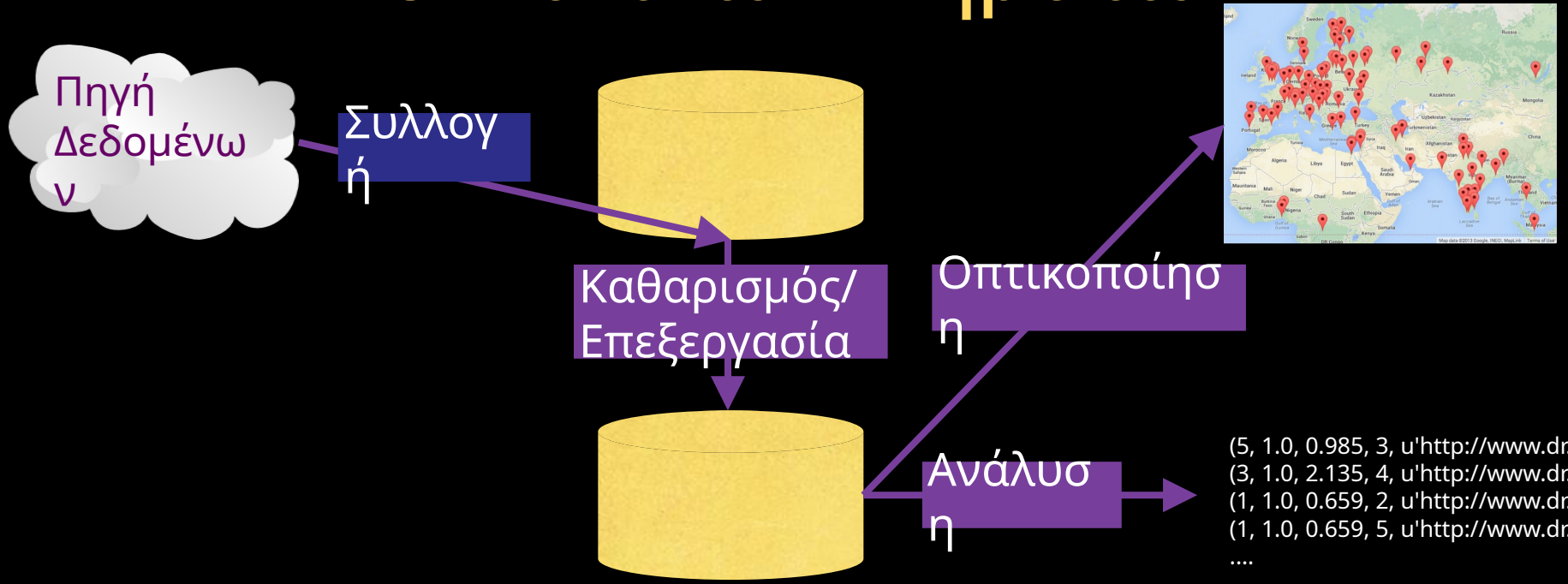
Ανάκτηση και Οπτικοποίηση Δεδομένων

Κεφάλαιο 16

Python για Όλους

www.py4e.com

Ανάλυση Δεδομένων Πολλαπλών Βημάτων



Τεχνολογίες Εξόρυξης Πολλών Δεδομένων

- https://hadoop.apache.org/
- http://spark.apache.org/
- https://aws.amazon.com/redshift/
- http://community.pentaho.com/
- ....

