ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
UNIVERSITY OF PATRAS

ΔΙΑΧΕΙΡΙΣΗ ΠΕΡΙΕΧΟΜΕΝΟΥ ΣΤΟΝ ΠΑΓΚΟΣΜΙΟ ΙΣΤΟ & ΓΛΩΣΣΙΚΑ ΕΡΓΑΛΕΙΑ

19/2/2013

1ο ΕΙΣΑΓΩΓΙΚΕΣ ΕΝΝΟΙΕΣ

Αντικείμενο του Μαθήματος

- Εφαρμογές και εργαλεία ΓλωσσικήςΤεχνολογίας με στόχο
  - τη βελτίωση της πρωτογενούς λειτουργίας των συστημάτων διαχείρισης δυναμικών δεδομένων κειμένου
  - την παροχή προηγμένων υπηρεσιών προς τους χρήστες που αλληλεπιδρούν με τα δεδομένα αυτά.

θεματικές ενότητες

- *Γλώσσα και Αναζήτηση Πληροφορίας*
- *Text Mining*
- *N-grams*
- *Data Mining*
- *Opinion Mining*
- *Δυναμική Εξαγωγή Πληροφορίας από τα Δεδομένα Διαδικτύου*
- *Οντολογίες*
- *Question Answering*
- *Ο πολυγλωσσικός Παγκόσμιος Ιστός*

Εξέταση του Μαθήματος

- εκπόνηση ενός project(80%)
  - Υποχρεωτική παράδοση
    - στην ημερομηνία που θα ανακοινωνθεί
  - Ο βαθμός του project δεν διατηρείται για την επόμενη χρονιά
- συμμετοχή σε προφορική εξέταση (20%)
  - Μπορεί να επαναληφθεί στην εξεταστική Σεπτεμβρίου

Που, Πότε, Ποιος, Τι…

- Που:
  - Αίθουσα Β3
- Πότε
  - Δευτέρα
  - 11:00 -13:00
- Ώρες γραφείου:
  - Δευτέρα 13:00 -14:00
- Email:
  - kozanid@ceid.upatras.gr
  - simaki@ceid.upatras.gr

- Προτεινόμενη βιβλιογραφία
  - "Υπολογιστική Γλωσσολογία" Αλεξανδρή Χριστίνα, εκδόσεις παπασωτηρίου
  - Speech and Language Processing (2nd Edition) by Daniel Jurafsky and James H. Martin (ISBN-13: 9780131873216)
  - *Dive Into Python*
  - *Natural Language Processing with Python*
- Επιπλέον εκπαιδευτικό υλικό στο site του μαθήματος
  - Διαφάνειες
  - Επιλεγμένες δημοσίεύσεις

