Εξαγωγή σχεσιακών πληροφοριών από τη Βικιπαίδια

Χρήστος Περιβολαρόπουλος

Τετάρτη 8 Ιουνίου 2016

Η εργασία αυτή έγινε υπο την επίβλεψη των κ. Κυριάκο Σγάρμπα του Τμήματος Ηλεκτρολόγων Μηχανικών και Τεχνολογίας Υπολογιστών Πανεπιστημίου Πατρών και κ. Boris Katz απο το InfoLab του CSAIL MiT.

Δομή

Ενας ωκεανός πληροφορίας

As of 2014 Google has indexed 200 Terabytes (TB) of data.

— *http://www.websitemagazine.com/*

The English Wikipedia is now one of 292 Wikipedia editions and holds the largest amount of articles, with more than 5,167,046,

— *wikipedia.org*

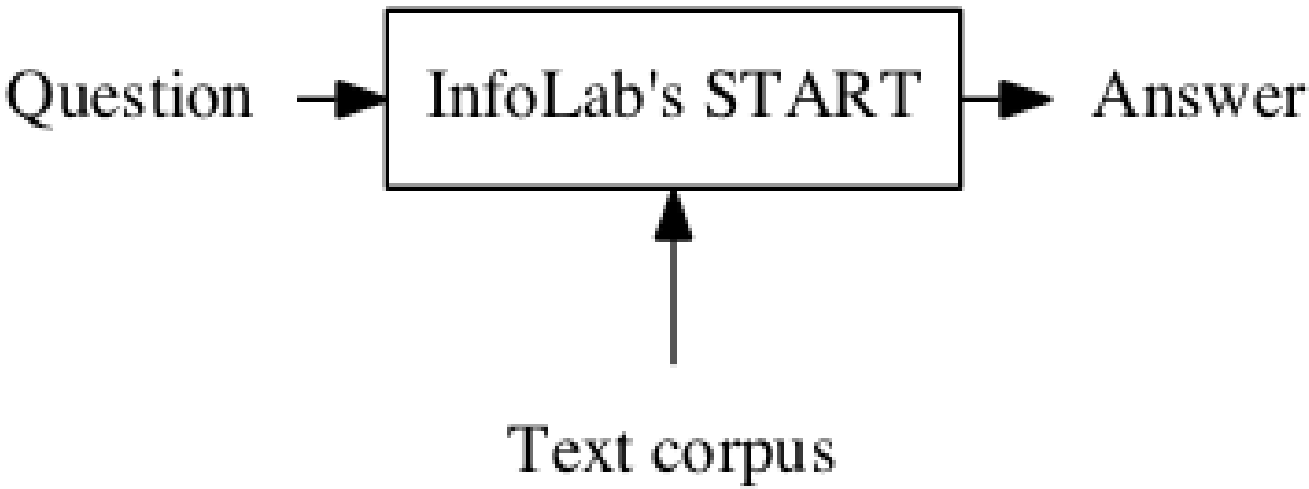
Εμπόδια στη χρήση της πληροφορίας

- Αδομητη πληροφορία κατανοήσιμη μονο απο ανθρώπους — πχ. ελεύθερο κείμενο, διαγράμματα, video
- Προυπόθεση απο το χρήστη να γνωρίζει τη δομη της πληροφορίας
- Προυπόθεση απο το χρήστη να γνωρίζει ειδικά εργαλεία και/ή τα μαθηματικά μοντέλα.

