NewSum: Όταν η εξαγωγή περιλήψεων έγινε πολυγλωσσική

Γιώργος Γιαννακόπουλος[1,2]

[1] ΕΚΕΦΕ "Δημόκριτος"
[2] SciFY ΑΜΚΕ

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)

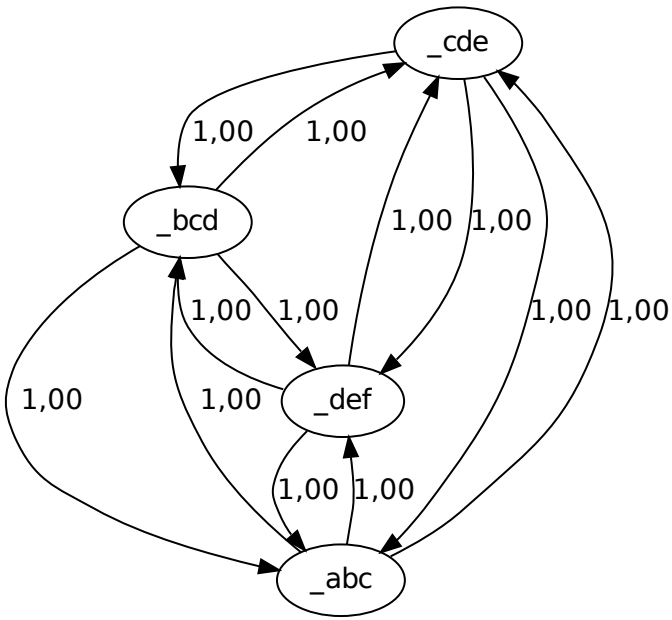
Σκοπός

- 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

Μια φορά κι έναν καιρό...