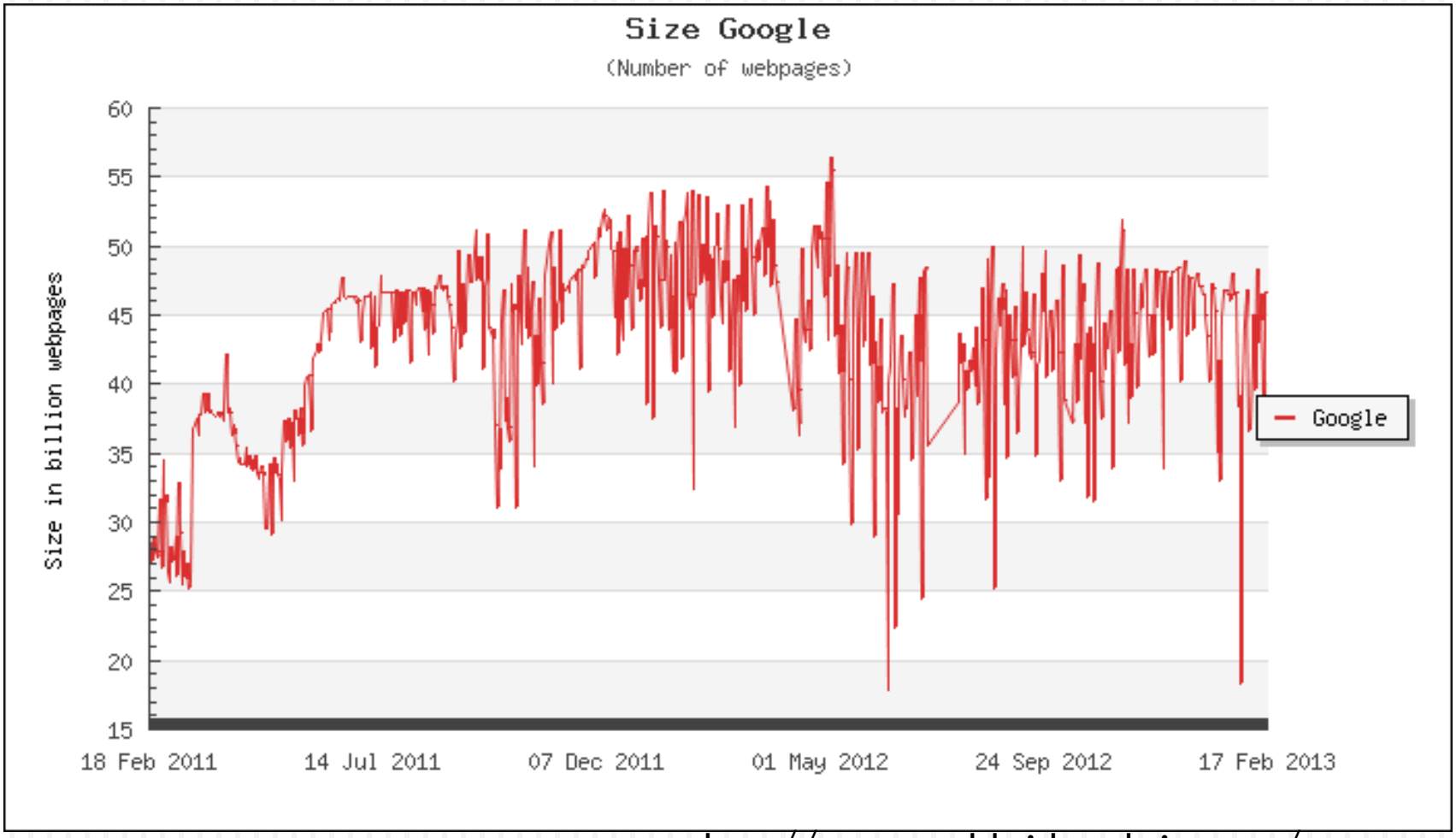
Επεξεργασία Φυσικής Γλώσσας

- Κλάδος της επιστήμης των υπολογιστών,
- Αξιοποιεί τεχνικές, τεχνητής νοημοσύνης και γλωσσολογίας,
- **Σκοπός :** η διευκόλυνση της αλληλεπίδρασης των υπολογιστών και των ανθρώπων με την χρήση φυσικών γλωσσών

Ιστορική Αναδρομή

- 1950 Turing test
- 1954 Georgetown experiment
- 1964-66 ELIZA bot
- During the 70's 'conceptual ontologies',
  - real-world information into computer-understandable data
- 1980, εισαγωγή machine learning αλγορίθμων για NLP
- statistical models,
- Recent research unsupervised / semi-supervised learning algorithms