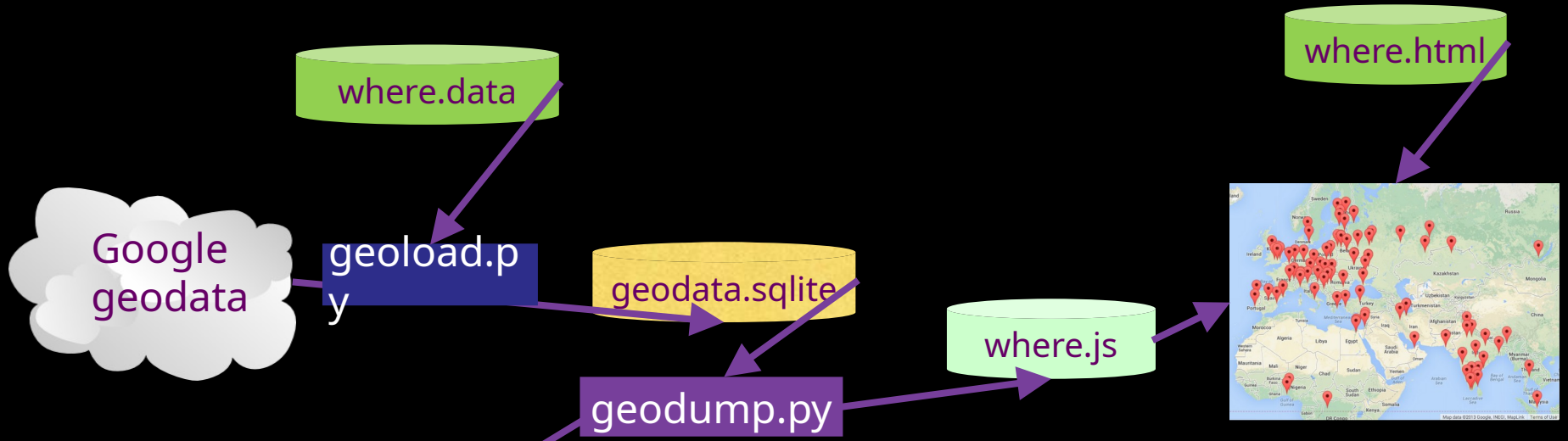
«Εξόρυξη Προσωπικών Δεδομένων»

Ο στόχος μας είναι να σας κάνουμε καλύτερους προγραμματιστές - όχι να σας κάνουμε ειδικούς εξόρυξης δεδομένων

GeoData

- Δημιουργία ενός Χάρτη Google από δεδομένα που έχουν εισαχθεί από τον χρήστη
- Χρήση του Google Geodata API
- Αποθηκεύει τα δεδομένα σε μια βάση δεδομένων για να αποφύγει τον περιορισμό πρόσβασης και να επιτρέπει την επανεκκίνηση
- Οπτικοποιημένα σε ένα πρόγραμμα περιήγησης χρησιμοποιώντας το Google Maps API

http://www.py4e.com/code3/geodata.zip

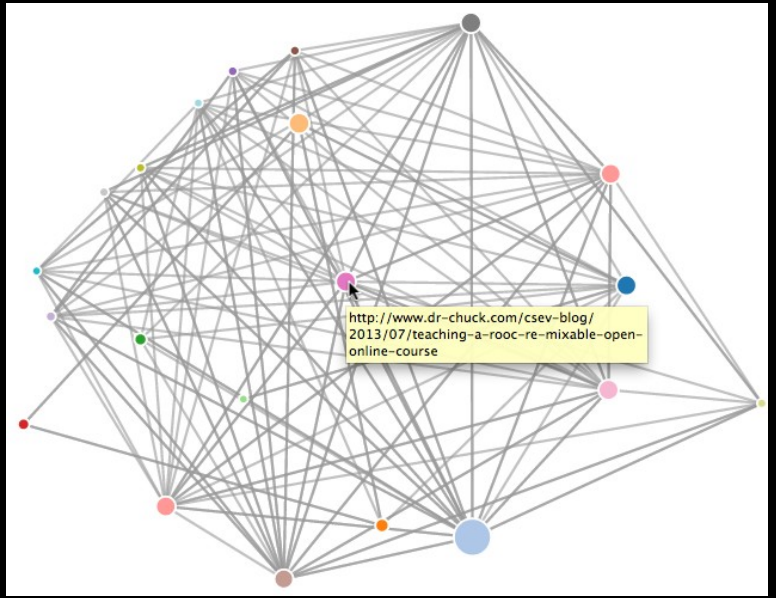


Northeastern University, ... Boston, MA 02115, USA 42.3396998 -71.08975
Bradley University, 1501 ... Peoria, IL 61625, USA 40.6963857 -89.6160811
...
Technion, Viazman 87, Kesalsaba, 32000, Israel 32.7775 35.0216667
Monash University Clayton ... VIC 3800, Australia -37.9152113 145.134682
Kokshetau, Kazakhstan 53.2833333 69.3833333
...
12 εγγραφές γράφτηκαν στο where.js
Ανοίξτε το where.html για να δείτε τα δεδομένα σε ένα πρόγραμμα περιήγησης