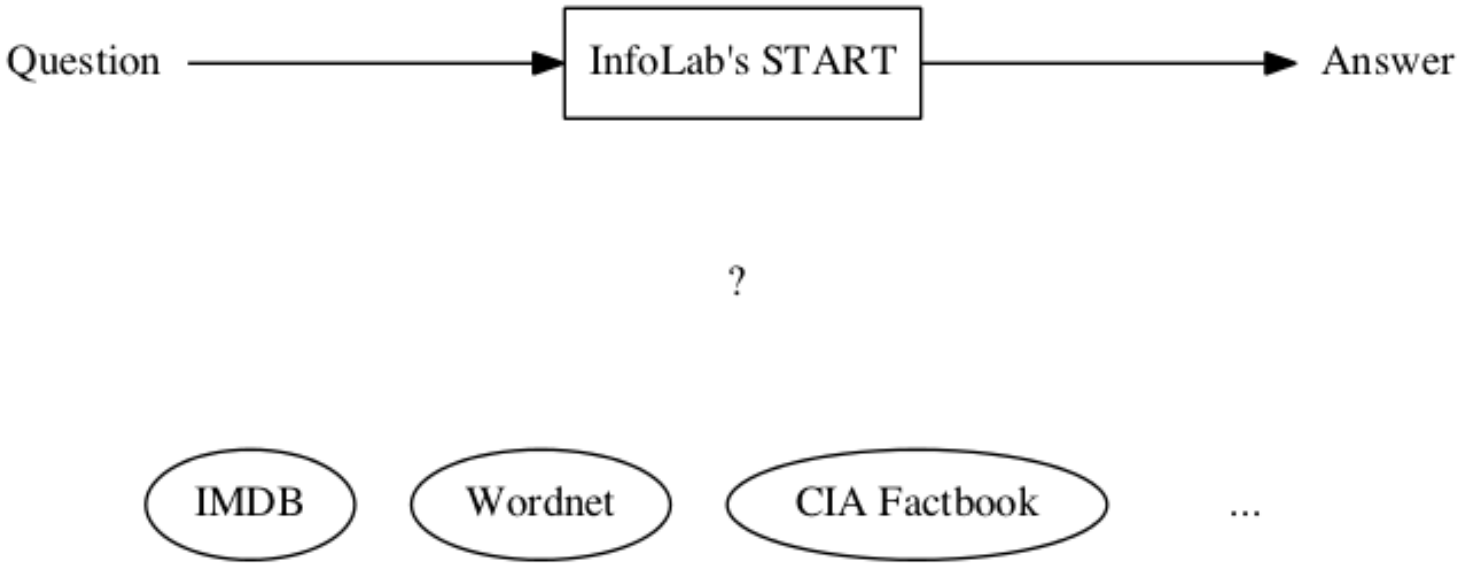
InfoLab's START

Προσβαση σε πολύπλοκες πληροφορίες απαντώντας σε ερωτήσεις σε φυσική γλώσσα.

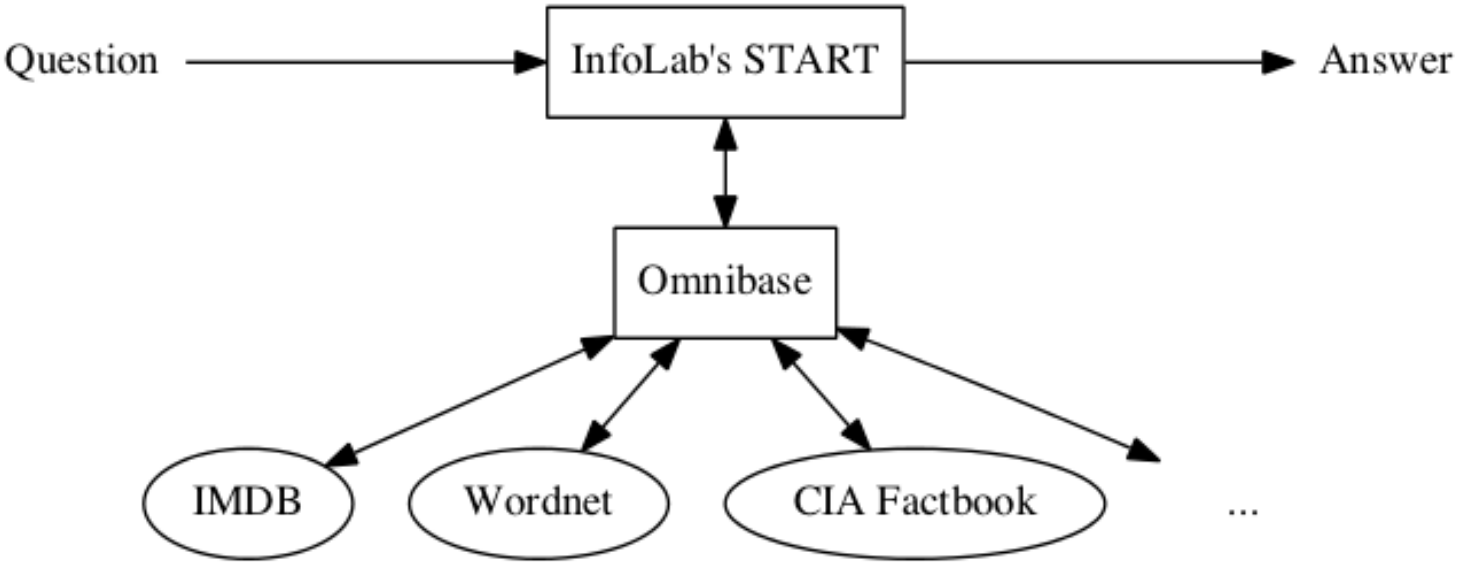
Βασική λειτουργία του START.



