# ΤΕΧΝΟΛΟΓΙΕΣ ΕΞΑΓΩΓΗΣ ΠΕΡΙΕΧΟΜΕΝΟΥ ΑΠΟ ΙΣΤΟΣΕΛΙΔΕΣ ΓΙΑ ΑΝΑΛΥΣΗ ΑΓΓΕΛΙΩΝ

Διπλωματική Εργασία ΠΜΣ Εφαρμοσμένης Πληροφορικής

**ΠΕΤΣΚΟΣ ΔΗΜΗΤΡΙΟΣ (mai16018)**

Επιβλέπων Καθηγητής: Στειακάκης Εμμανουήλ

# Περιεχόμενα

- Εισαγωγή – Σημαντικότητα Προβλήματος
- Θεωρητικό Υπόβαθρο (Web Crawling, Θέματα Υλοποίησης, Ηθικοί Φραγμοί και Νομιμότητα)
- Τεχνολογίες – Εργαλεία
- Βασική Αρχιτεκτονική Υλοποίησης
- Crawlers Ανάκτησης Δεδομένων από Αγγελίες
- Χαρτογράφηση Δεξιοτήτων (Mapping)
- Ανάλυση Δεδομένων με χρήση του Cluvio
- Συμπεράσματα, Περιορισμοί και Μελλοντικές Επεκτάσεις

# Εισαγωγή – Σημαντικότητα Προβλήματος

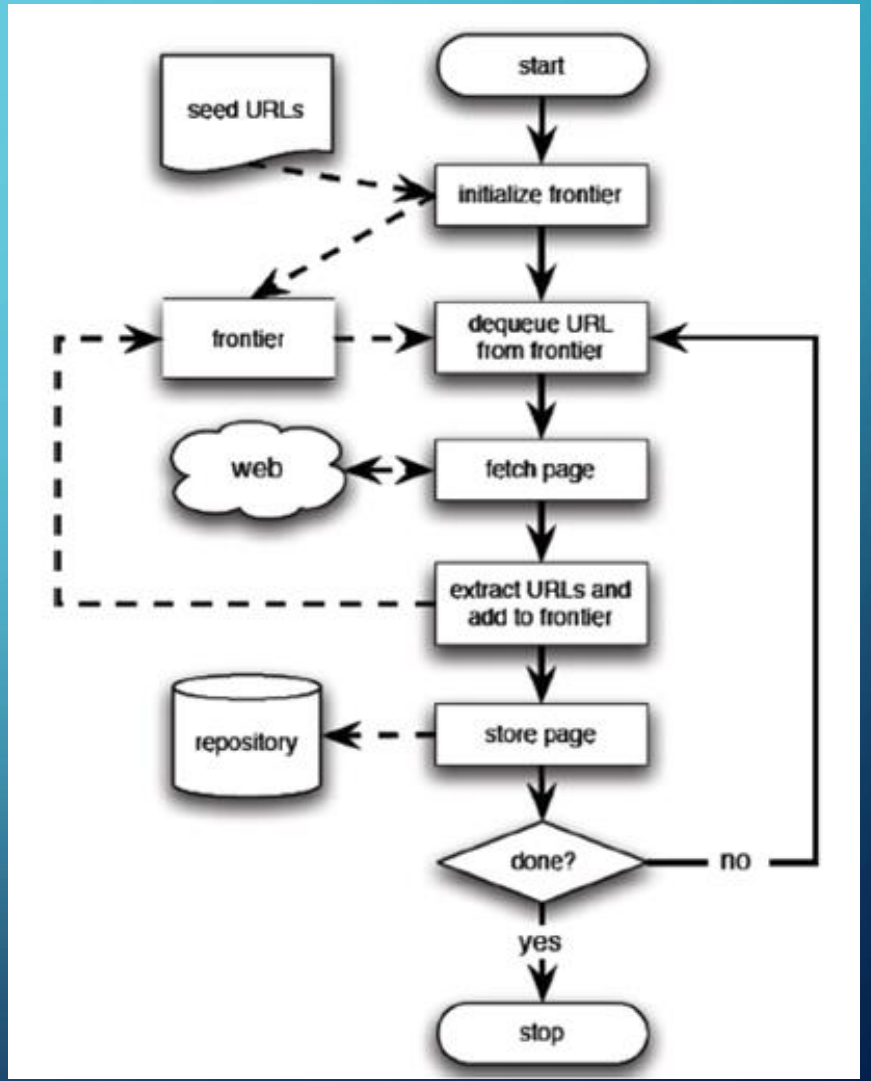
- Στη σημερινή εποχή έχει αλλάξει σε μεγάλο βαθμό ο τρόπος αναζήτησης εργασίας, λόγω της καθημερινής δημοσίευσης μεγάλου αριθμού αγγελιών στο διαδίκτυο
- Ο τεράστιος όγκος δημοσιευμένων αγγελιών προσθέτει δυσκολία στους αναζητούντες εργασία
- Μέσα στο πλήθος της πληροφορίας είναι πιθανό να μην είναι ξεκάθαρο ποιοι κλάδοι απασχόλησης παρουσιάζουν άνθηση και ποιες δεξιότητες έχουν μεγαλύτερη ζήτηση
- Μέσω μιας υλοποίησης ανάκτησης δεδομένων από αγγελίες και ανάλυσής τους είναι εφικτή η εξαγωγή συμπερασμάτων

# Web Crawling

- Web crawlers, scrapers, bots, robots, spiders
- Ένα σύστημα για την αυτοματοποιημένη μαζική λήψη ιστοσελίδων
- Αυτοματοποιημένο πρόγραμμα για αποστολή αιτημάτων σε ένα διακομιστή ιστού, λήψη δεδομένων (HTML), ανάλυση των δεδομένων για εξαγωγή πληροφοριών
- Χρήσιμη ικανότητα για την εργαλειοθήκη ενός προγραμματιστή

