

NewSum: Όταν η εξαγωγή περιλήψεων έγινε πολυγλωσσική

Γιώργος Γιαννακόπουλος^{1,2}

¹ ΕΚΕΦΕ “Δημόκριτος”

² SciFY AMKE

2018 – ggianna@iit.demokritos.gr

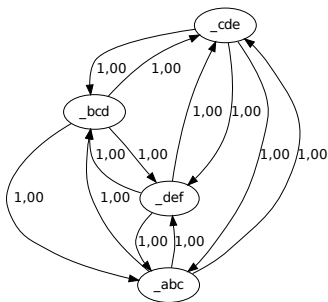
Σκοπός

- Να δούμε ένα παράδειγμα όπου η Τεχνητή Νοημοσύνη υπηρετεί τον πολίτη και επαγγελματία
- Να γνωρίσουμε μία πλευρά της ανάλυσης φυσικής γλώσσας (Natural Language Processing)

Σκοπός

- Να δούμε ένα παράδειγμα όπου η Τεχνητή Νοημοσύνη υπηρετεί τον πολίτη και επαγγελματία
- Να γνωρίσουμε μία πλευρά της ανάλυσης φυσικής γλώσσας (Natural Language Processing)
- Να δούμε πώς μία ιδέα ταξίδεψε από ένα λεωφορείο στις άκρες του κόσμου

Ο γράφος n-γραμμάτων (N-gram graph)



- Περιγράφει γειτονιές (γειτνιάσεις)
- Οι **Ακμές** είναι σημαντικές
- Οι κόμβοι είναι **μοναδικοί**

Γιατί;

- Πολλά εργαλεία ανάλυσης
- Λίγες στοχευμένες λύσεις
- Υπερβολικά πολλά και ξεχωριστά εργαλεία

The problems

- Συλλογή δεδομένων: αναζήτηση πηγών
- Περίληψη: Εξαγωγή της ουσίας, χωρίς επανάληψη
- Εντοπισμός τάσεων: προσδιορισμός του τι είναι σημαντικό
- Αυτόματη ταξινόμηση: σε ποια κατηγορία ανήκει κάθε νέο
- Δημιουργία “slugline” (υποτίτλου): ποιοι είναι οι εμπλεκόμενοι σε ένα γεγονός

Άλλες απαιτήσεις

- Αποτελεσματικό σε πολλά δεδομένα
- Εύκολα προσαρμόσιμο σε άλλες γλώσσες
- Εύκολα προσαρμόσιμο σε νέους τομείς

Η δική μας οπτική

Να φτιάξουμε κάτι **χρήσιμο, αποδοτικό και επαναχρησιμοποιούμενο.**

Η εργαλειοθήκη NewSum

Former NFL star Hernandez charged with double murder

Former New England Patriots football star Aaron Hernandez has been charged with murder in a 2012 Boston double homicide, Suffolk County District Attorney Daniel Conroy said at a press conference on Thursday.

Source:  Also found in [1](#).

Former National Football League star Aaron Hernandez was indicted on Thursday for the 2012 murder of two men in a Boston drive-by shooting, according to the Boston Globe, which cited two officials briefed on the case.

Source:  Also found in [1](#).

The ex-New England Patriot is already awaiting trial for an unrelated fatal shooting.

Source: [1](#)

All Sources:      [1](#).

Related tweets (1 found)

"Fact 5: Drug dealer" RT @Deadepin Aaron Hernandez may have tattooed himself with clues to 2012 double-murder: <http://t.co/LeWTGEFkFW>

21-5-2014
20:31:15

[View Message](#)

Συλλογή δεδομένων

- Ειδήσεις
 - RSS feeds
 - Προγραμματιστικές διεπαφές (APIs)
 - Ειδικές πηγές (CSV, JSON, scraping, κλπ.)
- Κοινωνικά δίκτυα
 - Twitter
 - Facebook

Ποιον βοηθά;

- Δημοσιογράφο
- Ερευνητή
- Πολίτη

Πώς βοηθά τον άνθρωπο;

- Διαβάζει ειδήσεις για εμάς
- Τις ξεχωρίζει σε θέματα
- Τις συνοψίζει, κρατώντας όλες τις γνώμες, αλλά με ελάχιστη επανάληψη

Ποιον βοηθά;

- Δημοσιογράφο
- Ερευνητή
- Πολίτη

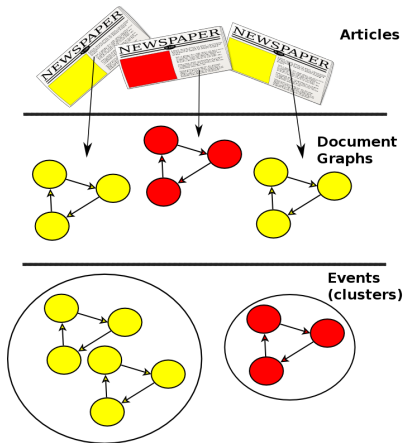
Πώς βοηθά τον άνθρωπο;

- Διαβάζει ειδήσεις για εμάς
- Τις ξεχωρίζει σε θέματα
- Τις συνοψίζει, κρατώντας όλες τις γνώμες, αλλά με ελάχιστη επανάληψη
- Μας επιτρέπει να **επαληθεύσουμε**

