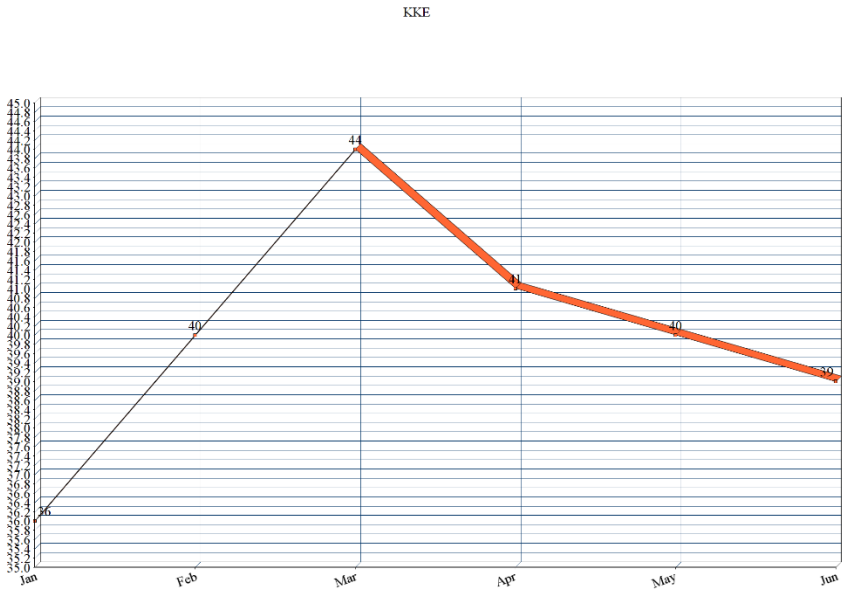
Το multiple R είναι στο 0.63 το οποίο δείχνει μια σχετικά καλή συσχέτιση ενώ η τιμή significance F είναι στο 0.25 σχετικά μακριά από το 5 % το οποίο δεν είναι ιδανικό.

Συνεχίζοντας με το ΚΚΕ, αρχικά με τις δημοσκοπήσεις.



Εικόνα 37: ΚΚΕ - αποτελέσματα δημοσκοπήσεων

Η ανάλυση των σχολίων και η επιστροφή από το Azure φέρνουν τα εξής αποτελέσματα



Εικόνα 38: KKE - αποτελέσματα ανάλυσης

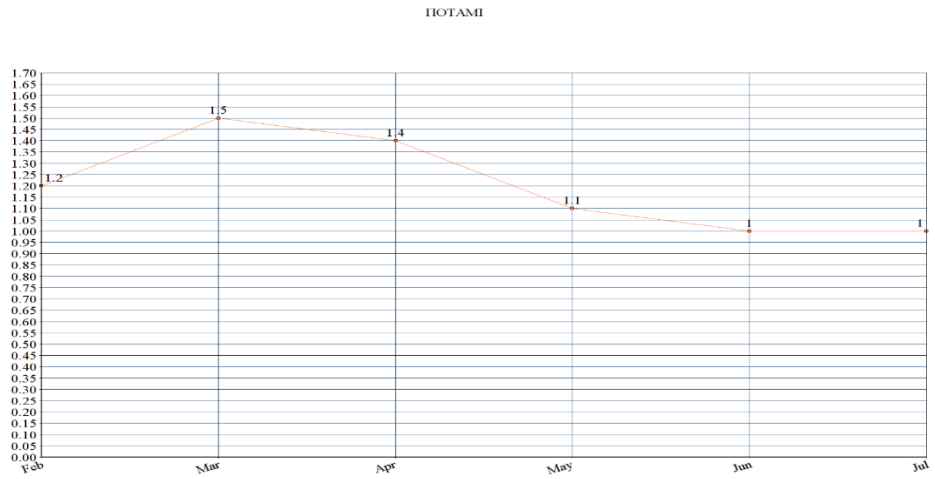
Συνεχίζουμε με τα αποτελέσματα από την ανάλυση παλινδρόμησης