2001

Ο γράφος ν-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

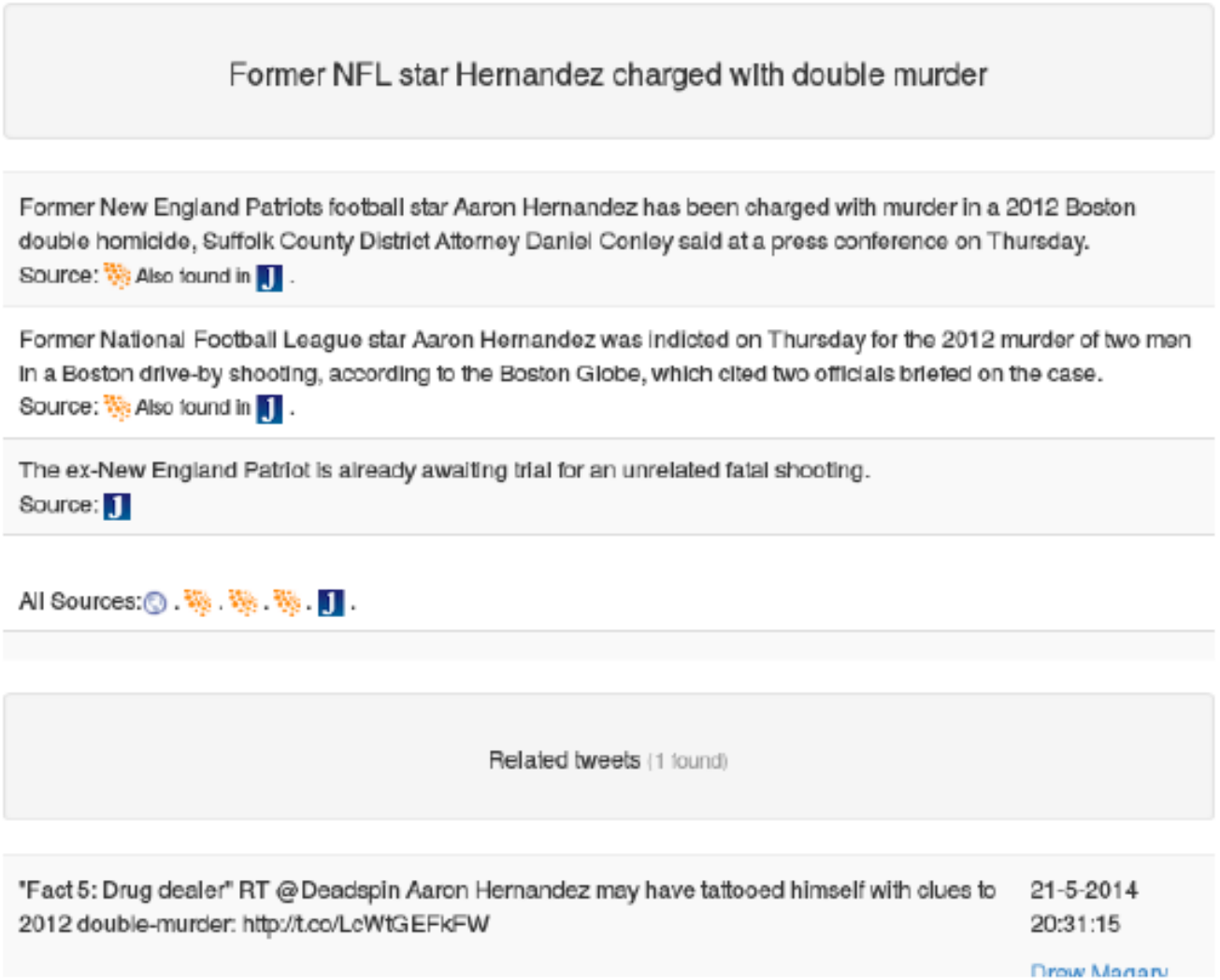
Άλλες απαιτήσεις

- 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



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

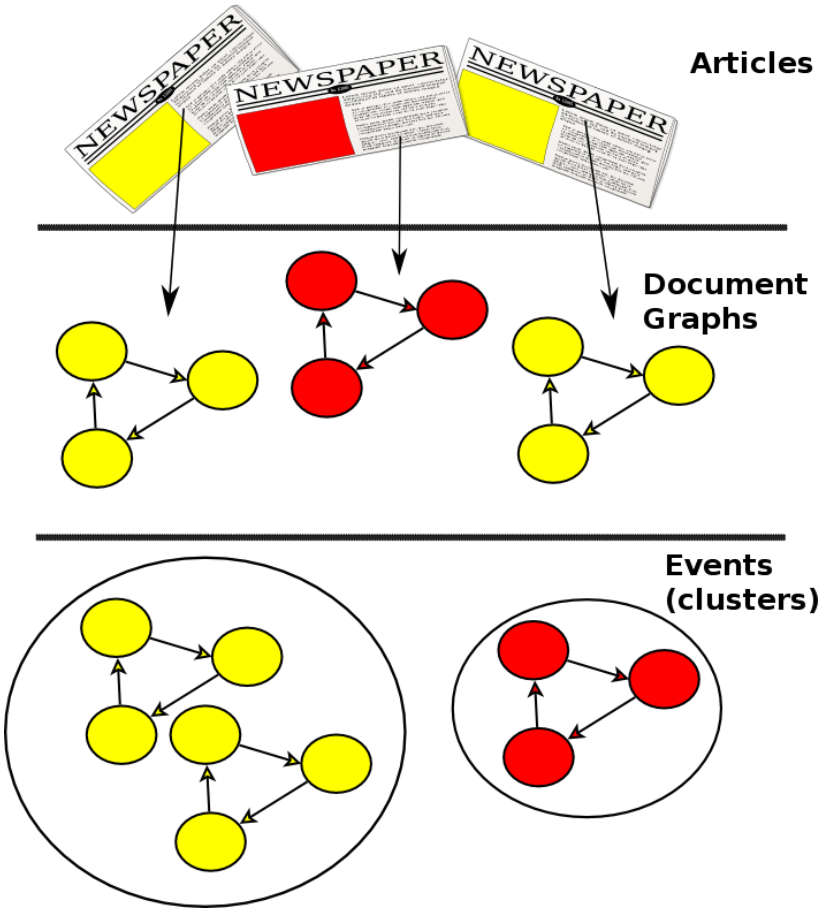
Πώς βοηθά τον άνθρωπο;