# Βασικός Αλγόριθμος

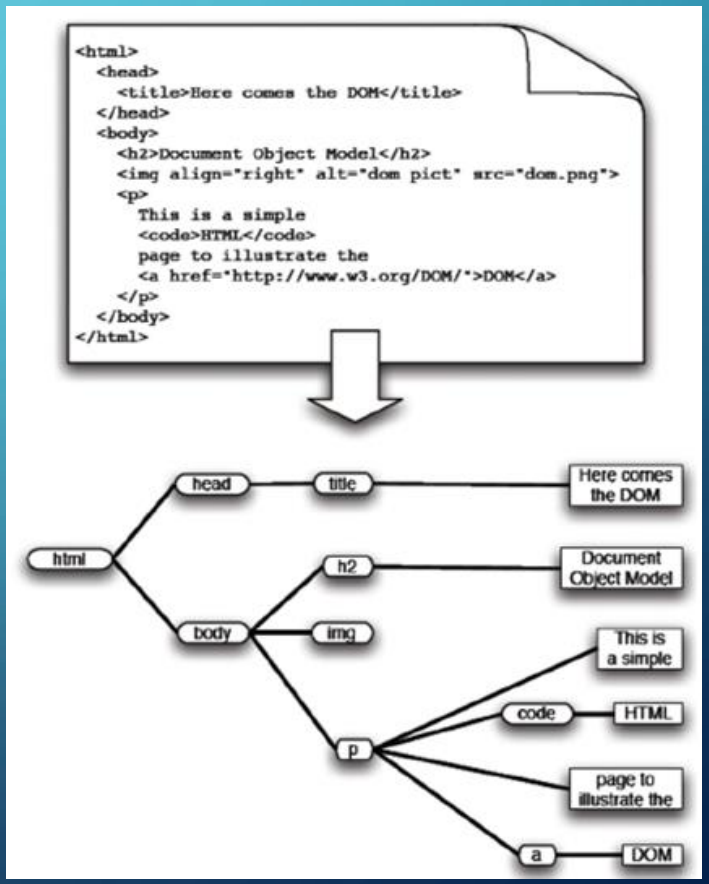
- Εκκίνηση από ένα σύνολο σελίδων σπόρου
- Χρήση συνδέσμων μέσα σε αυτές για ανάκτηση άλλων σελίδων
- Επανάληψη διαδικασίας μέχρις ότου να επισκεφθεί ένας επαρκής αριθμός σελίδων ή να επιτευχθεί άλλος στόχος




*Πηγή: Liu B. (2011)*

# Θέματα Υλοποίησης (1/2)

- Ανάκτηση:
  - ο crawler λειτουργεί ως Web client
- Ανάλυση:
  - ο crawler αναλύει το περιεχόμενο της σελίδας που λαμβάνεται
  - η δενδροειδής μορφή μιας σελίδας HTML (DOM) διευκολύνει την προσπέλαση των δεδομένων με χρήση κάποιου parser




*Πηγή: Liu B. (2011)*

# Θέματα Υλοποίησης (2/2)

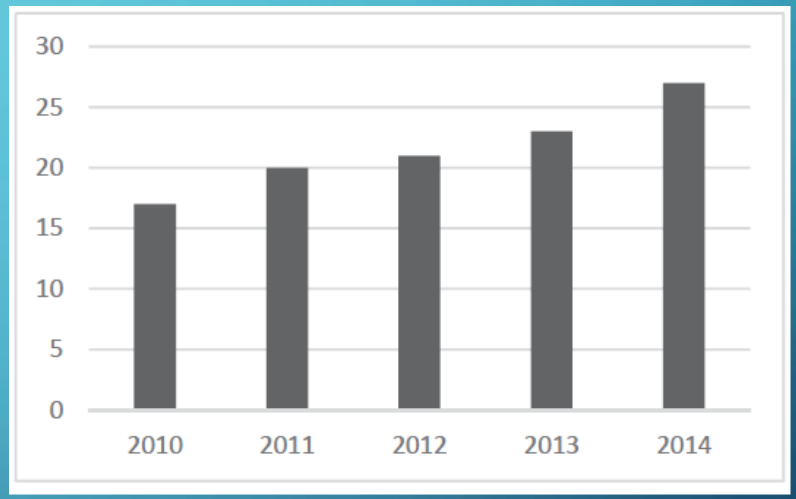
- Προ – επεξεργασία κειμένου:
  - αφαίρεση stopwords, δηλαδή των συχνά εμφανιζόμενων και ασήμαντων λέξεων που δεν προσθέτουν περιεχόμενο σε ένα έγγραφο (πχ. a, an, are, as, at κλπ.)
  - stemming, αναφέρεται στη διαδικασία μείωσης των λέξεων στις ρίζες τους (πχ. computer, computing και compute σε comput)
  - αφαίρεση αριθμών και σημείων στίξης (κατά περίπτωση)
  - αφαίρεση ετικετών HTML και διατήρηση κύριων μπλοκ περιεχομένου

# Ηθικοί Φραγμοί και Νομιμότητα

- Το web crawling βρίσκεται νομικά σε γκρίζα περιοχή
- Απαγορεύεται ρητά στους όρους χρήσης της υπηρεσίας των ιστοτόπων
  - monster.com: *"All Monster Users agree to not: (d) use any data mining, robots or similar data gathering or extraction methods"*
  - indeed.com: *"Unless you have been specifically permitted to do so in a separate, written agreement with Indeed, you agree that you will not crawl, scrape, reproduce, duplicate, copy, sell, trade or resell the Site for any purpose"*
- Ένας αποτελεσματικός crawler ασκεί σημαντική πίεση στους πόρους ενός διακομιστή, οδηγώντας ακόμα και σε denial of service
- Συστήνεται η συμμόρφωση με το robots.txt