Ο παγκόσμιος ιστό σήμερα

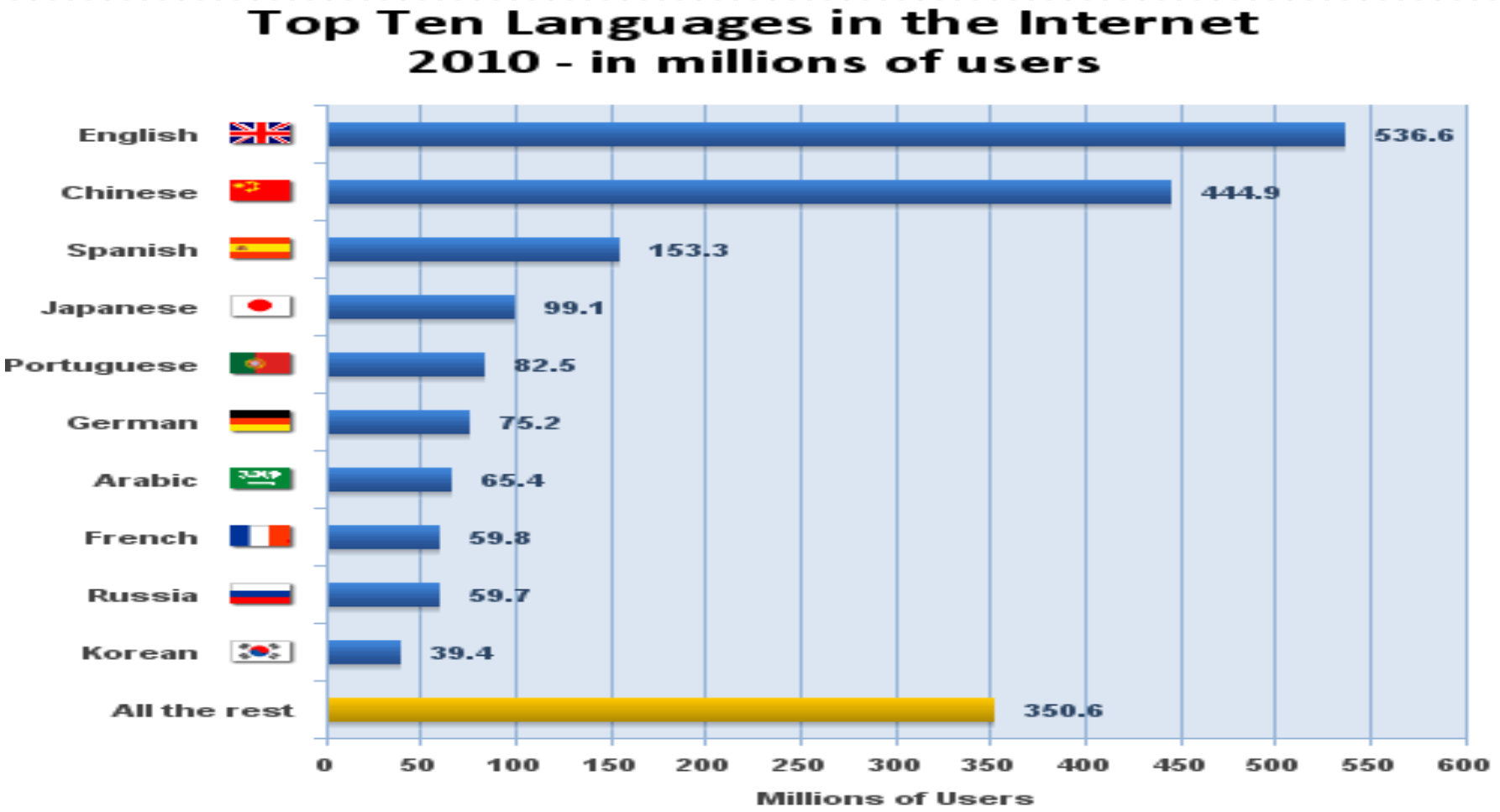


http://www.worldwidewebsize.com/

Ο παγκόσμιος ιστός σήμερα

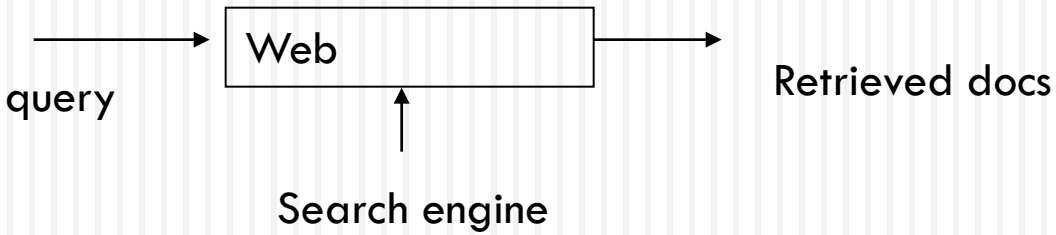
- Το μέγεθος του Παγκόσμιου Ιστού μεταβάλλεται συνεχώς.
- Surface Web
  - 11.5 billion web pages (2005)
    - http://www.cs.uiowa.edu/~asignori/web-size
  - 48 billion web pages (2013)
- Deep Web
  - Τουλάχιστον 538.5 billion deep web pages (2005)
- Γλώσσες στον Παγκόσμιο Ιστό (2002)
  - Αγγλικά 56.4%
  - Γερμανικά 7.7%
  - Γαλλικά 5.6%
  - Γιαπωνέζικα 4.9%

Χρήση διαφορετικών γλωσσών



Γλωσσική Ανάλυση στο Web

- Ο Παγκόσμιος Ιστός παρέχει πληροφορία
- Η πληροφορία περιέχεται κυρίως σε κείμενα
- Πρόσβαση στην πληροφορία μέσω ερωτημάτων προς μηχανές αναζήτησης



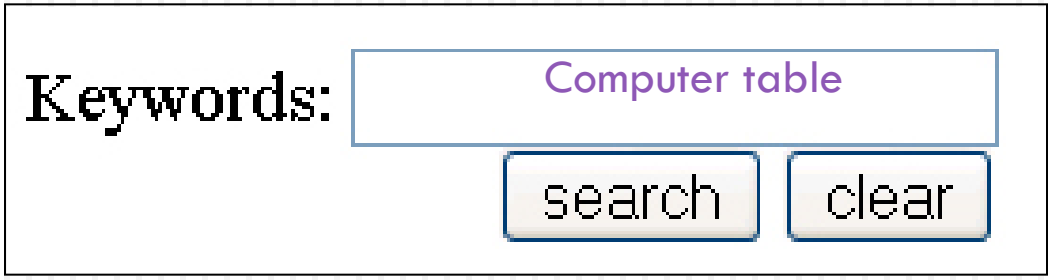
Τι είναι η Γλωσσική Ανάλυση;