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

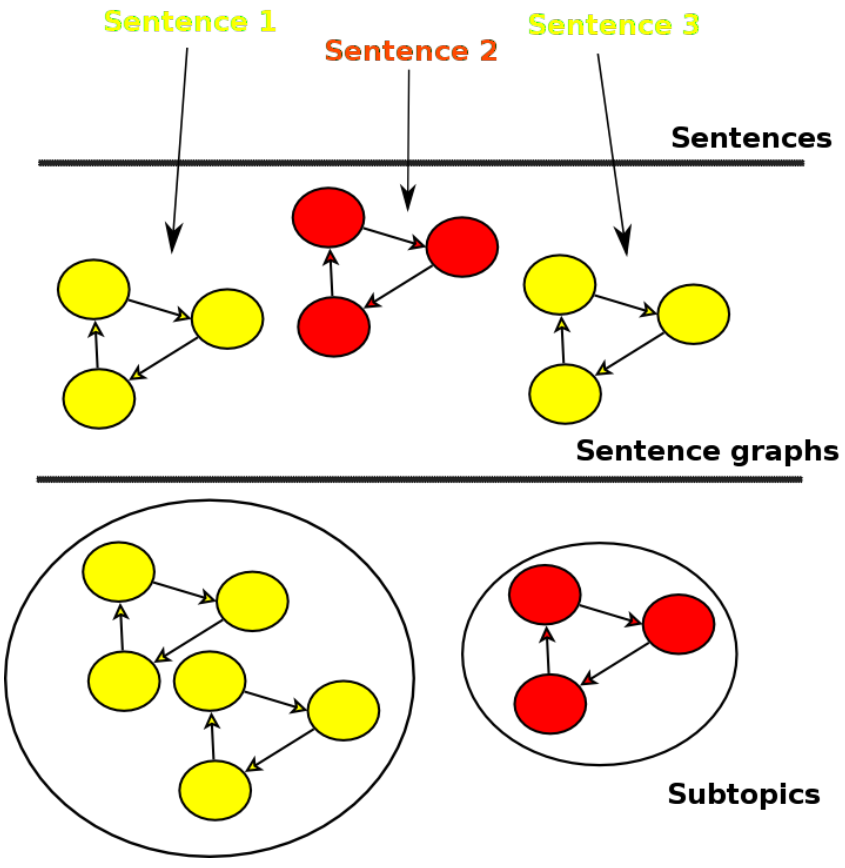
Εξαγωγή περίληψης

- 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

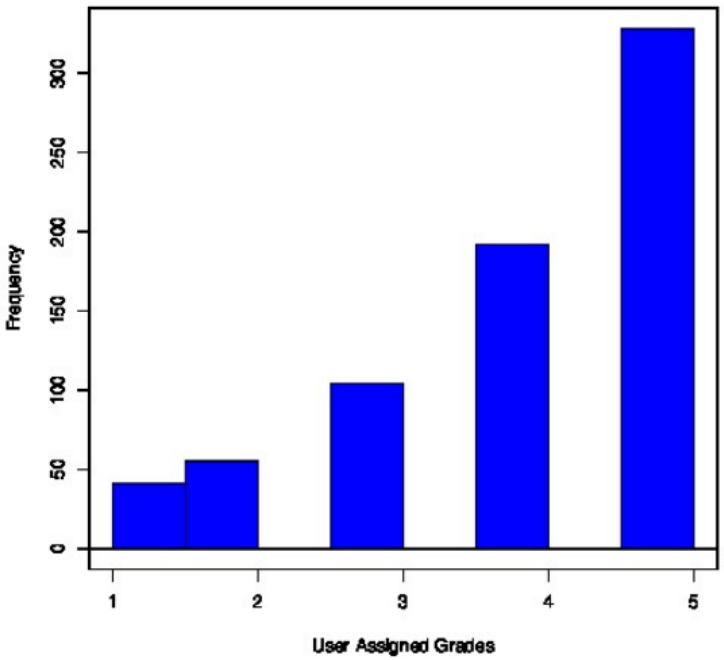


Τελικά είναι χρήσιμο – και σε διαφορετικές γλώσσες;

Ερώτηση	Η χρήση του συστήματος μου επέτρεψε να έχω συνολικότερη και πιο σωστή πληροφόρηση από πριν
Απαντήσεις	*1 – Διαφωνώ πλήρως* εώς *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)