http://www.py4e.com/code3/geodata.zip

Page Rank - Κατάταξη Σελίδας

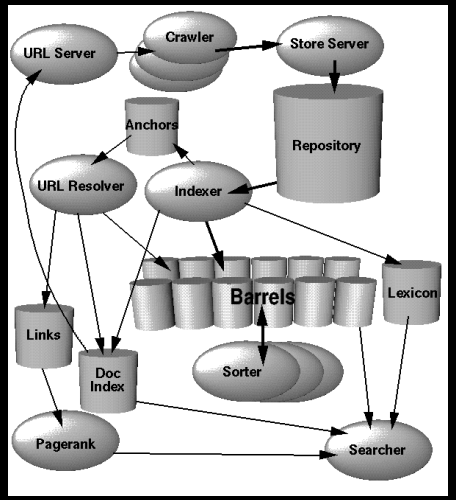
- Γράψτε ένα απλό πρόγραμμα ανίχνευσης ιστοσελίδων (crawler)
- Υπολογίστε μια απλή έκδοση του αλγορίθμου Page Rank της Google
- Οπτικοποιήστε το δίκτυο που προκύπτει



http://www.py4e.com/code3/pagerank.zip

Αρχιτεκτονική Μηχανών Αναζήτησης

- Ανίχνευση Ιστού
- Δημιουργία Ευρετηρίου
- Αναζήτηση



http://infolab.stanford.edu/~backrub/google.html

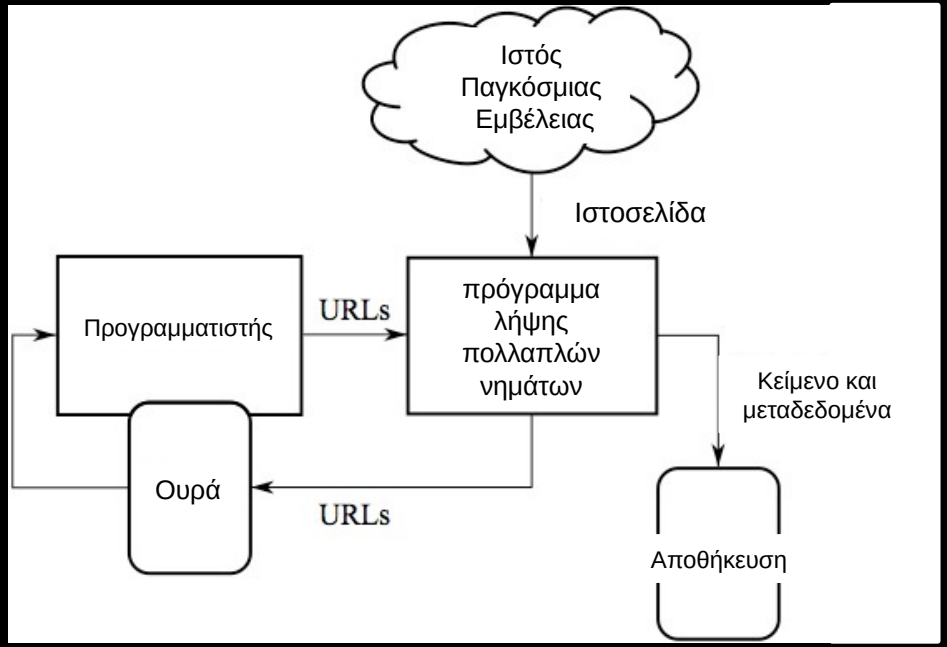
Πρόγραμμα Ανίχνευσης Ιστού

Το πρόγραμμα ανίχνευσης Ιστού είναι ένα πρόγραμμα υπολογιστή που περιηγείται στον Παγκόσμιο Ιστό με μεθοδικό, αυτοματοποιημένο τρόπο. Τα προγράμματα ανίχνευσης Ιστού χρησιμοποιούνται κυρίως για τη δημιουργία αντιγράφου όλων των σελίδων που επισκέπτεστε για μετέπειτα επεξεργασία από μια μηχανή αναζήτησης που θα ευρετηριάσει τις ληφθείσες σελίδες για να παρέχει γρήγορες αναζητήσεις.

http://en.wikipedia.org/wiki/Web_crawler

Ανιχνευτής Ιστού

- Ανακτά μια σελίδα
- Αναζητά συνδέσμους σε ολόκληρη τη σελίδα
- Προσθέτει τους συνδέσμους σε μια λίστα ιστότοπων «προς ανάκτηση»
- Επαναλαμβάνει...