- Επεξεργασία κειμένου
  - Λεξική: tokenization, PoS tagging
  - Συντακτική
  - Σημασιολογική
  - Λόγου (επίλυση αναφοράς)
- Επεξεργασία φωνής
  - Φωνητική μεταγραφή
  - Προσωδία
  - Εκφορά

Πού χρειάζεται η Γλωσσική Ανάλυση;

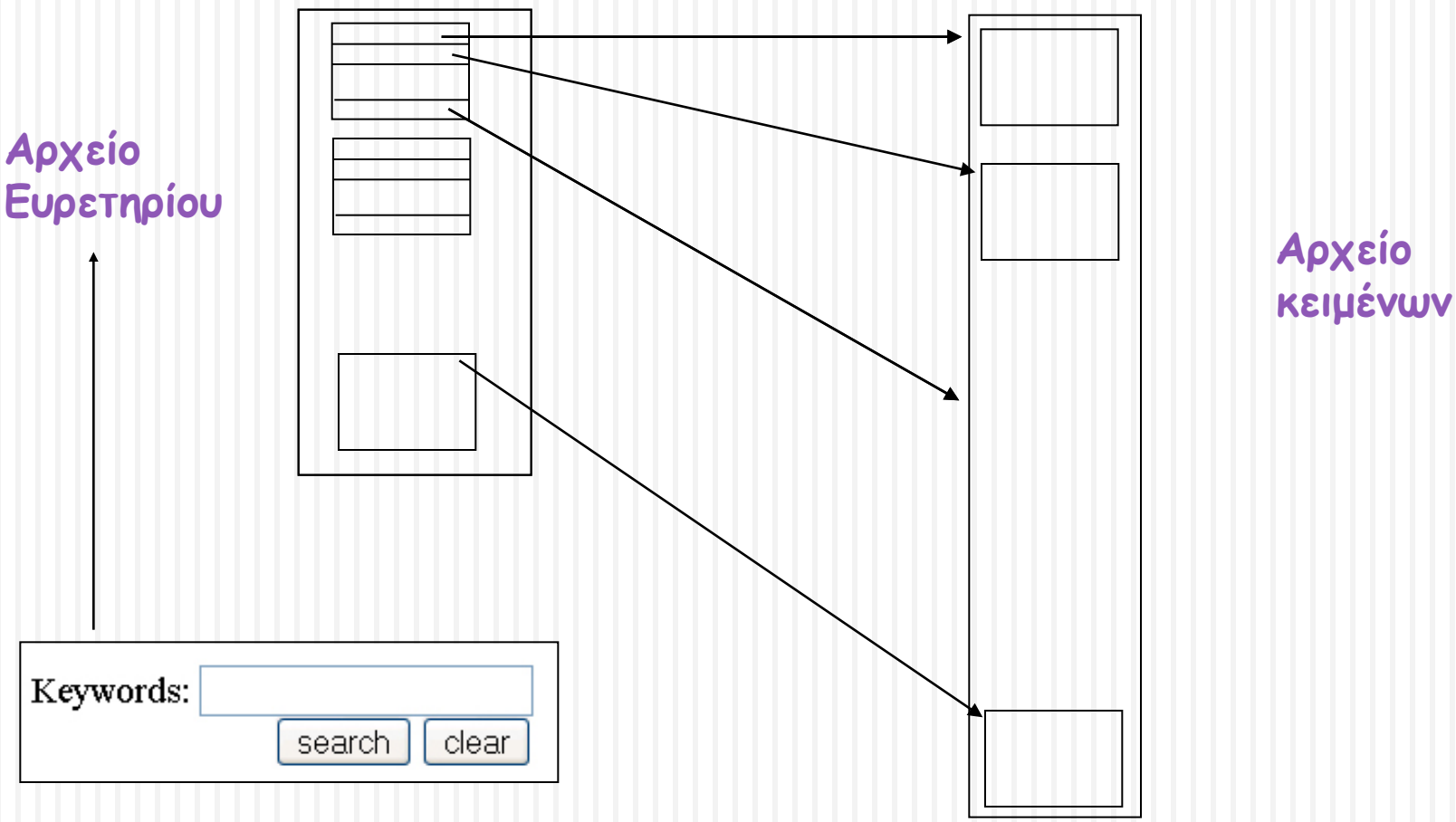
- Επεξεργασία ερωτημάτων
- Δεικτοδότηση και αναζήτηση ιστοσελίδων
- Μηχανική μετάφραση ιστοσελίδων
- Web Question - Answering
- Ανάλυση Weblogs
- Εξαγωγή πληροφορίας από δεδομένα
- Εξόρυξη γνώσης

Επεξεργασία Ερωτημάτων



- Περιέχει λέξεις (δομικά στοιχεία των γλωσσών)
- Οι λέξεις απαρτίζουν μεγαλύτερες μονάδες –φράσεις. Μερικές φορές αποτελούν από μόνες τους ολόκληρες προτάσεις.