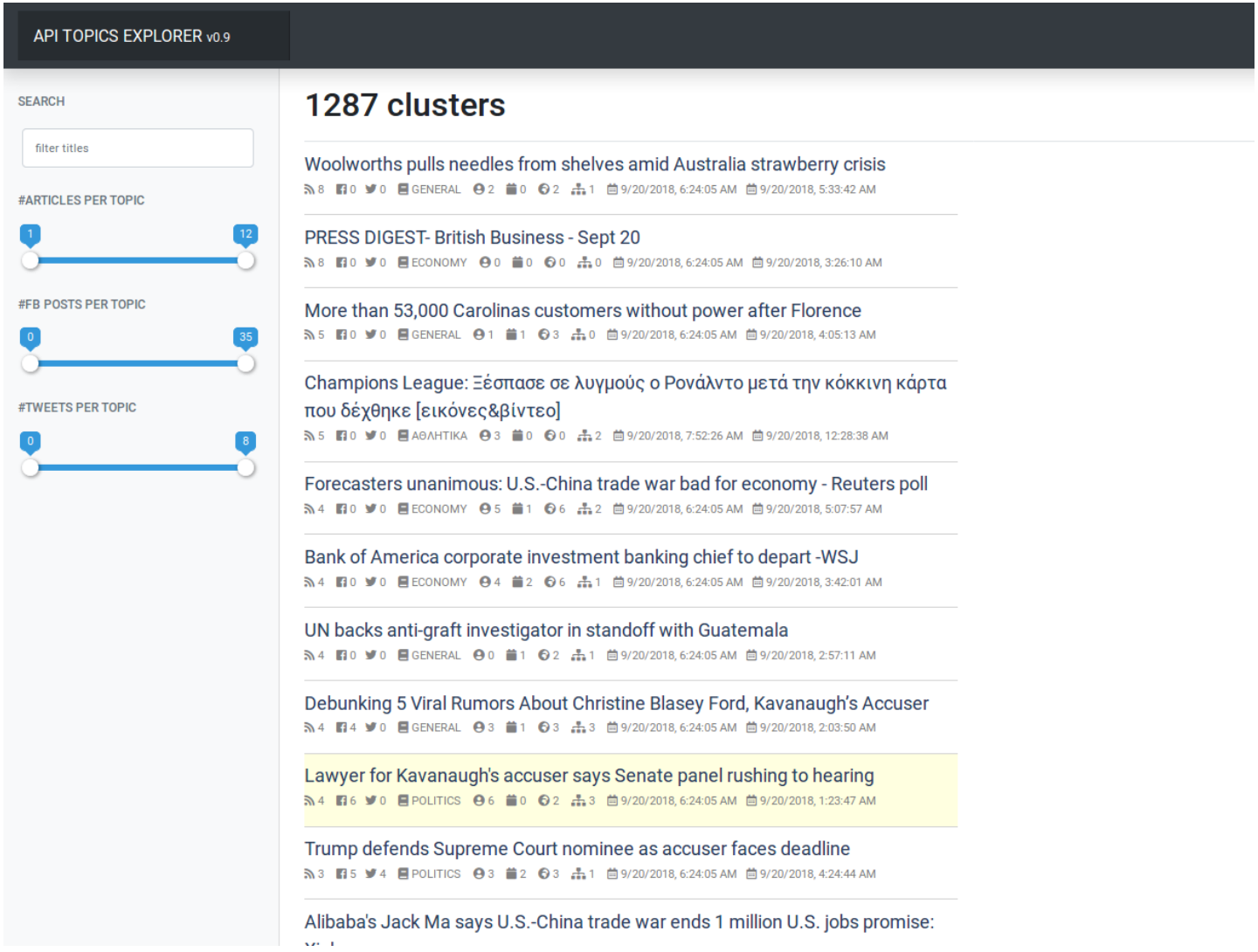
Το αποτέλεσμα

- Βράβευση στη Αθήνα
- Βράβευση ως “Εφαρμογή πιο άξια για επένδυση (fundable) ” στη Νίκαια της Γαλλίας
- Αξιοποίηση σε ερευνητικά έργα (NOMAD, StoryBot, ...)
- Χιλιάδες χρήστες της δωρεάν εφαρμογής