# Crawling Traps

- Λόγω της αύξησης της κίνησης των bots στο διαδίκτυο οι περισσότεροι ιστότοποι εφαρμόζουν τεχνολογίες ανίχνευσης και διάκρισης
  - αναγνώριση μέσω των κεφαλίδων HTTP
  - honeypots
  - ανίχνευση υπερβολικά γρήγορου ρυθμού αποστολής αιτημάτων
  - CAPTCHA




*Πηγή: Menshchikov A. et al (2017)*

# Τεχνολογίες - Εργαλεία

- Java
- Eclipse
- MySQL
- Selenium
- JSoup
- XPath
- JAXB API
- HTML Cleaner
- Cluvio

# Selenium

- Περιλαμβάνει το API WebDriver
- Επιτρέπει το άνοιγμα ενός browser προγραμματιστικά
- Δεν χρειάζεται να γίνει ρύθμιση των κεφαλίδων ή διαχείριση των cookies
- Υποστηρίζει Javascript

# JSoup

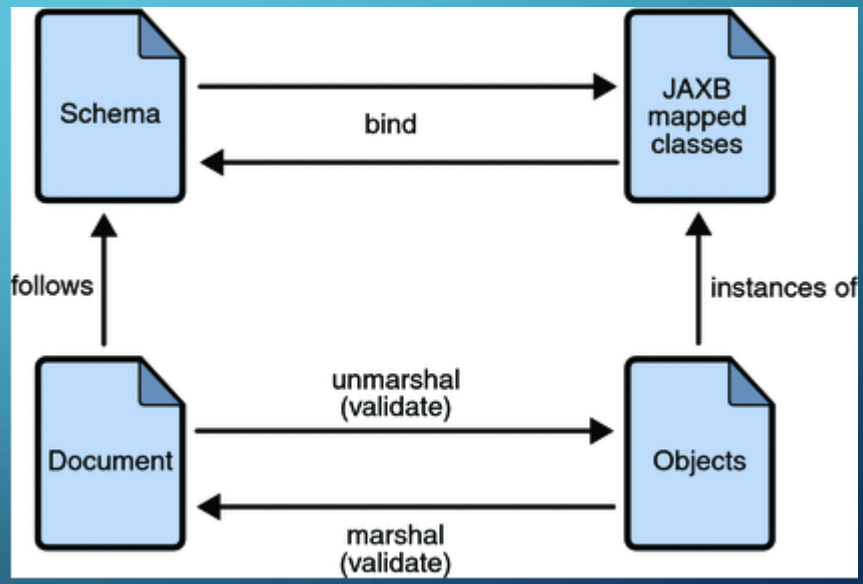
- Βιβλιοθήκη ανάκτησης ιστοσελίδων
- Ανάκτηση και ανάλυση HTML από ένα URL, ένα αρχείο ή ένα αλφαριθμητικό
- Αναζήτηση και εξαγωγή δεδομένων
- Διαχείριση των HTML στοιχείων
- Δυνατότητα εκκαθάρισης περιεχομένου που εισάγεται από τον χρήστη
- Παραγωγή τακτοποιημένου HTML κειμένου

# XPath

- Μια γλώσσα ερωτήματος (query language)
- Χρησιμοποιεί εκφράσεις διαδρομής τοποθεσίας (paths) για πλοήγηση σε έγγραφα XML
- Η πιο συνηθισμένη σύνταξη της XPath είναι
  //nodename[@attribute='value']

# JAXB API

- Παρέχει έναν αποδοτικό τρόπο χαρτογράφησης μεταξύ του κώδικα XML και της Java
- Αποσύνθεση περιεχομένου XML σε αναπαράσταση Java (Unmarshal)
- Σύνθεση της αναπαράστασης σε Java του περιεχομένου XML και πάλι σε περιεχόμενο XML (Marshal)




*Πηγή: Oracle (2010)*

# HTML Cleaner

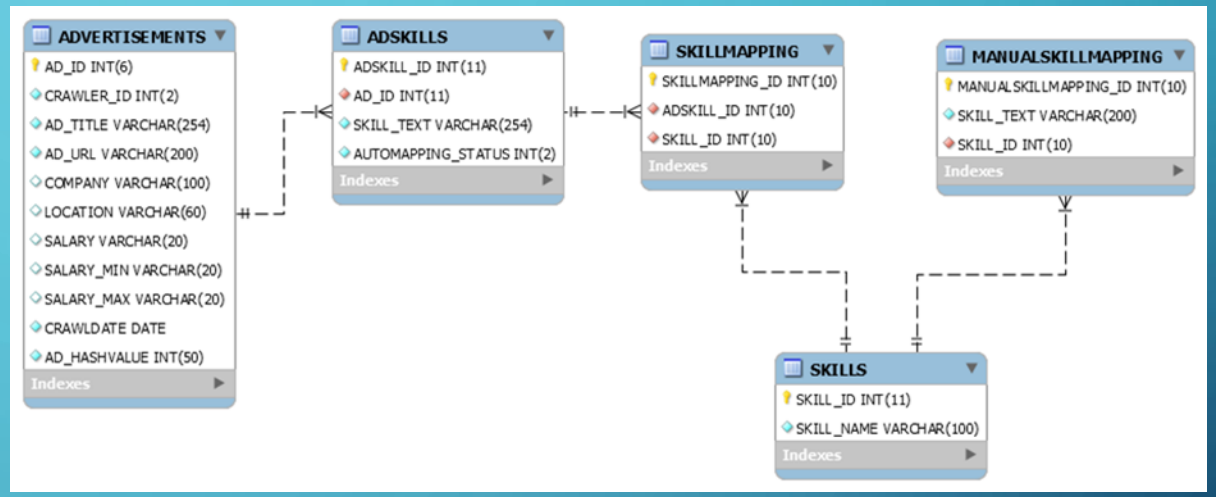
- HTML parser ανοιχτού κώδικα
- Αναδιατάσσει την HTML και παράγει καλά μορφοποιημένη XML
- Εξαγωγή της DOM δομής σε XML ώστε να μπορεί να γίνει επεξεργασία με χρήση XPath, XQuery και XSLT
- Μπορεί να χρησιμοποιηθεί μέσα από κώδικα Java