Δεικτοδότηση Ιστοσελίδων

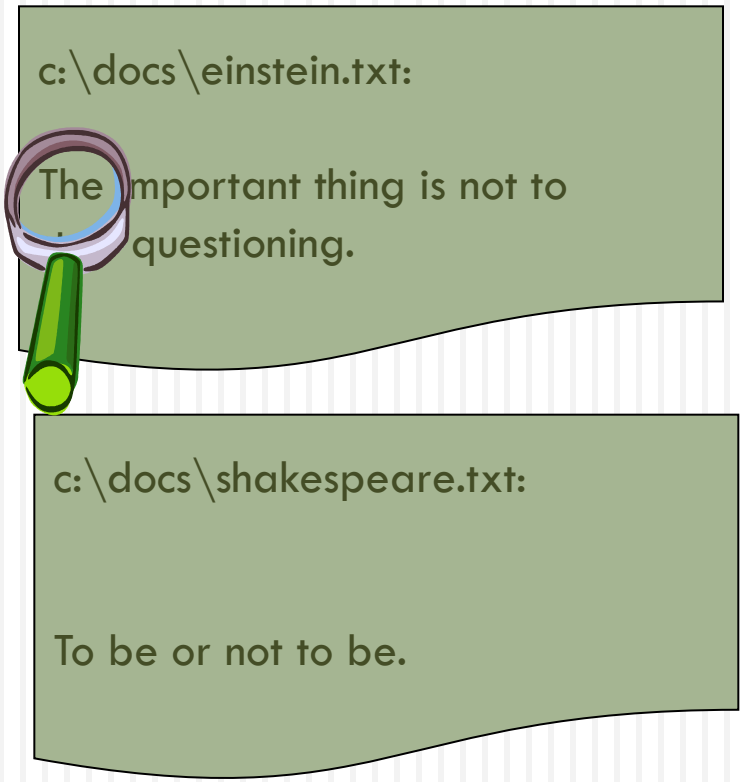


Αναζήτηση πληροφορίας

Ερώτημα: **not**

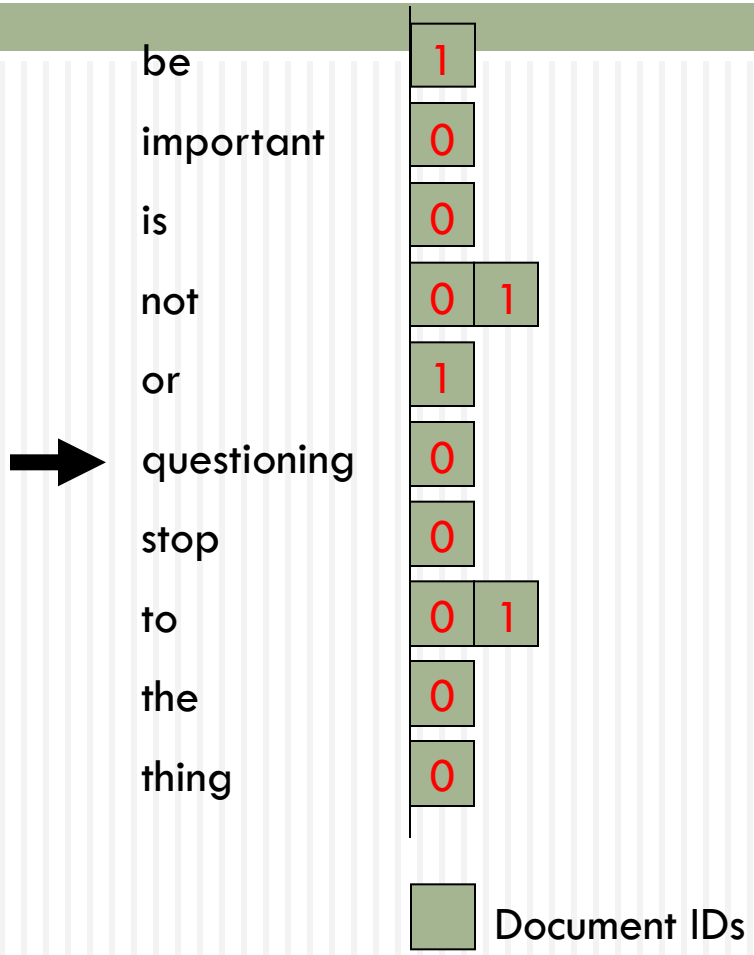
Σύγκριση συμβολοσειρών αργή!