SUMMARY OUTPUT					
Regression Statistics					
Multiple F	0.380668				
R Square	0.144908				
Adjusted I	-0.14012				
Standard E	2.053887				
Observati	5				
ANOVA					
	df	*SS*	*MS*	*F*	*gnificance*
Regressio	1	2.144639	2.144639	0.508395	0.527293
Residual	3	12.65536	4.218454		
Total	4	14.8			

Εικόνα 39: KKE - ανάλυση παλινδρόμησης

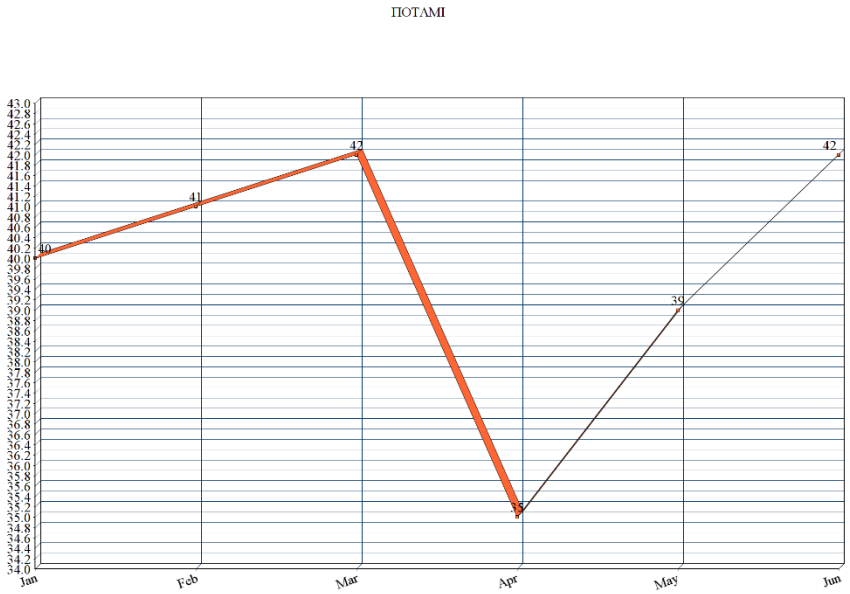
Τα στοιχεία 0.38 από το multiple R Και το 0.52 από το significance F δεν μας δείχνουν και ιδιαίτερα καλή συσχέτιση ανάμεσα στα δεδομένα.

Συνεχίζουμε με τα αποτελέσματα για το Ποτάμι, ξεκινώντας από τις δημοσκοπήσεις.



Εικόνα 40: Ποτάμι - αποτελέσματα δημοσκοπήσεων

Η ανάλυση των σχολίων και η επιστροφή από το Azure φέρνουν τα εξής αποτελέσματα



Εικόνα 41: Ποτάμι - αποτελέσματα ανάλυσης

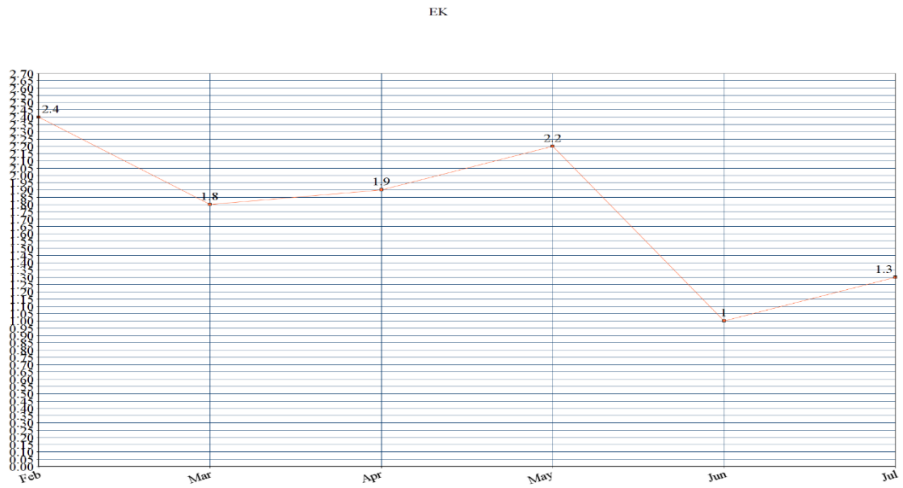
Βλέπουμε κάποια κοινά στοιχεία. Συνεχίζουμε με την ανάλυση παλινδρόμησης.