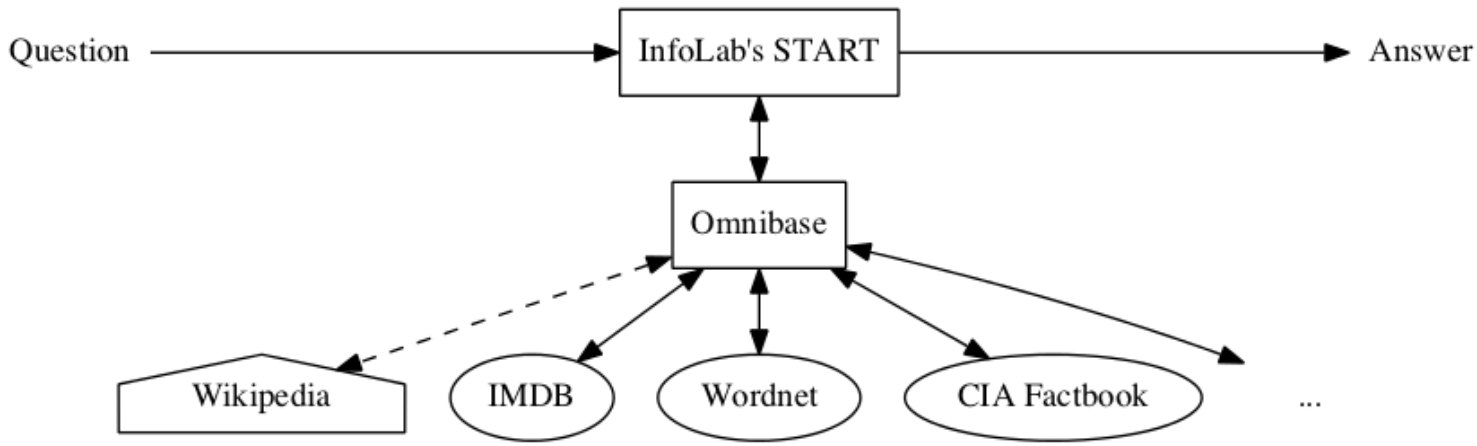
Αναγκη πρόσβασης στο διαδίκτυο.