Λύση: **Inverted index**



Inverted index

Query: **not**



c:\docs\einstein.txt: **0**

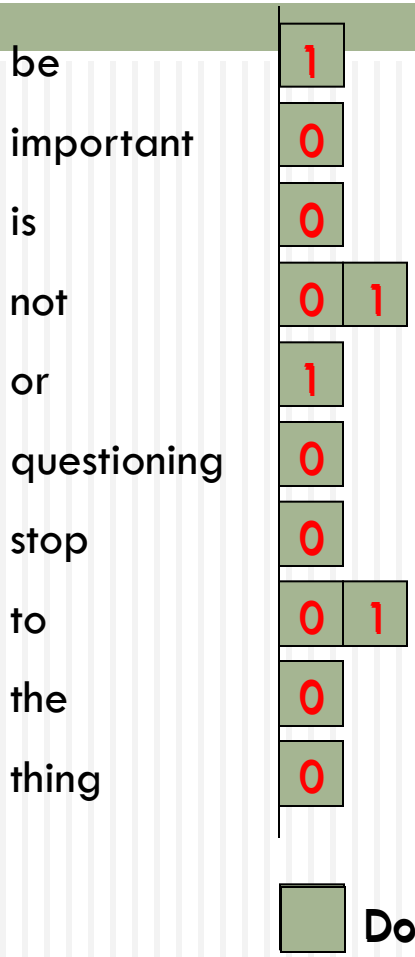
The important thing is not to stop questioning.

c:\docs\shakespeare.txt: **1**

To be or not to be.

Inverted index

Query: "not to"



c:\docs\einstein.txt: 0

0 1 2 3 4 5

The important thing is not to stop questioning.

6 7

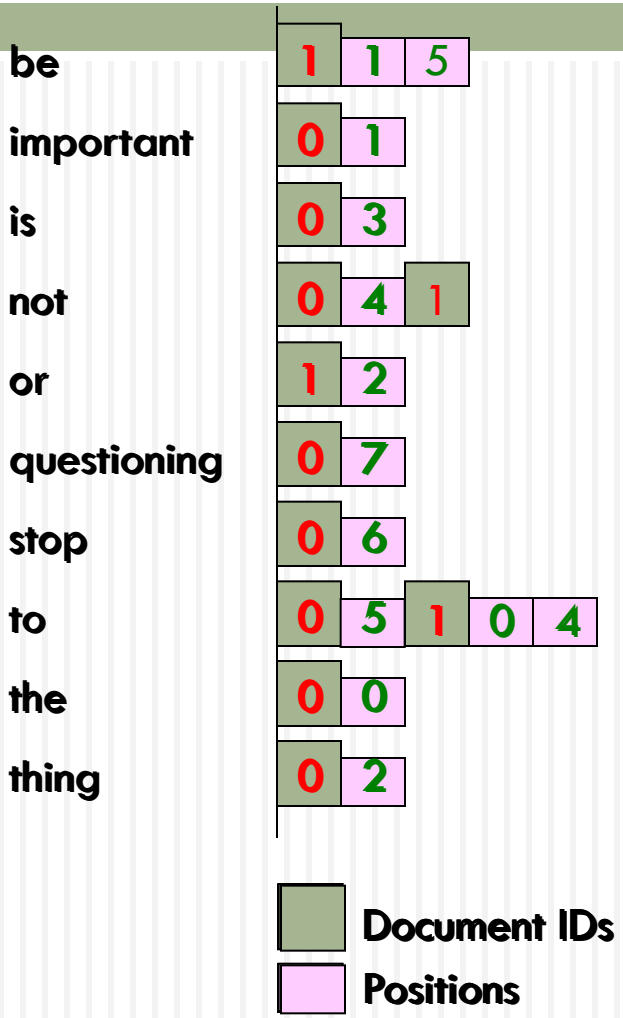
c:\docs\shakespeare.txt: 1

0 1 2 3 4 5

To be or not to be.

Inverted index

Query: "not to"



c:\docs\einstein.txt: 0

The important thing is not to stop questioning.

c:\docs\shakespeare.txt: 1

To be or not to be.

Μηχανική Μετάφραση

Killing Palestinians and wounding nine in the raids Sector

Nine Palestinians were wounded among civilians in an Israeli air raid in the neighborhood result in the Gaza Strip. This comes immediately after the killing of two prominent Al-Aqsa Martyrs Brigades in the Israeli occupying forces carried out air and infantry forces in the Balata camp in the West Bank.