SUMMARY OUTPUT					
Regression Statistics					
Multiple F	0.361409				
R Square	0.130617				
Adjusted I	-0.15918				
Standard E	3.175665				
Observati	5				
ANOVA					
	df	*SS*	*MS*	*F*	*gnificance I*
Regressio	1	4.545455	4.545455	0.450721	0.550063
Residual	3	30.25455	10.08485		
Total	4	34.8			

Εικόνα 42: Ποτάμι - ανάλυση παλινδρόμησης

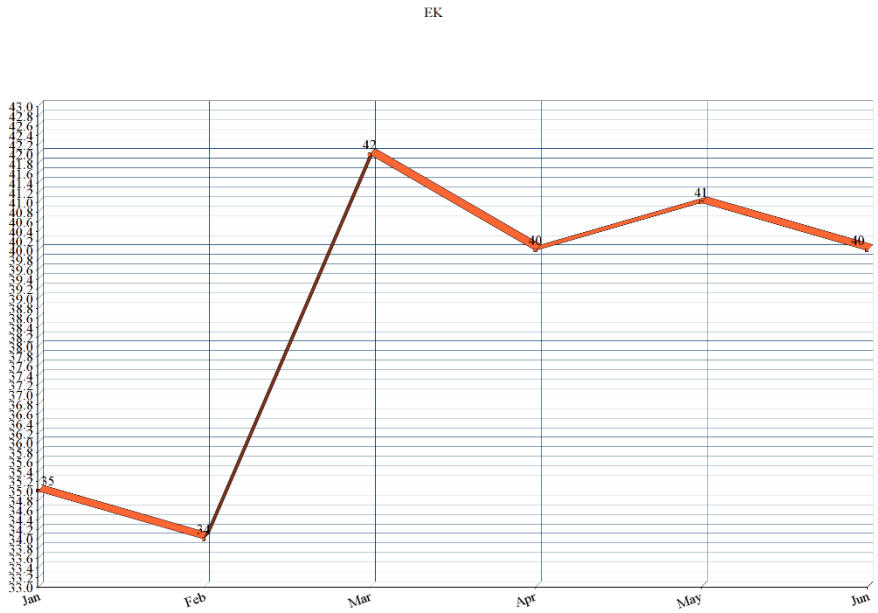
Το multiple R βρίσκεται στο 0.36 και το significance F στο 0.55 στοιχεία που δείχνουν χαμηλή συσχέτιση των δεδομένων.

Συνεχίζουμε με τα αποτελέσματα για την Ένωση Κεντρώων, ξεκινώντας από τις δημοσκοπήσεις.



Εικόνα 43: ΕΚ - αποτελέσματα δημοσκοπήσεων

Η ανάλυση των σχολίων και η επιστροφή από το Azure φέρνουν τα εξής αποτελέσματα



Εικόνα 44: ΕΚ - αποτελέσματα ανάλυσης

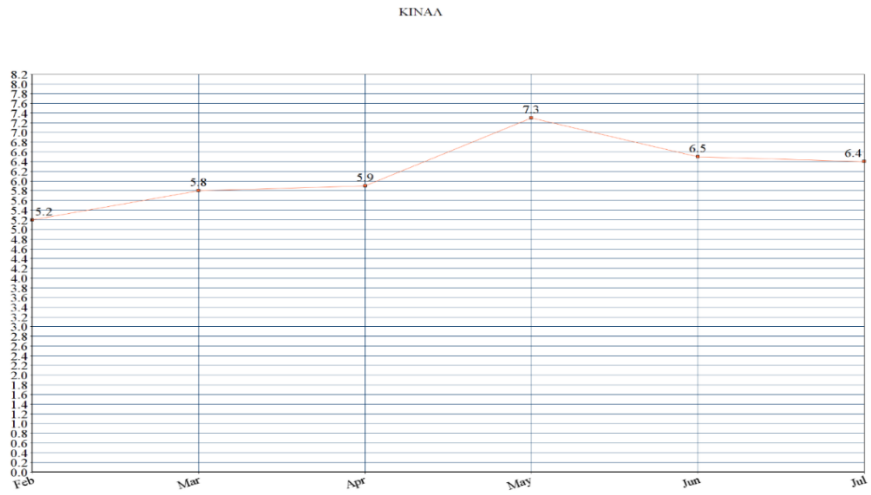
Συνεχίζουμε με τα αποτελέσματα των δημοσκοπήσεων.