# Cluvio

- Πλατφόρμα ανάλυσης μετρικών και BI
- Επιτρέπει την σύνδεση με βάση δεδομένων
- Υποστηρίζει όλες τις δημοφιλείς SQL βάσεις, συμπεριλαμβανομένης της MySQL
- Εύκολη δημιουργία αναφορών και γραφημάτων με τη χρήση ενός ερωτήματος SQL ή ενός σεναρίου (script) σε R
- Δυνατότητα συγκέντρωσης των αναφορών και των γραφημάτων σε ένα dashboard

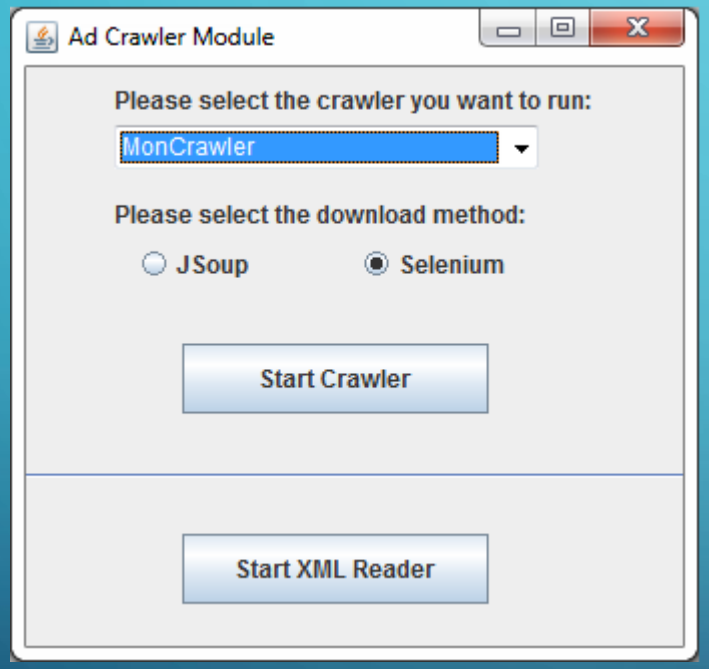