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 1ο



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

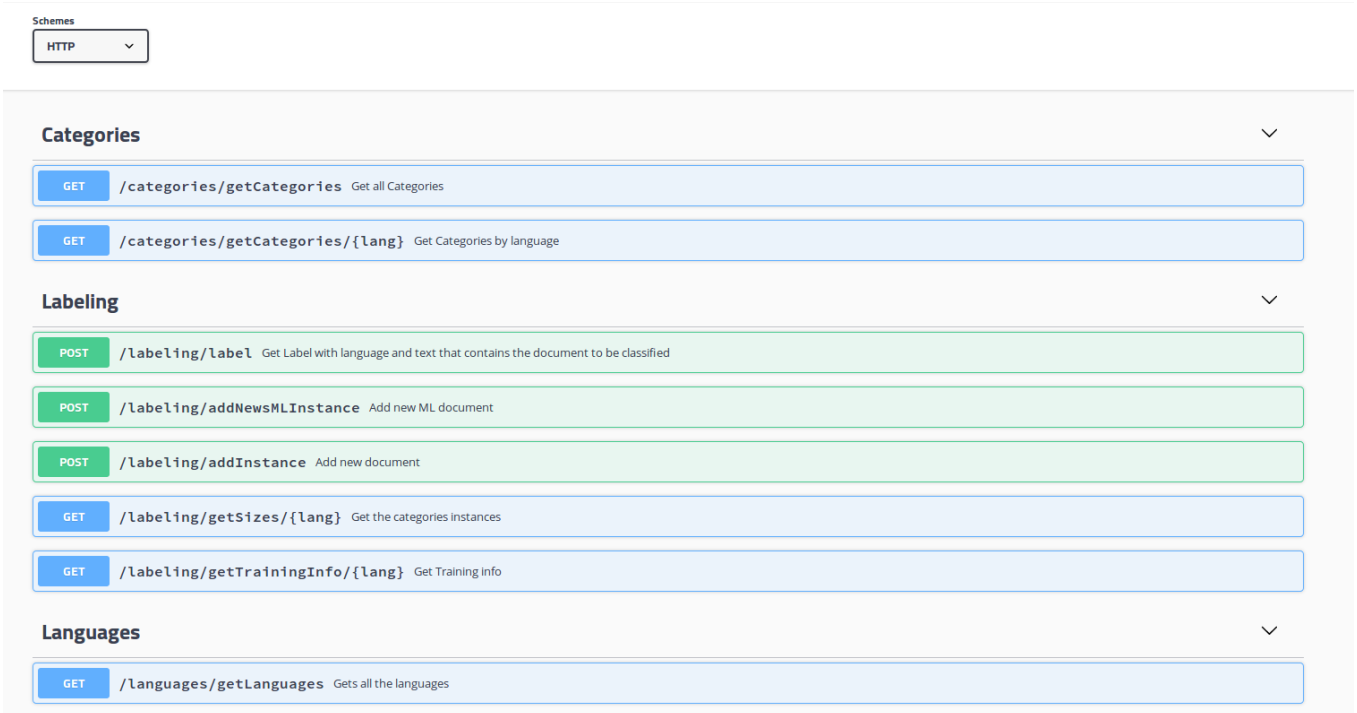
Trump says doesn't mind if Kavanaugh vote delayed to allow accuser to be heard

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

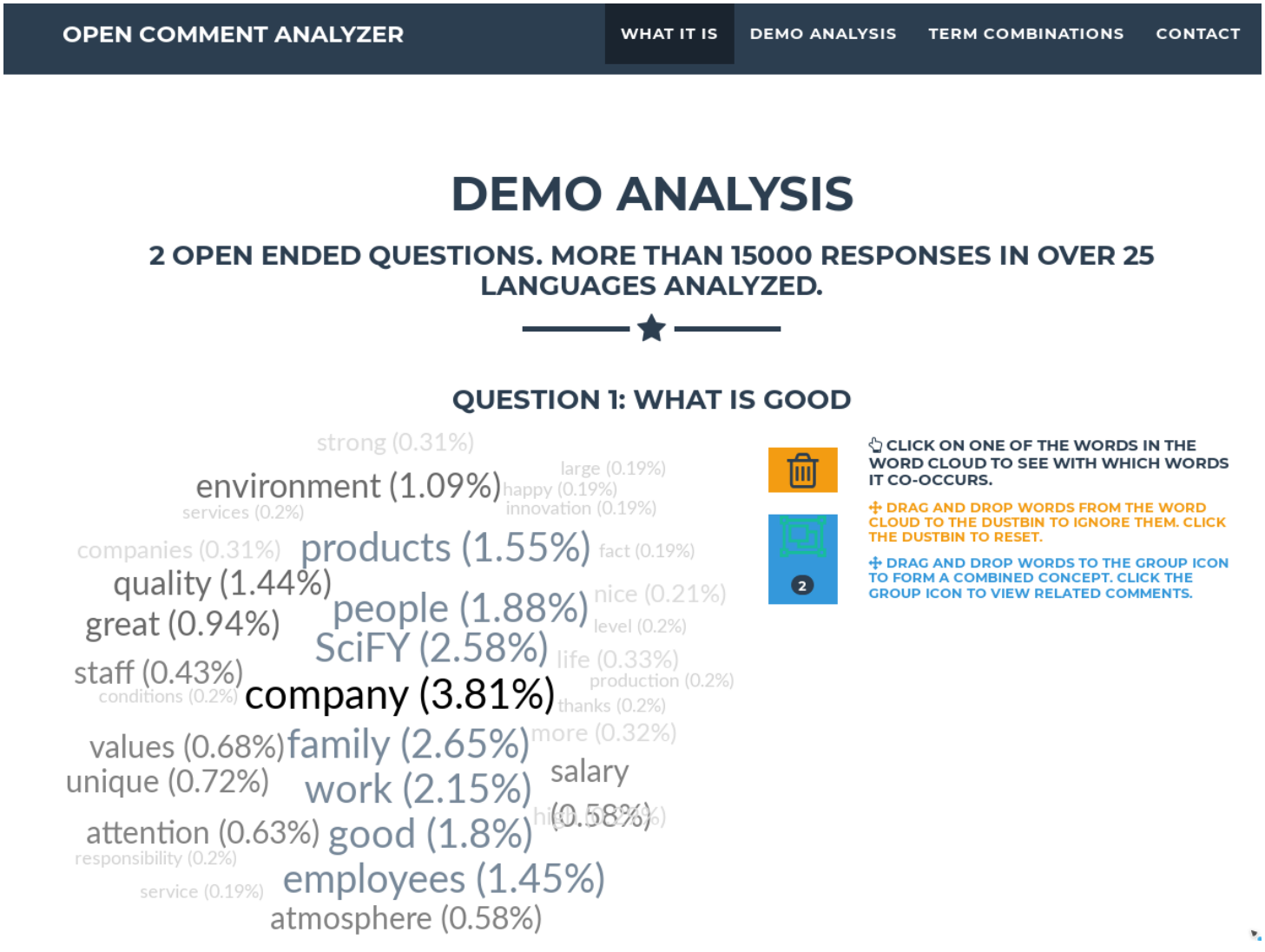
Gas explosions spark fires in dozens of homes in US