SUMMARY OUTPUT					
Regression Statistics					
Multiple F	0.178679				
R Square	0.031926				
Adjusted I	-0.29077				
Standard E	3.556613				
Observati	5				
ANOVA					
	df	*SS*	*MS*	*F*	*gnificance F*
Regressio	1	1.251502	1.251502	0.098937	0.773716
Residual	3	37.9485	12.6495		
Total	4	39.2			

Εικόνα 45: ΕΚ - ανάλυση παλινδρόμησης

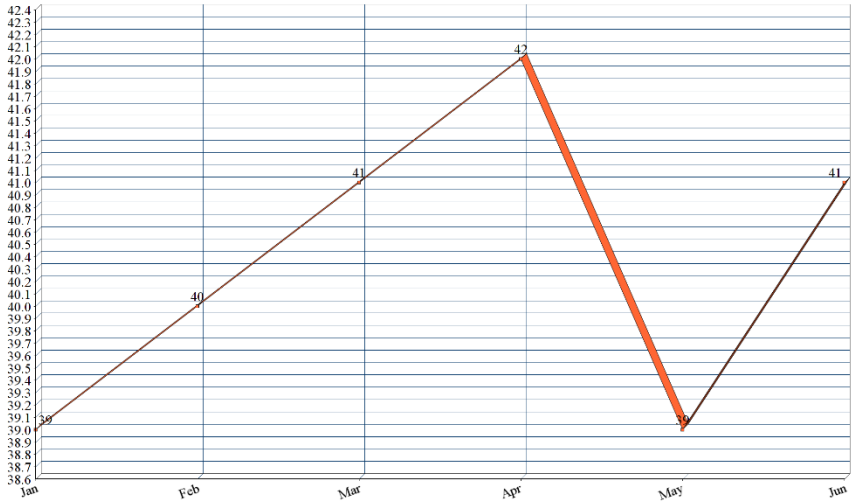
Πάλι παρατηρούμε ότι οι συντελεστές Multiple R και significance F (0.17 και 0.77 αντίστοιχα) δεν είναι καλοί δείκτες συσχέτισης των δεδομένων.

Συνεχίζουμε με το Κίνημα Αλλαγής, ξεκινώντας με τα στοιχεία των δημοσκοπήσεων.



Εικόνα 46: ΚΙΝΑΛ - αποτελέσματα δημοσκοπήσεων

Η ανάλυση των σχολίων και η επιστροφή από το Azure φέρνουν τα εξής αποτελέσματα



Εικόνα 47: ΚΙΝΑΛ - αποτελέσματα ανάλυσης

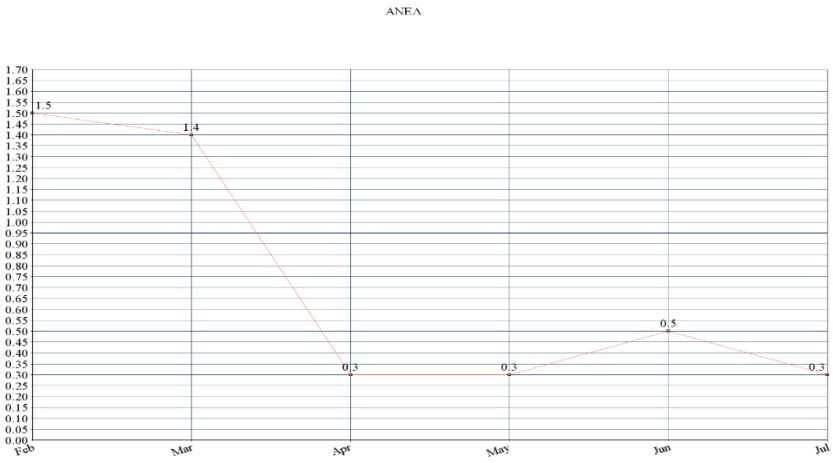
Συνεχίζουμε με τα αποτελέσματα από την ανάλυση παλινδρόμησης.