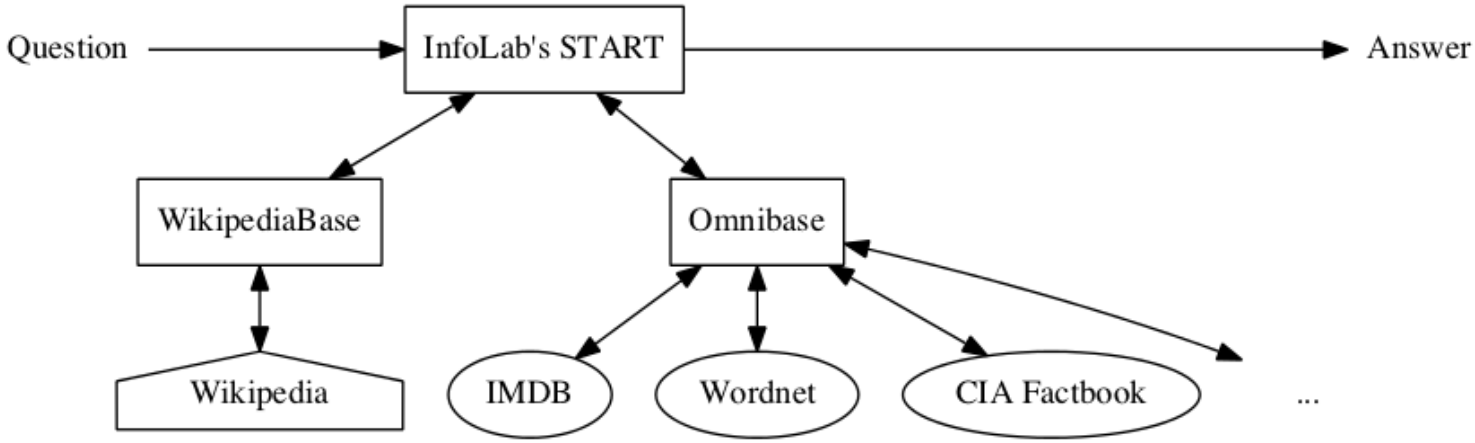
Omnibase


Question
InfoLab's START
Answer
Omnibase
IMDB
Wordnet
CIA Factbook
...