# Βασική Αρχιτεκτονική Υλοποίησης

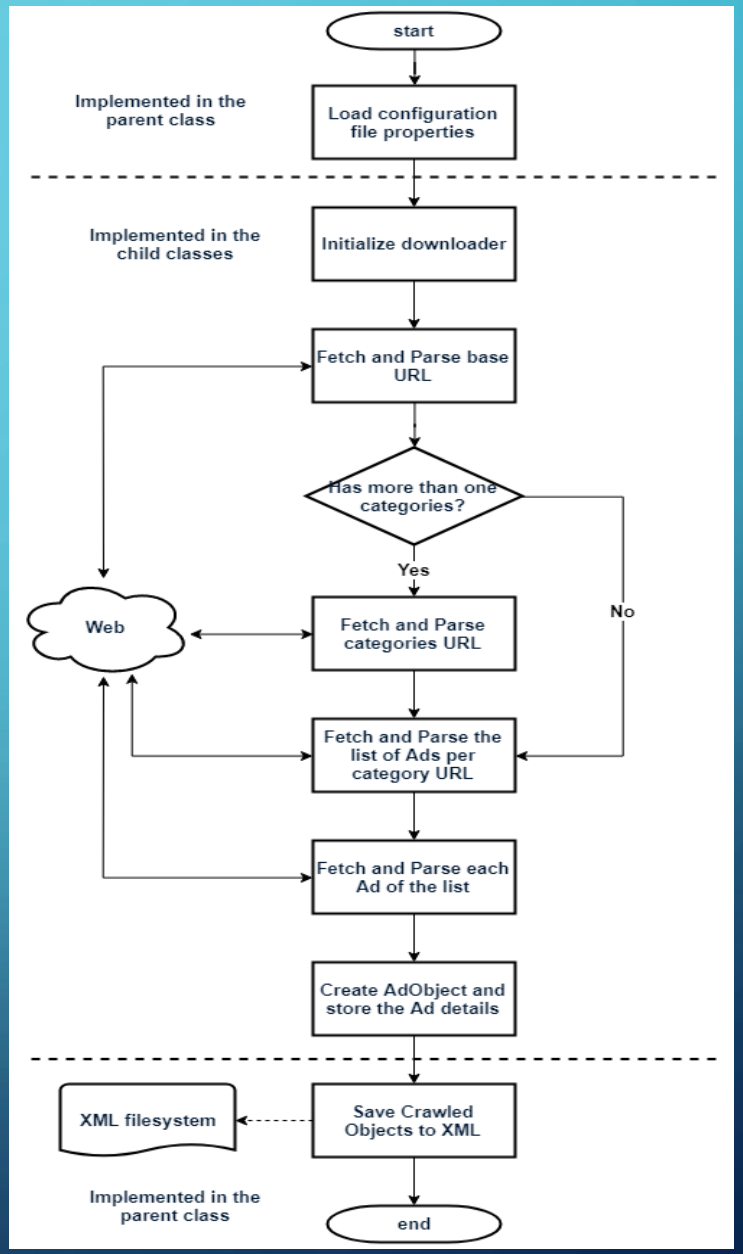
# Βάση Δεδομένων



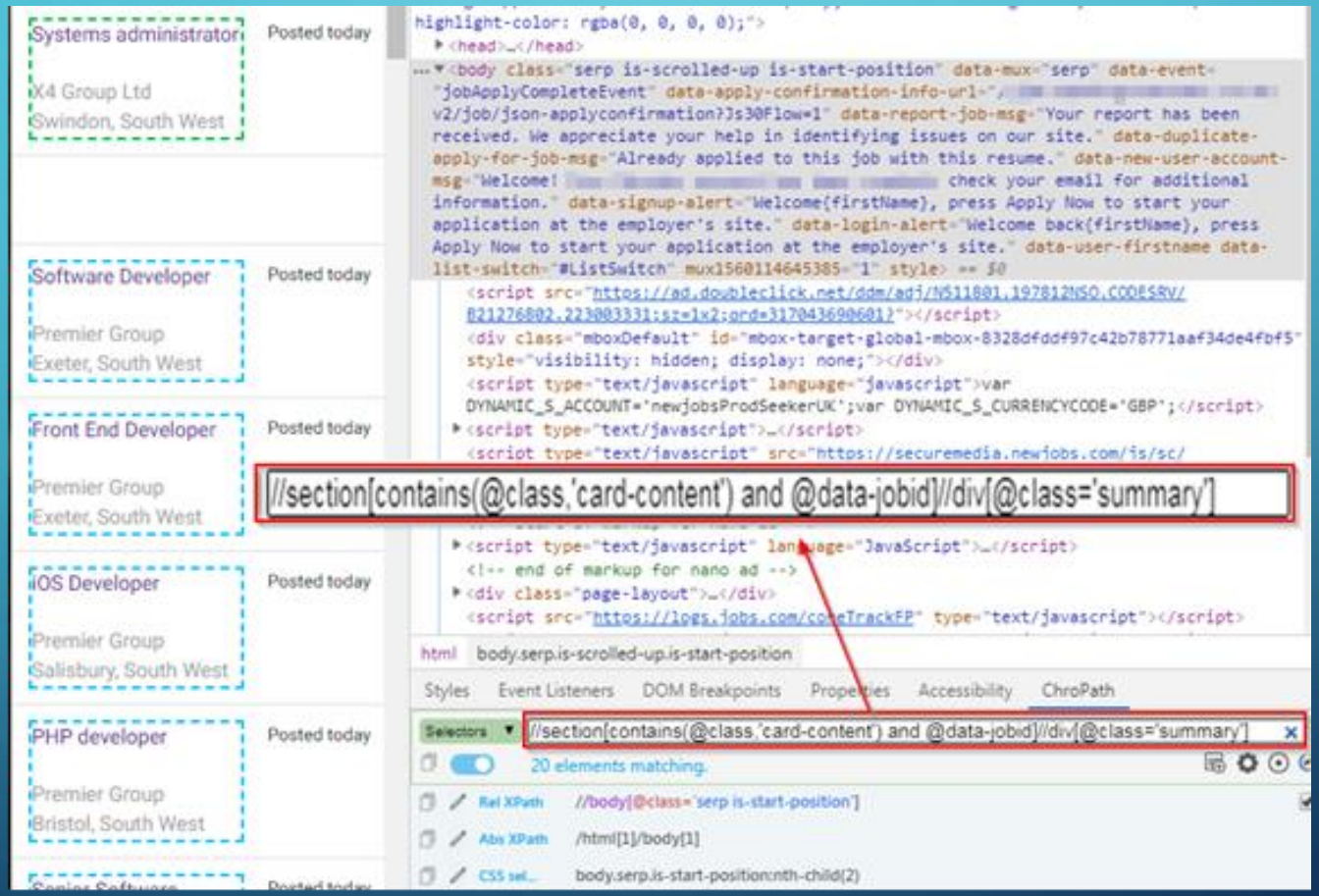

- ADVERTISEMENTS: οι αγγελίες με τις κύριες πληροφορίες τους
- ADSKILLS: οι απαιτούμενες δεξιότητες κάθε αγγελίας
- SKILLS: οι γενικές κατηγορίες δεξιοτήτων
- SKILLMAPPING: χαρτογράφηση δεξιοτήτων σε κάποιες γενικές κατηγορίες
- MANUALSKILLMAPPING: χειροκίνητη αντιστοίχηση των γενικών κατηγοριών δεξιοτήτων