SUMMARY OUTPUT					
Regression Statistics					
Multiple F	0.438949				
R Square	0.192676				
Adjusted I	-0.07643				
Standard E	1.182946				
Observati	5				
ANOVA					
	df	*SS*	*MS*	*F*	*gnificance F*
Regressio	1	1.001917	1.001917	0.715982	0.459618
Residual	3	4.198083	1.399361		
Total	4	5.2			

Εικόνα 48: ΚΙΝΑΛ - ανάλυση παλινδρόμησης

Για ακόμη μία φορά τα στοιχεία Multiple R (0.43) και significance F (0.45) δεν δείχνουν καλή συσχέτιση των δεδομένων.

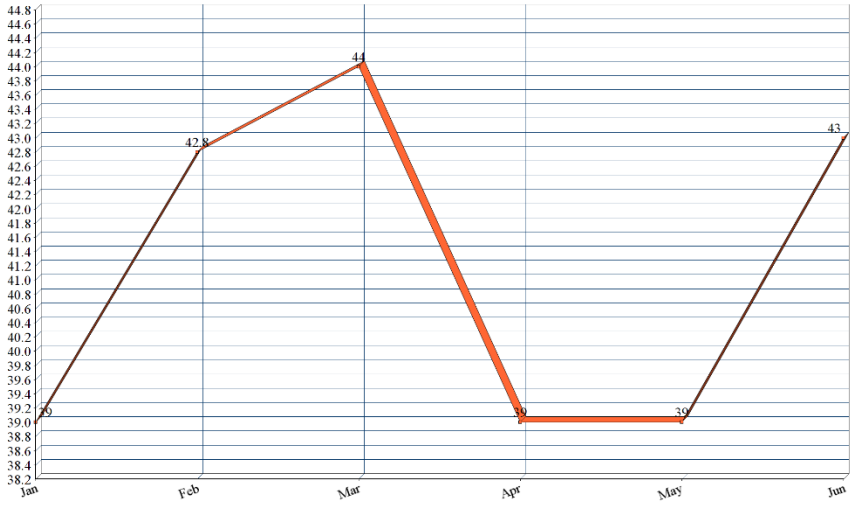
Τελευταίο κόμμα, οι Ανεξάρτητοι Έλληνες, ξεκινώντας από τα αποτελέσματα των δημοσκοπήσεων.



Εικόνα 49: ΑΝΕΛ - αποτελέσματα δημοσκοπήσεων

Η ανάλυση των σχολίων και η επιστροφή από το Azure φέρνουν τα εξής αποτελέσματα

ΑΝΕΛ



Εικόνα 50: ΑΝΕΛ - αποτελέσματα ανάλυσης

Παρατηρούμε ότι δεν υπάρχει τόσο καλή συσχέτιση σύμφωνα με τα διαγράμματα. Συνεχίζουμε με την ανάλυση παλινδρόμησης.

SUMMARY OUTPUT					
Regression Statistics					
Multiple F	0.187369				
R Square	0.035107				
Adjusted I	-0.28652				
Standard E	2.700374				
Observati	5				
ANOVA					
	df	*SS*	*MS*	*F*	*gnificance F*
Regressio	1	0.795947	0.795947	0.109153	0.762838
Residual	3	21.87605	7.292018		
Total	4	22.672			

Εικόνα 51: ΑΝΕΛ - ανάλυση παλινδρόμησης

Παρατηρούμε ότι τα στοιχεία δεν είναι καθόλου καλά και δείχνουν ότι συμφωνούν με τα προηγούμενα διαγράμματα. Οι τιμές 0.18 και 0.76 για το Multiple R και το significance F αντίστοιχα δεν δείχνουν ικανοποιητική συσχέτιση.