http://en.wikipedia.org/wiki/Web_crawler

Πολιτική Ανίχνευσης Ιστού

- μια πολιτική επιλογής που δηλώνει ποιες σελίδες θα κατεβάσετε,
- μια πολιτική επαν-επίσκεψης που δηλώνει πότε πρέπει να ελέγχεται για αλλαγές στις σελίδες,
- μια πολιτική ευγένειας που δηλώνει πώς να αποφύγετε την υπερφόρτωση Ιστοσελίδων και
- μια πολιτική παραλληλισμού που δηλώνει τον τρόπο συντονισμού των κατανεμημένων ανιχνευτών Ιστού

robots.txt

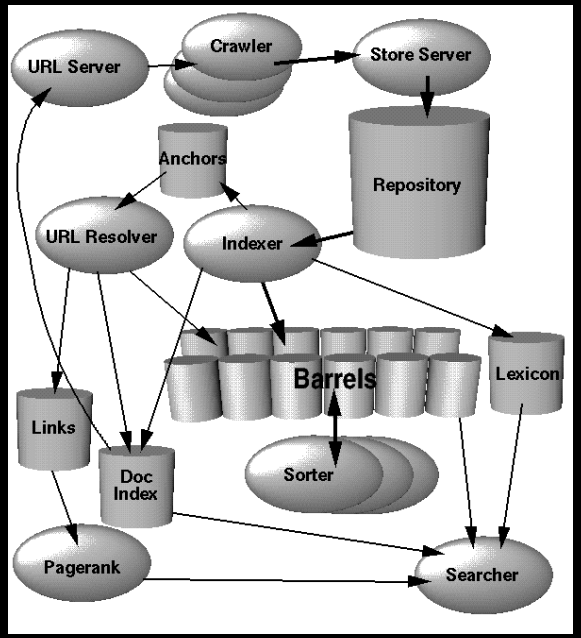
- Ένας τρόπος επικοινωνίας ενός ιστότοπου με προγράμματα ανίχνευσης ιστού
- Ένα άτυπο και εθελοντικό πρότυπο
- Μερικές φορές οι άνθρωποι κάνουν μια «παγίδα αράχνης» (Spider Trap) για να πιάσουν «κακές» αράχνες

```
User-agent: *
Disallow: /cgi-bin/
Disallow: /images/
Disallow: /tmp/
Disallow: /private/
```

http://en.wikipedia.org/wiki/Robots_Exclusion_Standard
http://en.wikipedia.org/wiki/Spider_trap