# Crawling Module

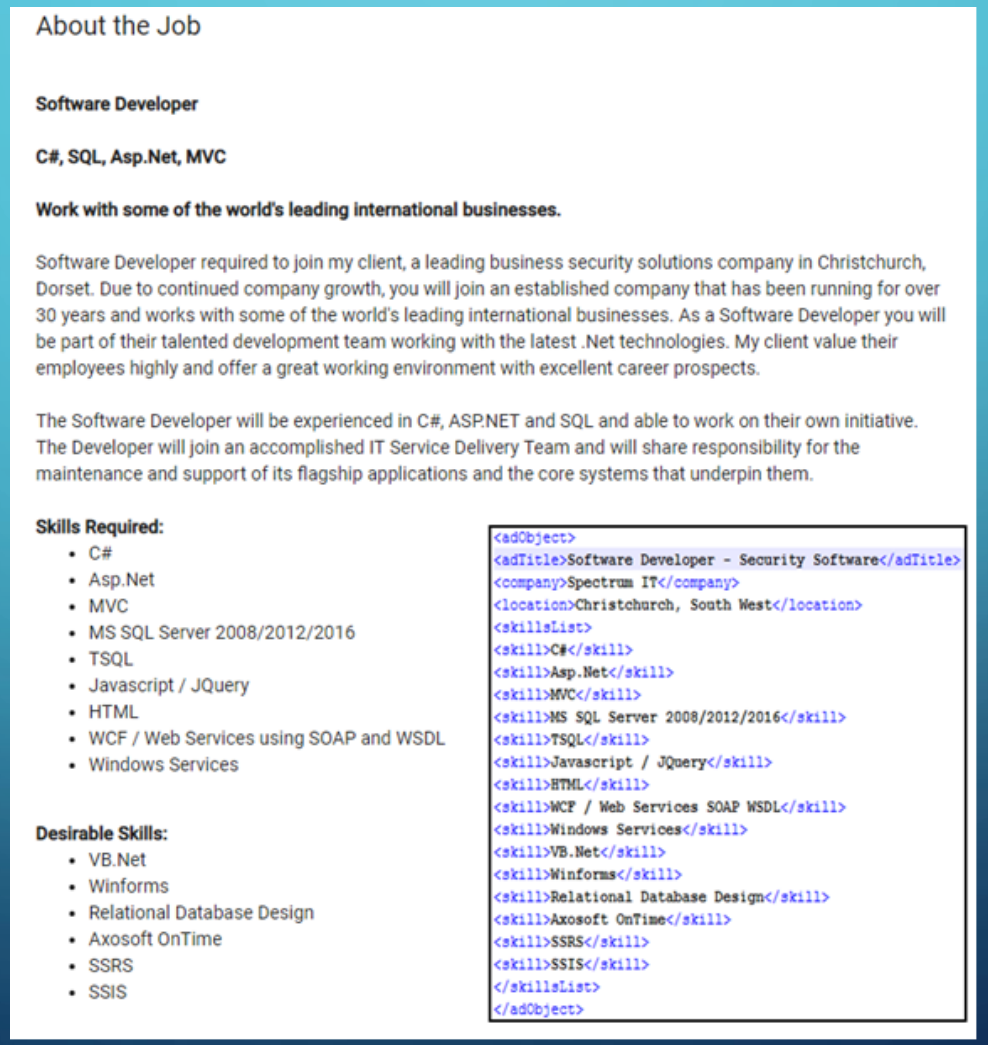


# Parsing με XPath



# Μετάβαση από HTML σε XML



About the Job

**Software Developer**

**C#, SQL, Asp.Net, MVC**

**Work with some of the world's leading international businesses.**

Software Developer required to join my client, a leading business security solutions company in Christchurch, Dorset. Due to continued company growth, you will join an established company that has been running for over 30 years and works with some of the world's leading international businesses. As a Software Developer you will be part of their talented development team working with the latest .Net technologies. My client value their employees highly and offer a great working environment with excellent career prospects.

The Software Developer will be experienced in C#, ASP.NET and SQL and able to work on their own initiative. The Developer will join an accomplished IT Service Delivery Team and will share responsibility for the maintenance and support of its flagship applications and the core systems that underpin them.

**Skills Required:**

- C#
- Asp.Net
- MVC
- MS SQL Server 2008/2012/2016
- TSQL
- Javascript / JQuery
- HTML
- WCF / Web Services using SOAP and WSDL
- Windows Services

**Desirable Skills:**

- VB.Net
- Winforms
- Relational Database Design
- Axosoft OnTime
- SSRS
- SSIS

```
<adObject>
<adTitle>Software Developer - Security Software</adTitle>
<company>Spectrum IT</company>
<location>Christchurch, South West</location>
<skillsList>
<skill>C#</skill>
<skill>Asp.Net</skill>
<skill>MVC</skill>
<skill>MS SQL Server 2008/2012/2016</skill>
<skill>TSQL</skill>
<skill>Javascript / JQuery</skill>
<skill>HTML</skill>
<skill>WCF / Web Services SOAP WSDL</skill>
<skill>Windows Services</skill>
<skill>VB.Net</skill>
<skill>Winforms</skill>
<skill>Relational Database Design</skill>
<skill>Axosoft OnTime</skill>
<skill>SSRS</skill>
<skill>SSIS</skill>
</skillsList>
</adObject>
```

# Εισαγωγή στη Βάση

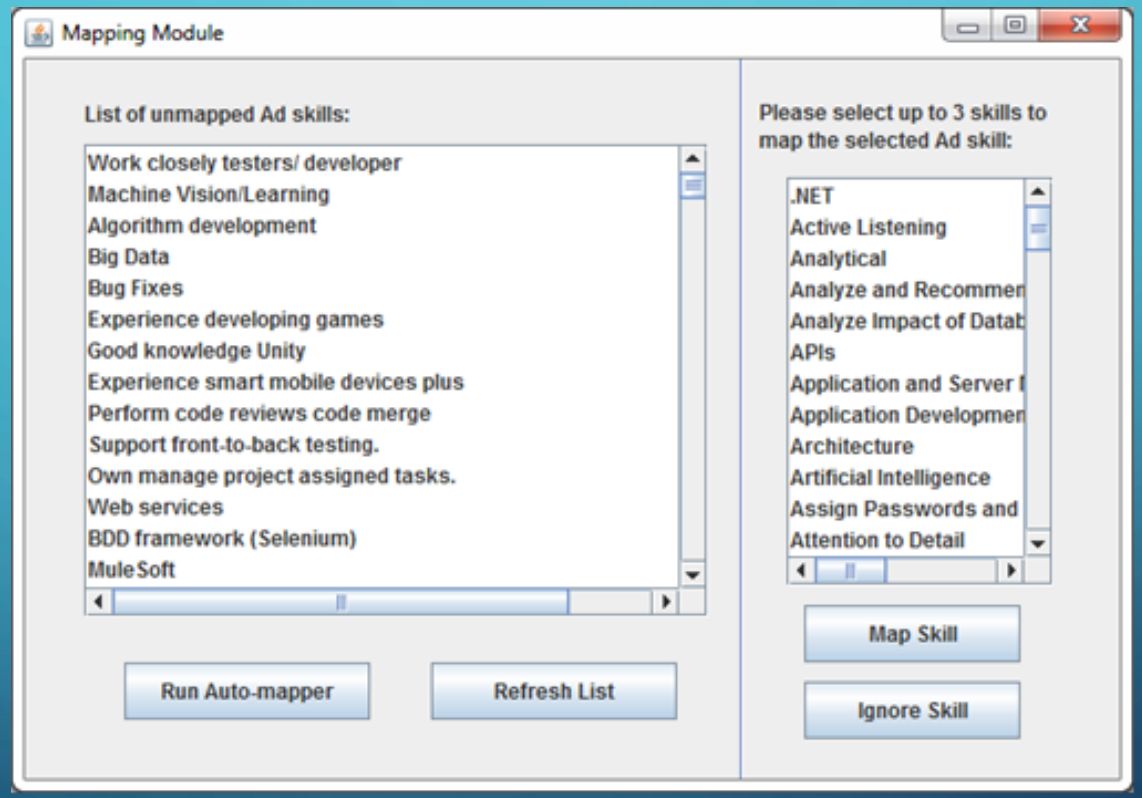
```
SELECT * FROM ADVERTISEMENTS a, ADSKILLS b WHERE a.AD_ID = b.AD_ID AND a.AD_ID = 310
```