3.3 Σύγκριση των δύο μεθόδων

Παρατηρούμε ότι στη δεύτερη μέθοδο, ή οποία θα έπρεπε κανονικά να εμφανίζει καλύτερα αποτελέσματα, κάτι τέτοιο δεν ισχύει. Η μετατόπιση του χρόνου και η συσχέτιση των αποτελεσμάτων των σχολίων με τις δημοσκοπήσεις ενός μήνα μετά επιστρέφει χειρότερα αποτελέσματα. Αυτό μπορεί να οφείλεται στο γεγονός ότι οι δημοσκοπήσεις βρίσκονται συνήθως προς το τέλος του μήνα και είναι πιθανό η ίδια η δημοσίευση με την έρευνα που γίνεται να πραγματοποιείται αρκετά γρήγορα και προς από τα μέσα μέχρι ή τα τέλη του μήνα οπότε να συμπίπτει με τα ίδια τα σχόλια που καταγράφονται στις ιστοσελίδες.

Συμπεράσματα

Η προσπάθεια που έγινε για να πραγματοποιηθεί μια σωστή σύγκριση στις πολιτικές τάσεις ανάμεσα στα σχόλια των χρηστών σε ιστοσελίδες και στις πραγματικές δημοσκοπήσεις των εταιρειών θεωρείται αξιόλογη έως ένα σημείο, ενώ παράγοντες όπως ηλικιακές ομάδες ατόμων που γράφουν σε ιστοσελίδες και οπαδοί ή αντίπαλοι κομμάτων επηρεάζουν τα αποτελέσματα σε τέτοιο βαθμό ώστε να μην διαθέτουμε τα αποτελέσματα που επιθυμούμε. Παραμένουν οι δημοσκοπήσεις ο καλύτερος τρόπος να εξάγουμε αποτέλεσμα για τις πολιτικές τάσεις του κοινού, ενώ η έρευνα που διεξάχθηκε στη συγκεκριμένη διπλωματική είναι ένας αρκετά ορθός συμπληρωματικός τρόπος να εξάγουμε κάποιο συμπέρασμα ή να αναφερθούμε πιο συγκεκριμένα στον κόσμο που κινείται στο διαδίκτυο.

Βιβλιογραφία

Βλαχάβας Ιωάννης, Κεφαλάς Πέτρος, Βασιλειάδης Νικόλαος, Κόκκορας Φώτης,Σακελλαρίου Ηλίας, 2011. *Τεχνητή Νοημοσύνη Γ' 'Εκδοση*. Εκδόσεις Πανεπιστημίου Μακεδονίας, pp.1-4. *[1]*

Liu B., 2010. *Sentiment Analysis And Subjectivity*. [online] www.cs.uic.edu. Available at: <https://www.cs.uic.edu/~liub/FBS/NLP-handbook-sentiment-analysis.pdf> [Accessed 26 June 2020]. *[2]*

Βλαχάβας Ιωάννης, Κεφαλάς Πέτρος, Βασιλειάδης Νικόλαος, Κόκκορας Φώτης,Σακελλαρίου Ηλίας, 2011. *Τεχνητή Νοημοσύνη Γ' 'Εκδοση*. Εκδόσεις Πανεπιστημίου Μακεδονίας, pp.335-336. *[3]*

Hedley J., 2009. *Jsoup Java HTML Parser 1.13.1 API*. [online] jsoup.org. Available at: <https://jsoup.org/apidocs/> [Accessed 26 June 2020]. *[4]*

Aydin O., 2018. *R Web Scraping Quick Start Guide: Techniques And Tools To Crawl And Scrape Data From Websites*. pp.7-8. *[5]*

Sirosh J., 2014. *Microsoft Azure Machine Learning Combines Power Of Comprehensive Machine Learning With Benefits Of Cloud*. [online] blogs.microsoft.com. Available at: <https://blogs.microsoft.com/blog/2014/06/16/microsoft-azure-machine-learning-combines-power-of-comprehensive-machine-learning-with-benefits-of-cloud/> [Accessed 26 June 2020]. *[6]*

Harthshorn S., 2017. *Linear Regression And Correlation: A Beginner's Guide*. p.1. *[7]*

Cheusheva S., 2019. *Linear Regression Analysis In Excel*. [online] www.ablebits.com. Available at: <https://www.ablebits.com/office-addins-blog/2018/08/01/linear-regression-analysis-excel/#regression-analysis-in-Excel> [Accessed 26 June 2020]. *[8]*

DavidJ.C.MacKay, 2005. InformationTheory, Inference,andLearningAlgorithms. P.468 [9]