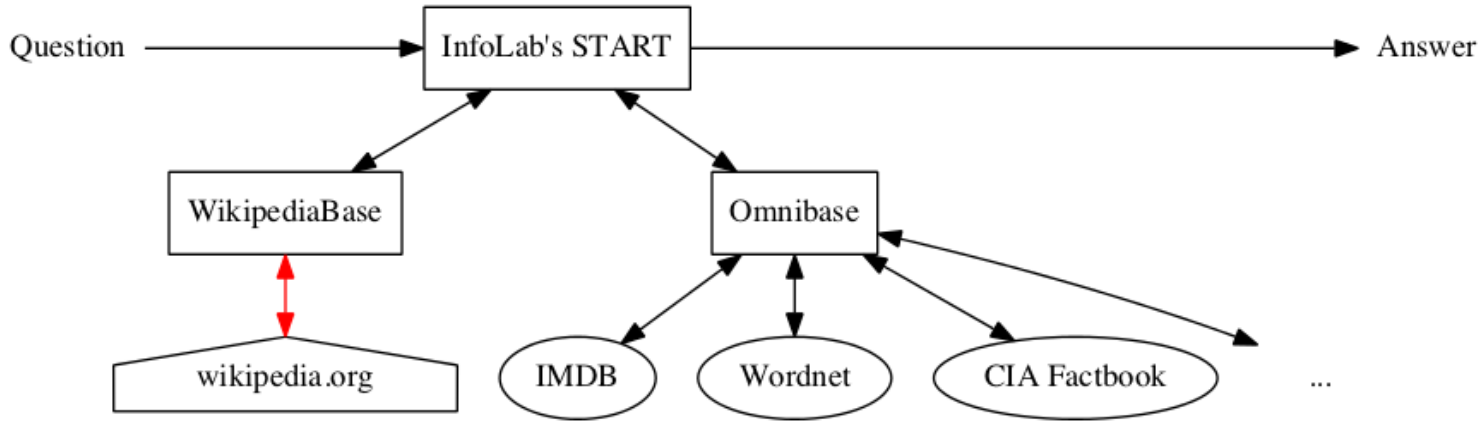
Η wikipedia χρήζει ιδιέταιρης προσωχής



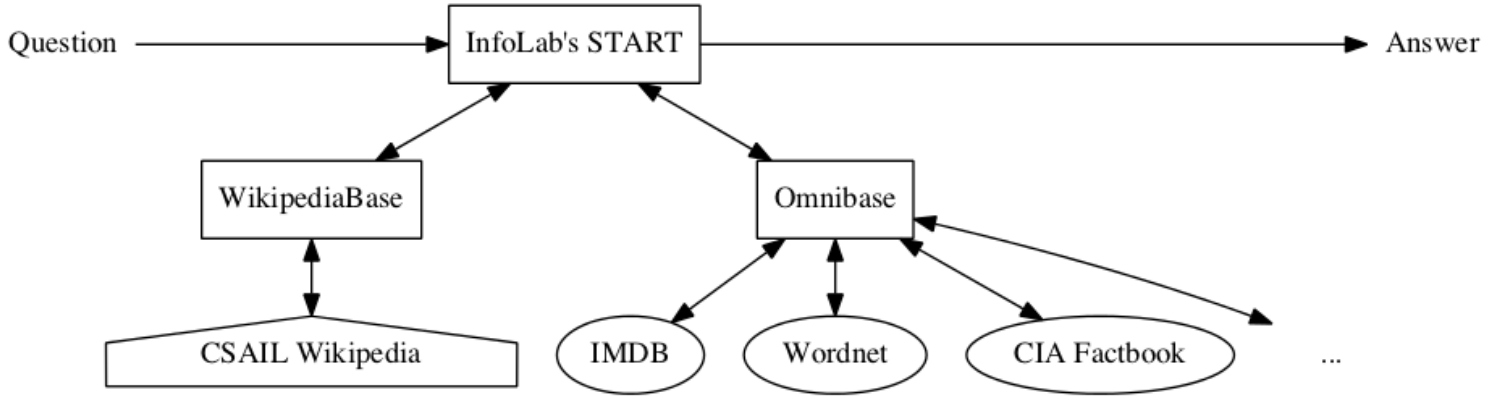
WikipediaBase



Βαρειά χρήση του wikipedia.org



Wikipedia Mirror



Πληροφορίες που διαχειρίζεται κυρίως το WikipediaBase

- Χαρακτηριστικά οντοτήτων

```
(get "wikibase-person" "Barack Obama" (:ID
"BIRTH-DATE"))
=> ((:yyyymmdd 19610804))
```

- Κατηγορίες/κλάσεις οντοτήτων

```
(get-classes "Cardinal (bird)")
=> ("wikibase-term" "wikipedia-paragraphs"
"wikipedia-taxobox")
```

- Συνώνυμα

Infobox

Gerhard Gentzen

Gerhard Gentzen in Prague, 1945.

Born	November 24, 1909 Greifswald, Germany
Died	August 4, 1945 (aged 35) Prague, Czechoslovakia
Nationality	German
Fields	Mathematics
Alma mater	University of Göttingen
Doctoral advisor	Paul Bernays

Infobox

Λαμβάνουμε:

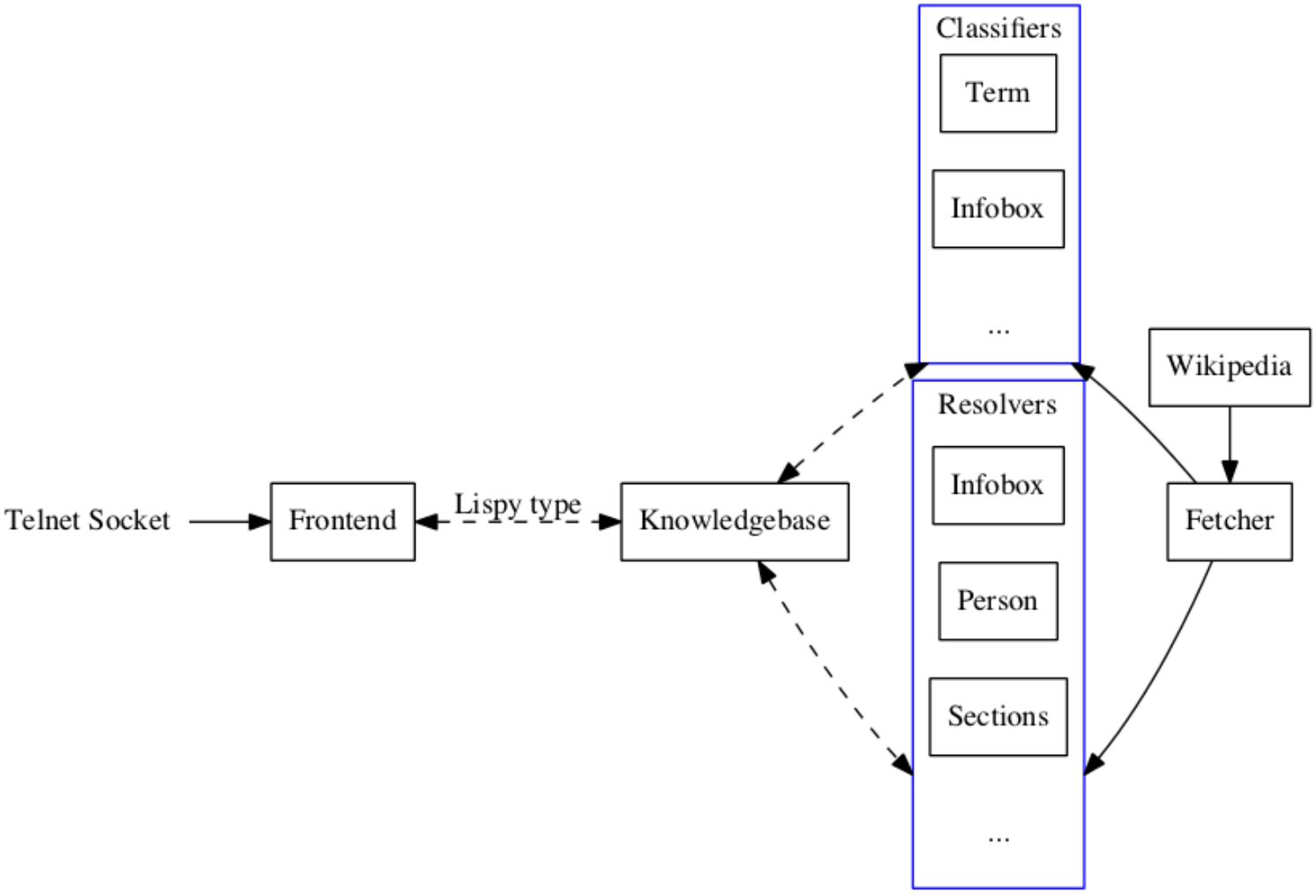
- Χαρκτηριστικά από τον πίνακα του infobox
- Την κατηγορία της οντότητας απο τον τύπο του infobox.

Ιεραρχία των infoboxes

Προέρχεται απο τη σελίδα List of infoboxes. πχ.
`Template:Infobox officeholder` ειναι υπο-κλάση του `Template:Infobox person` και ετσι:

```
> (get-classes "Hillary Rodham Clinton")
("wikibase-term" "wikipedia-paragraphs"
"wikibase-person" "wikipedia-officeholder"
"wikipedia-person")
```

WikipediaBase data pipeline



Resolvers

- Infobox — Χαρακτηριστικά απο το infobox
- Person — Αν το αρθρο αναφέρεται σε άνθρωπο, το γενος του, η ημ.γεννησης κτλ
- Sections — Αυτούσιο κείμενο του άρθρου
- Term — συντεταγμένες, εικόνες, αριθμό, κύρια ονόματα, περίληψη αρθρου, URL καο αριθμός λέξεων
- Error — Ενδειξη οτι δεν βρέθηκε.

Classifiers

- Term — παντα `wikipedia-term`
- Infobox — απο τον τύπο του infobox
- Person — Ευρετικές για το αν το άρθρο αναφέρεται σε πρόσωπο.

Provider/Acquirer model

Διαχωρισμός της λογικής που σχετίζεται με το χειρισμό δεδομένων απο τη λογική που καθορίζει τις πηγές της.

$$\bigcup_{p \in providers} \{o \,:\, o \in provided(p)\}$$

ή dict πρόσβαση

$$\bigcup_{p \in providers} \{(k, v) \,:\, (k, v) \in provided(p)\}$$

Συνώνυμα

Περισσοτερα απο ενα σύμβολα μπορει αν αντιστοιχούν στην ίδια οντότητα.

Μερικά σύμβολα αντιστοιχούν σε αρθρα της wikipedia αλλα οχι σε οντότητες χρήσιμες στο START.