Show all | Restore column order | Number of rows: 25 | Filter rows: Search this table

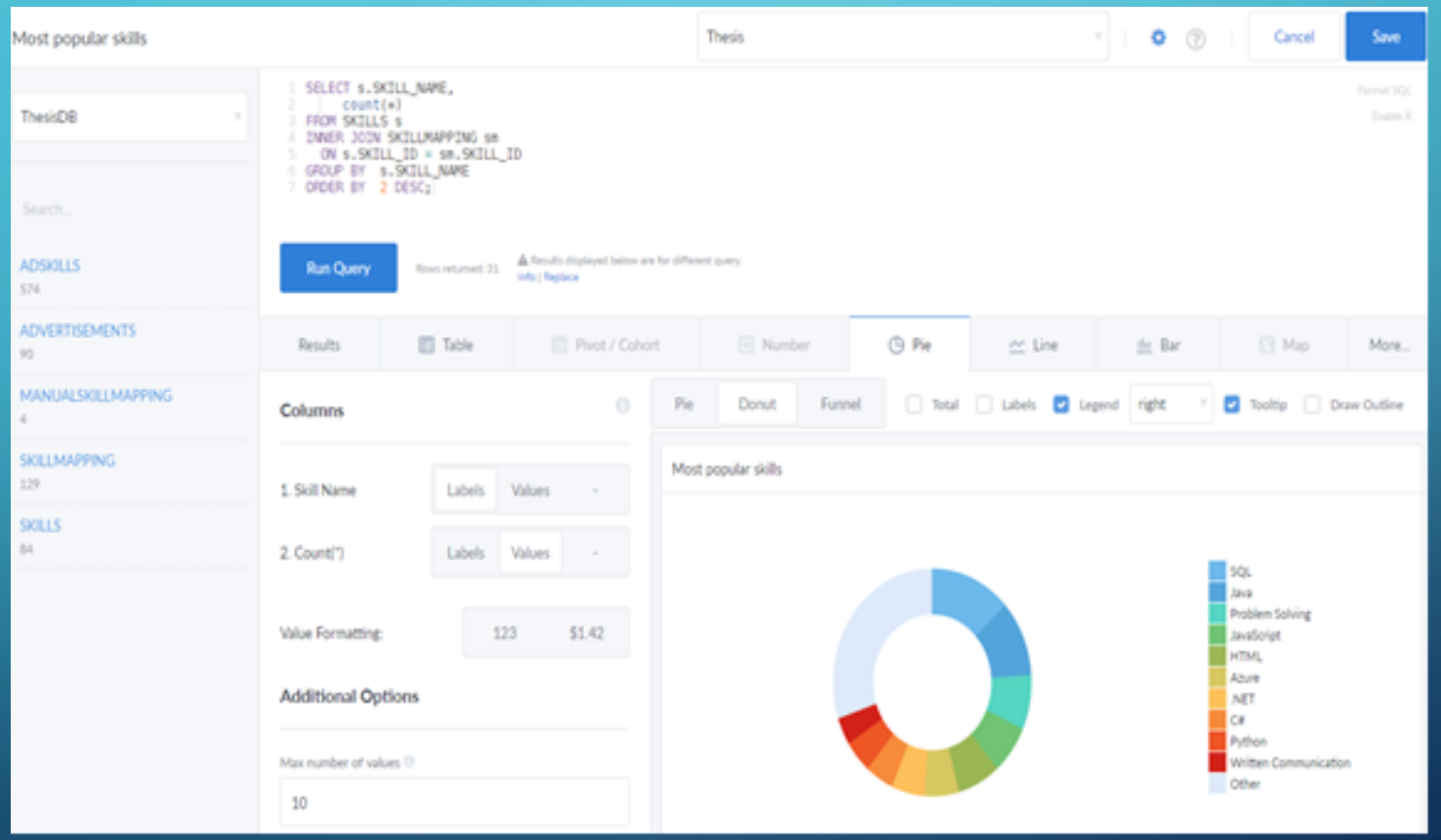
Options

| AD_ID | CRAWLER_ID | AD_TITLE | COMPANY | LOCATION | ADSKILL_ID | SKILL_TEXT |
|---|---|---|---|---|---|---|
| 310 | 1 | Software Developer - Security Software | Spectrum IT | Christchurch, South West | 1316 | C# |
| 310 | 1 | Software Developer - Security Software | Spectrum IT | Christchurch, South West | 1317 | Asp.Net |
| 310 | 1 | Software Developer - Security Software | Spectrum IT | Christchurch, South West | 1318 | MVC |
| 310 | 1 | Software Developer - Security Software | Spectrum IT | Christchurch, South West | 1319 | MS SQL Server 2008/2012/2016 |
| 310 | 1 | Software Developer - Security Software | Spectrum IT | Christchurch, South West | 1320 | TSQL |
| 310 | 1 | Software Developer - Security Software | Spectrum IT | Christchurch, South West | 1321 | Javascript / JQuery |
| 310 | 1 | Software Developer - Security Software | Spectrum IT | Christchurch, South West | 1322 | HTML |
| 310 | 1 | Software Developer - Security Software | Spectrum IT | Christchurch, South West | 1323 | WCF / Web Services SOAP WSDL |
| 310 | 1 | Software Developer - Security Software | Spectrum IT | Christchurch, South West | 1324 | Windows Services |
| 310 | 1 | Software Developer - Security Software | Spectrum IT | Christchurch, South West | 1325 | VB.Net |
| 310 | 1 | Software Developer - Security Software | Spectrum IT | Christchurch, South West | 1326 | Winforms |