Bashir meets Fraser, the Security Council will not impose forces Darfur

Is scheduled to meet with Sudanese President Omar al-Bashir Jenday Fraser Assistant Minister for Foreign Affairs of the American attempt to persuade officials in Khartoum, Sudanese Darfur deployment of the nationalities. For his part, US Ambassador to the United Nations that it has no intention of the Security Council to impose its forces in the province.

Rmsfield and Cheney insist on keeping the American forces in Iraq

Called American Defense Minister Donald Rmsfield Americans to show patience on Iraq. I take Vice President Dick Cheney calls Democrats withdrawal of American forces from Iraq link and the possibility of early withdrawal of attacks inside the United States.

Killing civilians and wounding officer suicide attack in Afghanistan

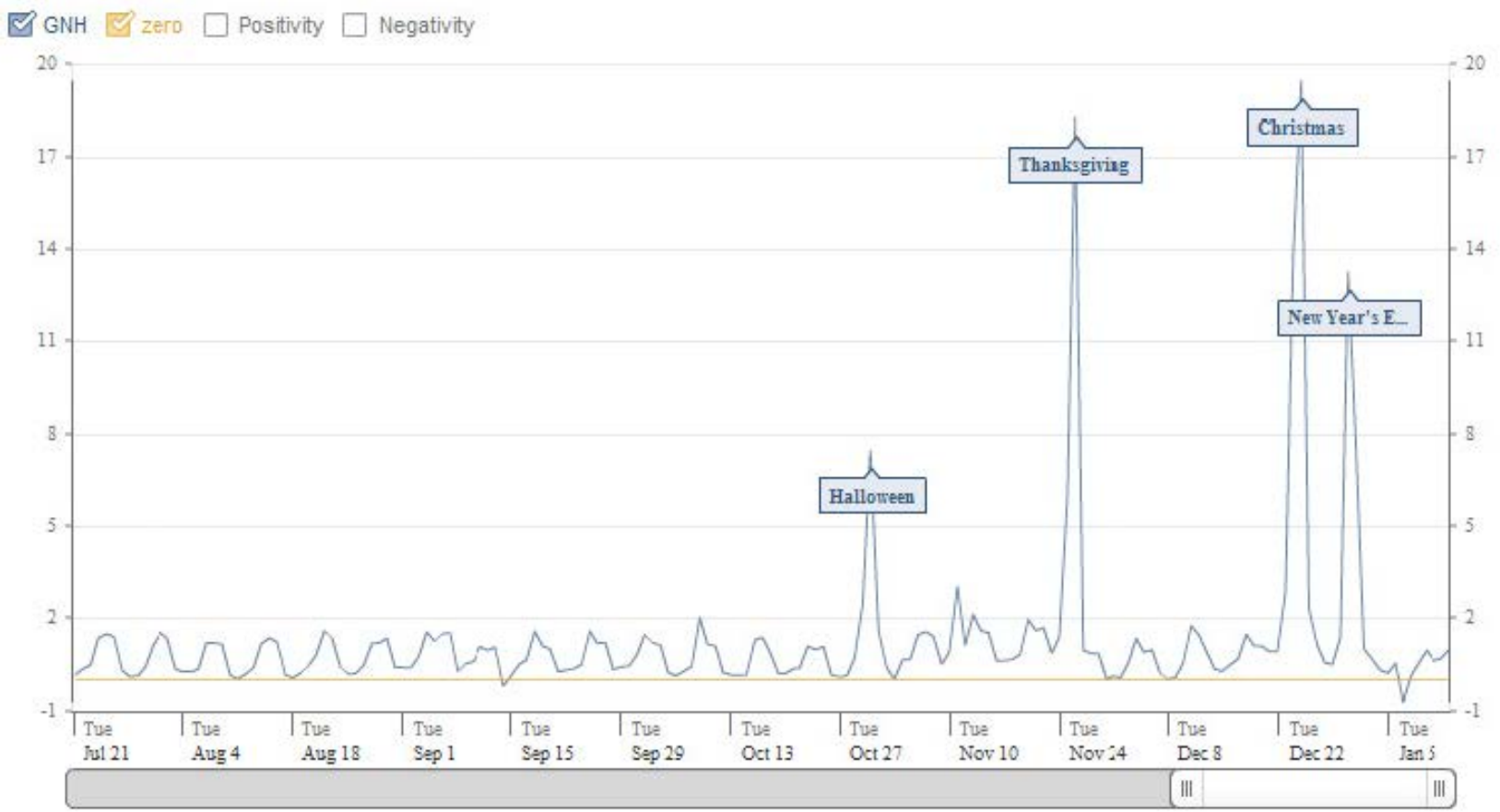
The international force to help establish security (ISAF) killed civilians and the wounding of an officer in an attack against Afghan forces convoy south Atlantic Afghanistan. In the capital Kabul, a hand grenade exploded at the passage of manufacture French patrol was not reported injuries or damage.