Παραδείγματα συνωνύμων

- Raven (Journal) $\equiv$ Raven
- Russian language/Russian alphabet $\equiv$ Russian alphabet $\equiv$ Russian language
- Obama $\equiv$ Barack Obama
- Beatles $\equiv$ The Beales

Παραδειγματα μη αποδεκτών συνωνύμων

- File:Venn0001.svg
- List of infoboxes
- Alexander_Pushkin#Legacy
- Abraham House $\equiv$ A. House $\not\equiv$ House

Αναγνώριση χρονικών διαστημάτων

```
> I am on the 12th bus,
I will be here from  30th of september 2006 to
18.7.2007
```

Αναγνώριση χρονικών διαστημάτων

```
> I am on the 12th bus,
I will be here from  30th of september 2006 to
18.7.2007
((30, 9, 2006), (18, 7, 2007))
```

Αναγνώριση χρονικών διαστημάτων

```
> I am on the 12th bus,
I will be here from  30th of september 2006 to
18.7.2007
((30, 9, 2006), (18, 7, 2007))
```

- tag:month, tag:fullname

Αναγνώριση χρονικών διαστημάτων

```
> I am on the 12th bus,
I will be here from  30th of september 2006 to
18.7.2007
((30, 9, 2006), (18, 7, 2007))
```

- tag:month, tag:fullname
- tag:day, tag:numeric

Αναγνώριση χρονικών διαστημάτων

```
> I am on the 12th bus,
I will be here from  30th of september 2006 to
18.7.2007
((30, 9, 2006), (18, 7, 2007))
```

- tag:month, tag:fullname
- tag:day, tag:numeric
- tag:year, tag:4digit

Αναγνώριση χρονικών διαστημάτων

```
> I am on the 12th bus,
I will be here from  30th of september 2006 to
18.7.2007
((30, 9, 2006), (18, 7, 2007))
```

- tag:month, tag:fullname
- tag:day, tag:numeric
- tag:year, tag:4digit
- {day,numeric} of {month,fullname} {tag:year}

Αναγνώριση χρονικών διαστημάτων

```
> I am on the 12th bus,
I will be here from  30th of september 2006 to
18.7.2007
((30, 9, 2006), (18, 7, 2007))
```

- tag:month, tag:fullname
- tag:day, tag:numeric
- tag:year, tag:4digit
- {day,numeric} of {month,fullname} {tag:year}

Αναγνώριση χρονικών διαστημάτων

```
> I am on the 12th bus,
I will be here from  30th of september 2006 to
18.7.2007
((30, 9, 2006), (18, 7, 2007))
```

- tag:month, tag:fullname
- tag:day, tag:numeric
- tag:year, tag:4digit
- {day,numeric} of {month,fullname} {tag:year}
- {day,number}.{month,number}.{year}

Αναγνώριση χρονικών διαστημάτων

```
> I am on the 12th bus,
I will be here from  30th of september 2006 to
18.7.2007
((30, 9, 2006), (18, 7, 2007))
```

- tag:month, tag:fullname
- tag:day, tag:numeric
- tag:year, tag:4digit
- {day,numeric} **of** {month,fullname} {tag:year}
- {day,number}.{month,number}.{year}
- **{date} to {date}**

Τι είνει το Wikipedia Mirror

Ενα προγραμμα που παράγει κλώνους της wikipedia που τρεχουν σε ενα τοπικό μηχάνημα.

Διαδικασία

Διαδικασία

- Στήσιμο του server stack

Διαδικασία

- Στήσιμο του server stack
- Εγκατάσταση του Mediawiki

Διαδικασία

- Στήσιμο του server stack
- Εγκατάσταση του Mediawiki
- Κατέβασμα και αποκωδικοποίηση των wikipedia dumps

Διαδικασία

- Στήσιμο του server stack
- Εγκατάσταση του Mediawiki
- Κατέβασμα και αποκωδικοποίηση των wikipedia dumps
- Φόρτωση των dumps στη wikipedia

Διαδικασία

- Στήσιμο του server stack
- Εγκατάσταση του Mediawiki
- Κατέβασμα και αποκωδικοποίηση των wikipedia dumps
- Φόρτωση των dumps στη wikipedia
- Ρύθμιση του mediawiki να μιμείται τη wikipedia