| 310 | 1 | Software Developer - Security Software | Spectrum IT | Christchurch, South West | 1327 | Relational Database Design |
| 310 | 1 | Software Developer - Security Software | Spectrum IT | Christchurch, South West | 1328 | Axosoft OnTime |
| 310 | 1 | Software Developer - Security Software | Spectrum IT | Christchurch, South West | 1329 | SSRS |
| 310 | 1 | Software Developer - Security Software | Spectrum IT | Christchurch, South West | 1330 | SSIS |

# Mapping Module

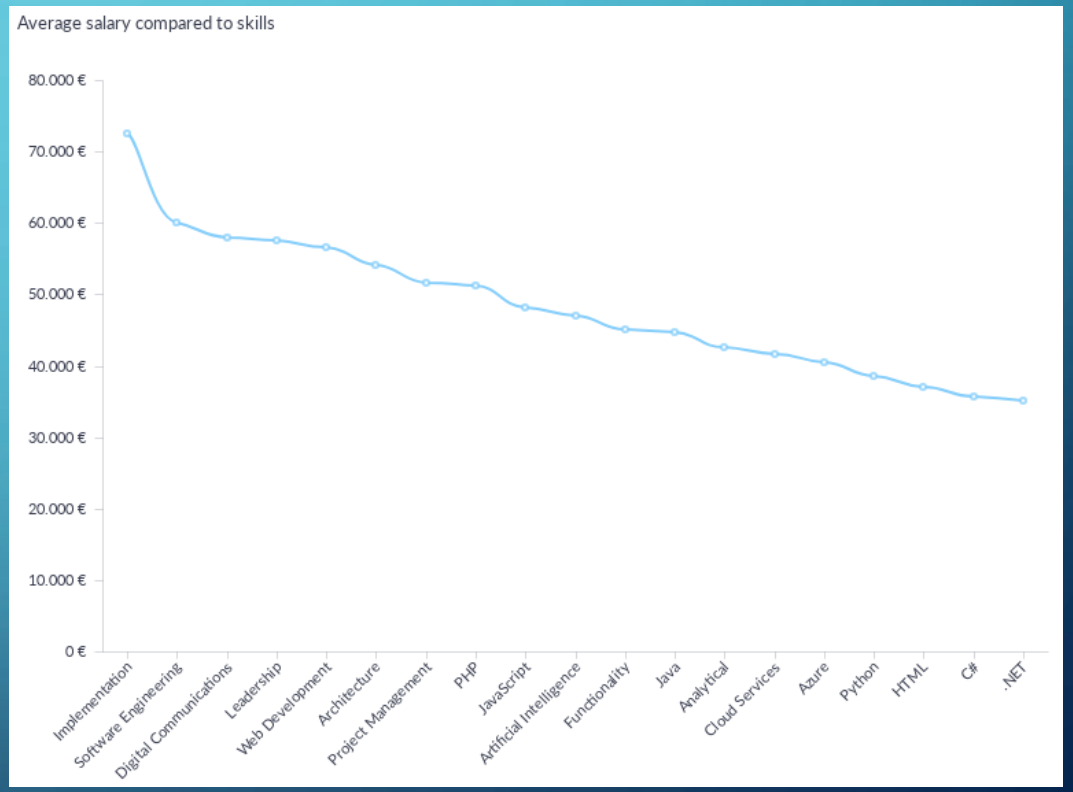


# Ανάλυση Δεδομένων με χρήση του Cluvio



# Παραδείγματα Γραφημάτων





# Συμπεράσματα και Περιορισμοί

- Η αυτοματοποιημένη εξαγωγή περιεχομένου από ιστοσελίδες αγγελιών είναι εφικτή
- Μέσω της συλλογής δεδομένων και της δημιουργίας γραφημάτων προστίθεται ένα επιπλέον μέσο στο οπλοστάσιο των αναζητούντων εργασία
- Η περιήγηση και η ανάκτηση είναι μεν αυτοματοποιημένη δεν σημαίνει όμως πως είναι μια διαδικασία που δεν χρειάζεται καμία επίβλεψη
- Παρά την ύπαρξη των τεχνολογικών μέσων για την αυτοματοποιημένη ανάκτηση οι περιορισμοί τίθενται από τους όρους χρήσης της εκάστοτε υπηρεσίας, το οποίο αποτελεί και τον σημαντικότερο ηθικό περιορισμό

# Μελλοντικές Επεκτάσεις

- Σε περίπτωση που απαιτείται η ανάκτηση των δεδομένων σε σύντομα χρονικά διαστήματα χρειάζεται μια διαφορετική προσέγγιση
- Η ανάλυση περιεχομένου και η αυτόματη χαρτογράφηση μπορεί να επεκταθεί με τη χρήση NLP ή ειδικών αλγορίθμων για την συσχέτιση αλφαριθμητικών

Σας ευχαριστώ πολύ
για την προσοχή σας!