Αρχιτεκτονική Google

- Ανίχνευση Ιστού
- Δημιουργία Ευρετηρίου
- Αναζήτηση

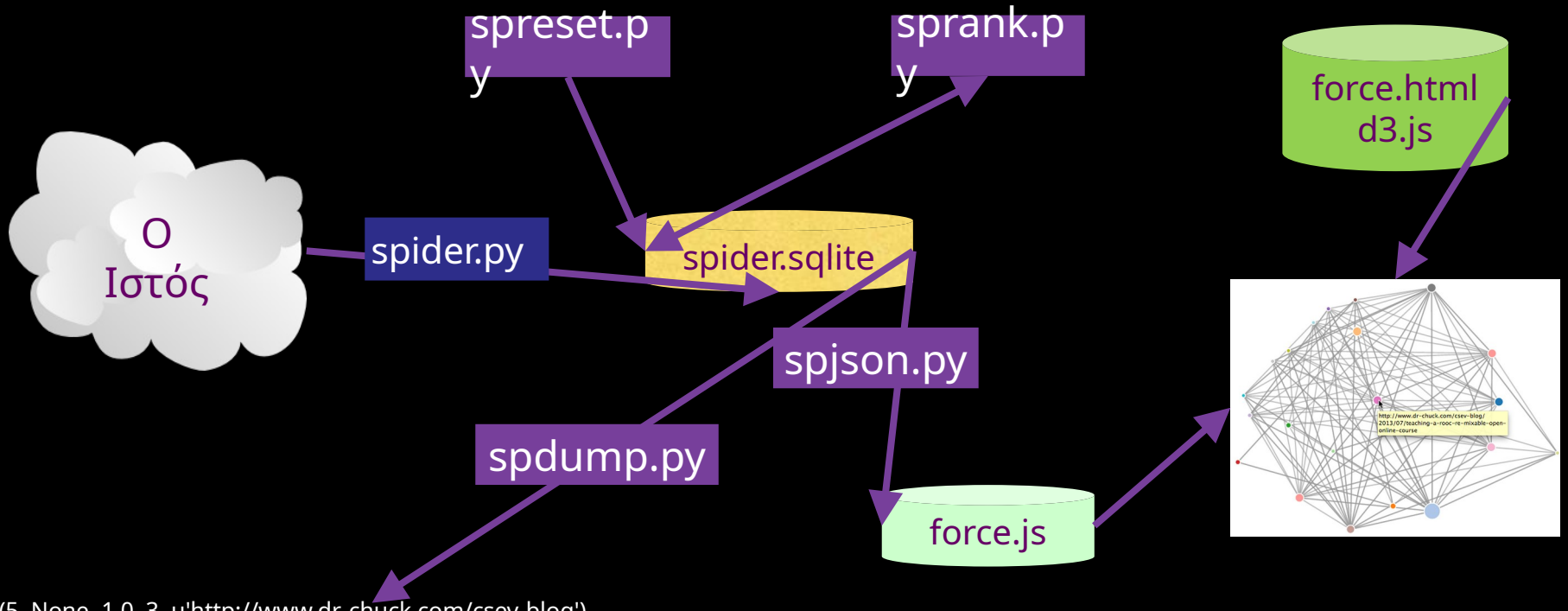


http://infolab.stanford.edu/~backrub/google.html

Ευρετηρίαση Αναζήτησης

Η ευρετηρίαση των μηχανών αναζήτησης συλλέγει, αναλύει και αποθηκεύει δεδομένα για να διευκολύνει τη γρήγορη και ακριβή ανάκτηση πληροφοριών. Ο σκοπός της αποθήκευσης ενός ευρετηρίου είναι η βελτιστοποίηση της ταχύτητας και της απόδοσης στην εύρεση σχετικών εγγράφων για ένα ερώτημα αναζήτησης. Χωρίς ευρετήριο, η μηχανή αναζήτησης θα σαρώνει κάθε έγγραφο στον ιστό, το οποίο θα απαιτούσε σημαντικό χρόνο και υπολογιστική ισχύ.

http://en.wikipedia.org/wiki/Index_(search_engine)



(5, None, 1.0, 3, u'http://www.dr-chuck.com/csev-blog')
(3, None, 1.0, 4,
u'http://www.dr-chuck.com/dr-chuck/resume/speaking.htm')
(1, None, 1.0, 2, u'http://www.dr-chuck.com/csev-blog/')
(1, None, 1.0, 5, u'http://www.dr-chuck.com/dr-chuck/resume/index.htm')
4 γραμμές..

http://www.py4e.com/code3/pagerank.zip

Λίστες Αλληλογραφίας - Gmane

- Ανιχνεύει το αρχείο μιας λίστας αλληλογραφίας
- Κάνει κάποια ανάλυση / καθαρισμό
- Οπτικοποιεί τα δεδομένα ως νέφος λέξεων και γραμμές

http://www.py4e.com/code3/gmane.zip

Προσοχή: Αυτό το Σύνολο Δεδομένων είναι > 1GB

- Μην κατευθύνετε απλώς αυτήν την εφαρμογή στο gmane.org και την αφήσετε να εκτελεστεί
- Δεν υπάρχει όριο τιμών - αυτοί είναι καλοί άνθρωποι

Χρησιμοποιήστε αυτό για τις δοκιμές σας:

http://mbox.dr-chuck.net/sakai.devel/4/5

gmane.py

content.sqlite

gword.htm
d3.js

gword.js

mapping.sqlite

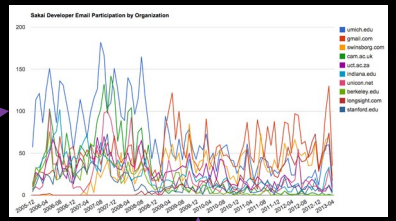
gmodel.py

gword.py

content.sqlite

gbasic.py

gline.py



gline.js

gline.htm
d3.js

How many to dump? 5
Loaded messages= 51330 subjects= 25033 senders= 1584
Top 5 Email list participants
steve.swinsburg@gmail.com 2657
azeckoski@unicon.net 1742
ieb@tfd.co.uk 1591
csev@umich.edu 1304
david.horwitz@uct.ac.za 1184
...

http://www.py4e.com/code3/gmane.zip

Ευχαριστίες / Συνεισφορές

Αρχική ανάπτυξη : Charles Severance, University of Michigan School of Information

Απόδοση στα Ελληνικά: Κιουρτίδου Δ. Κωνσταντία

… Εισαγάγετε νέους Μεταφραστές και άτομα που έχουν συνεισφέρει εδώ

Συνέχεια…