Διαδικασία

- Στήσιμο του server stack
- Εγκατάσταση του Mediawiki
- Κατέβασμα και αποκωδικοποίηση των wikipedia dumps
- Φόρτωση των dumps στη wikipedia
- Ρύθμιση του mediawiki να μιμείται τη wikipedia
- Ρύθμιση της βάσης δεδομένων για επιδοσεις συγκρίσιμες με wikipedia.org

Mediawiki Stack

- Wikipedia configuration, database (MySQL) etc
- Mediawiki
- PHP
- Apache
- Linux

Mediawiki Stack

- Wikipedia configuration, **database (MySQL)** etc
- **Mediawiki**
- **PHP**
- **Apache**
- **Linux**

Bitnami!

Περιεχόμενα βάσης δεδομένων

Μηνιαία snapshots ολόκληρης της βάσης (εκτός απ' τα extensions και τους χρήστες).

`https://dumps.wikimedia.org/enwiki/latest/`

Mwdumper

Μετατροπή απο XML σε MySQL: **mwdumper**

xerces bug

Stack overflow στην αποκωδικοποίηση κάποιων άρθρων απο την XML library (xerces).

xerces bug

Stack overflow στην αποκωδικοποίηση κάποιων άρθρων απο την XML library (xerces).

- Γνωστό bug αλλα δυσκολα αναπαράξιμο.

xerces bug

Stack overflow στην αποκωδικοποίηση κάποιων άρθρων απο την XML library (xerces).

- Γνωστό bug αλλα δυσκολα αναπαράξιμο.
- Μονο του το αρθρο δεν έσπαγε.

xerces bug

Stack overflow στην αποκωδικοποίηση κάποιων άρθρων απο την XML library (xerces).

- Γνωστό bug αλλα δυσκολα αναπαράξιμο.
- Μονο του το αρθρο δεν έσπαγε.
- Καλυμμένο με κενα έσπαγε.

xerces bug

Stack overflow στην αποκωδικοποίηση κάποιων άρθρων απο την XML library (xerces).

- Γνωστό bug αλλα δυσκολα αναπαράξιμο.
- Μονο του το αρθρο δεν έσπαγε.
- Καλυμμένο με κενα έσπαγε.
- Αφαιρεμένο απο το XML αρχείο δούλευε.

xerces bug

Stack overflow στην αποκωδικοποίηση κάποιων άρθρων απο την XML library (xerces).

- Γνωστό bug αλλα δυσκολα αναπαράξιμο.
- Μονο του το αρθρο δεν έσπαγε.
- Καλυμμένο με κενα έσπαγε.
- Αφαιρεμένο απο το XML αρχείο δούλευε.Αυτοματοποιήσαμε αυτη τη διαδικασία.

Επιδόσεις

- *CPU:* Xeon E5-1607 3GHz 4-Core 64 bit
- *Main memory:* 64G
- *HDD:* (spinning disk) 500GB + 2Tb

Επιδόσεις δημιουργίας της βάσης (ανάλογα τη φορά):

- Προεπεξεργασία των dumps 10 λεπτά
- Δημιουργία της βάσης 10 ώρες

Runtime Επιδόσεις

Απαγορευτικές για την ομαλή λειτουργία του START. Για το άρθρο του Barack Obama:

- Χωρίς βελτιστοποιήσεις 10s κατ ευθείαν στη βαση
- Με καποιες βελτιστοποιήσεις 7s κατ ευθείαν στη βαση

Ρυθμίσεις MySQL

Βελτιστοποιήσεις στη βάση που επιχειρήθηκαν:

- `innodb_buffer_pool_size` $\in (16, 5GB)$`fsync`
- `innodb`$_i o_c$*apacitybandwidth*.

Ευχαριστίες

Ευχαριστώ τους καθηγητές μου κ Σγαρμπα, κ Katz και κ Φακωτάκη.

Mandatory Dijkstra quote

> The question of whether Machines Can Think... is about as relevant as the question of whether Submarines Can Swim.
>
> — Edsger W. Dijkstra