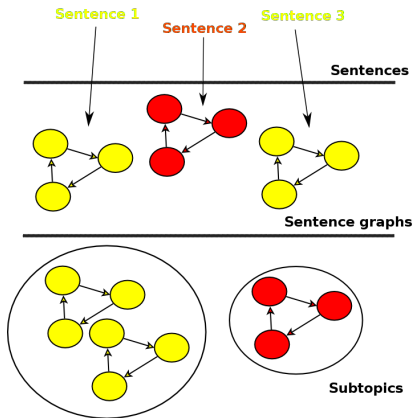
Εξαγωγή περίληψης

- Ανάλυση 2 βημάτων
 - Ομαδοποίηση εγγράφων σε θέματα
 - Ομαδοποίηση προτάσεων σε υπο-θέματα
- Σύνδεση θεμάτων από διαφορετικές πηγές (π.χ. RSS με tweets)

Εντοπισμός γεγονότων



Αναγνώριση υπο-θεμάτων ενός γεγονότος



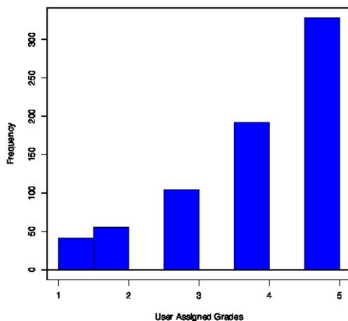
Τελικά είναι χρήσιμο – και σε διαφορετικές γλώσσες;

Ερώτηση Η χρήση του συστήματος μου επέτρεψε να έχω συνολικότερη και πιο σωστή πληροφόρηση από πριν

Απαντήσεις 1 – Διαφωνώ πλήρως έως 5 – Συμφωνώ πλήρως

Τι μάθαμε

- 267 αξιολογήσεις για Αγγλικά, 453 για Ελληνικά
- Αγγλικά: 3.73 (st. dev. 1.34)
- Ελληνικά: 4.14 (st. dev. 1.07)













Προϊόν

- Προγραμματιστική διεπαφή (API)
- Διάφορες εφαρμογές
 - NewSum στο κινητό (δωρεάν και διαθέσιμο)
 - NewSum στο διαδίκτυο (www.newsumontheweb.org)
 - Εφαρμογές σε
 - ανάλυση σχολίων
 - ανάλυση διαβουλεύσεων
 - Γαλλικά, Γερμανικά, Αραβικά, Εβραϊκά, Τούρκικα, ...

Τι κάνουμε... Μέρος 2ο

3 clusters (1287 total)





Doubts arise over whether Trump court nominee's accuser will testify

 4  13  3  GENERAL  2  2  2  3  9/20/2018, 6:24:05 AM  9/19/2018, 12:12:55 AM

Trump says doesn't mind if Kavanaugh vote delayed to allow accuser to be heard

 4  10  3  POLITICS  2  1  4  2  9/20/2018, 6:24:05 AM  9/17/2018, 11:53:04 PM

Gas explosions spark fires in dozens of homes in US

 1  11  3  WORLD  0  0  1  0  9/20/2018, 6:24:05 AM  9/14/2018, 4:28:00 AM

Για του λόγου το αληθές...

Τι κάνουμε... Μέρος 3ο

Schemes
HTTP ▾

Categories ▾

- GET /categories/getCategories Get all Categories
- GET /categories/getCategories/{lang} Get Categories by language

Labeling ▾

- POST /labeling/label Get Label with language and text that contains the document to be classified
- POST /labeling/addNewsMLInstance Add new ML document
- POST /labeling/addInstance Add new document
- GET /labeling/getSizes/{lang} Get the categories instances
- GET /labeling/getTrainingInfo/{lang} Get Training info

Languages ▾

- GET /languages/getLanguages Gets all the languages

Τι κάνουμε... Τελευταίο

OPEN COMMENT ANALYZER

WHAT IT IS

DEMO ANALYSIS

TERM COMBINATIONS

CONTACT

DEMO ANALYSIS

2 OPEN ENDED QUESTIONS. MORE THAN 15000 RESPONSES IN OVER 25 LANGUAGES ANALYZED.



QUESTION 1: WHAT IS GOOD



CLICK ON ONE OF THE WORDS IN THE WORD CLOUD TO SEE WITH WHICH WORDS IT CO-OCCURS.



DRAW AND DROP WORDS FROM THE WORD CLOUD TO THE DUSTBIN TO IGNORE THEM. CLICK THE DUSTBIN TO RESET.

DRAW AND DROP WORDS TO THE GROUP ICON TO FORM A COMBINED CONCEPT. CLICK THE GROUP ICON TO VIEW RELATED COMMENTS.

Αξιοποίηση

- Συζητήσεις για δημιουργία τεχνοβλαστού (spin-off)
- Εμπορική αξιοποίηση
- Χρηματοδότηση από επενδυτικά ταμεία ρίσκου (VC/funds)

Εργαλειοθήκη NewSum

Πρόκληση: Θα δουλεύει καλά στα Κινέζικα;
Ευχαριστώ πολύ!

NewSum: Όταν η εξαγωγή περιλήψεων έγινε
πολυγλωσσική

Γιώργος Γιαννακόπουλος^{1,2}

¹ ΕΚΕΦΕ “Δημόκριτος”

² SciFY AMKE

2018 – ggianna@iit.demokritos.gr