Web Question - Answering

Ανάλυση Weblogs

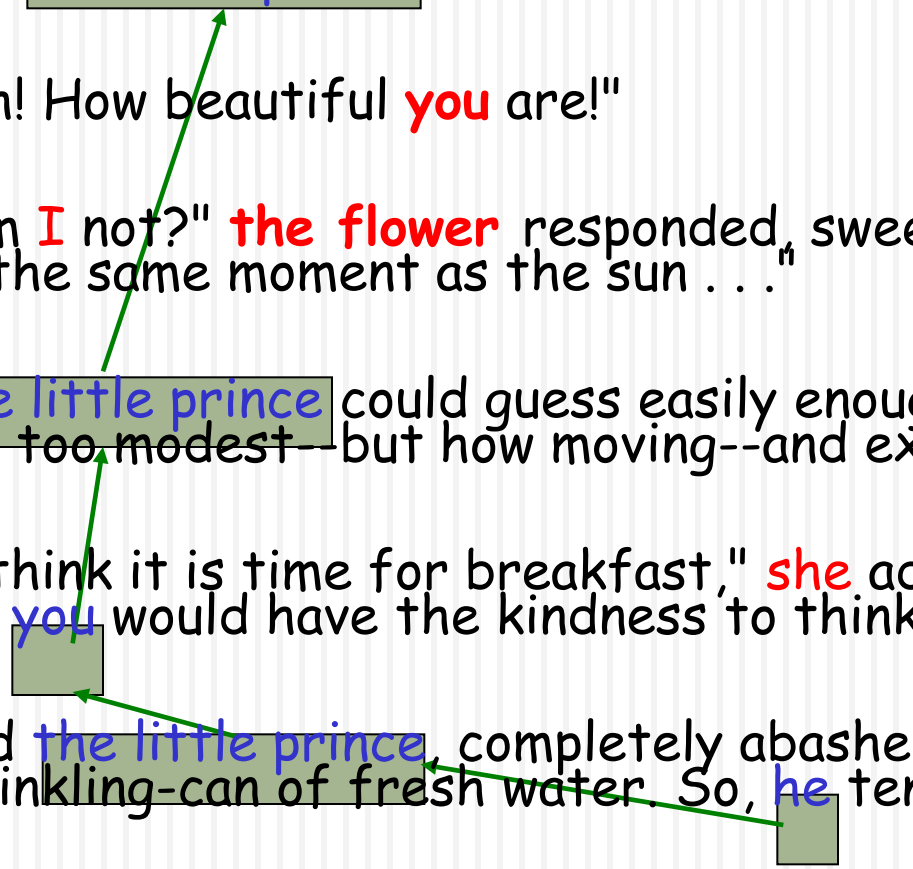
- Εξόρυξη γνώσης από Weblogs, forums, blogs, και άλλα μέσα online πληροφορίας που δημιουργούν οι χρήστες
  - Πληροφορία για marketing προϊόντων
  - Αποτύπωση πολιτικών απόψεων
  - Ανάλυση κοινωνικών δικτύων
  - Buzz analysis (επικαιρότητα, τι συζητούν τώρα οι χρήστες του web).

Facebook's "Gross National Happiness Index"



Επίλυση αναφοράς

- But the little prince could not restrain admiration:
- "Oh! How beautiful **you** are!"
- "Am I not?" **the flower** responded, sweetly. "And I was born at the same moment as the sun . . ."
- The little prince could guess easily enough that she was not ~~any too modest--~~but how moving--and exciting--she was!
- "I think it is time for breakfast," she added an instant later. "If you would have the kindness to think of my needs--"
- And the little prince, completely abashed, went to look for a sprinkling-can of fresh water. So, he tended the flower.



Αναγκαιότητα Χρήσης Γλωσσικών Εργαλείων

- Οι φυσικές γλώσσες χαρακηρίζονται από ασάφεια

ΜτΛ → **ΑΠΑΝΤΗΣΕΙΣ (ουσιαστικό ή ρήμα;)**

Σημασιολογική → **ΠΟΝΤΙΚΙ (τρωκτικό, μυς ή εξάρτημα υπολογιστή;)**

Πραγματολογική → **Φέρε τη γάτα με το γάντι**

Ασάφεια

- Βρείτε 3 διαφορετικά νοήματα για την πρόταση:
  - Είδα τον άντρα στο πάρκο με το τηλεσκόπιο

Σχετικά με το Project

- Εργασία 2 ατόμων
- Γλώσσα υλοποίησης
  - Python 2.7.3
  - Πακέτο επεξεργασίας φυσικής γλώσσας nltk 2.0
  - Περισσότερα στο Φροντιστήριο

Ερωτήσεις

Ευχαριστώ!!!