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

Τι κάνουμε... Μέρος 3ο



Τι κάνουμε... Τελευταίο

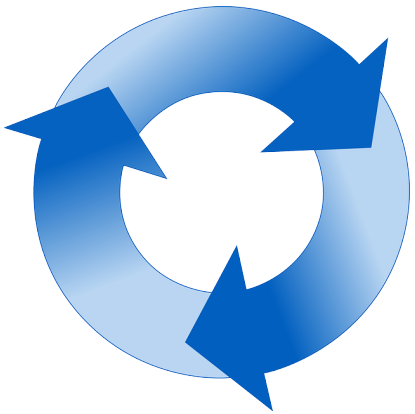


Αξιοποίηση

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

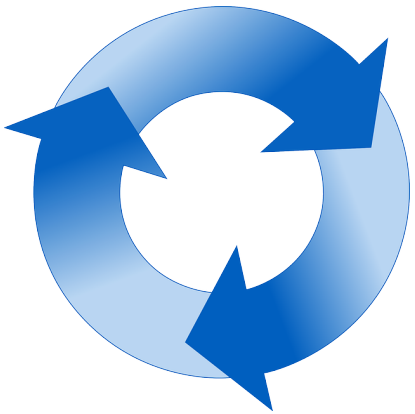
Τι να θυμόμαστε

- NewSum: εξαγωγή περιλήψεων...



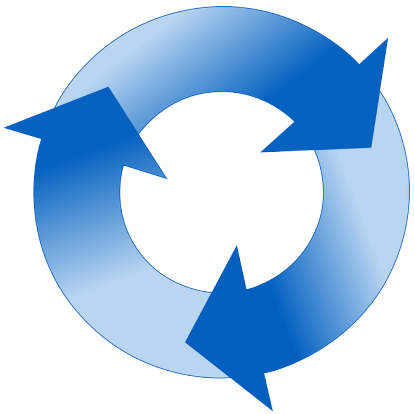
Τι να θυμόμαστε

- NewSum: εξαγωγή περιλήψεων...
- ανεξάρτητη γλώσσας...



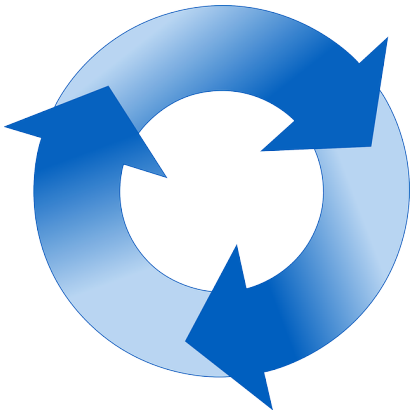
Τι να θυμόμαστε

- NewSum: εξαγωγή περιλήψεων...
- ανεξάρτητη γλώσσας...
- ...και όχι μόνο για ειδήσεις!



Τι να θυμόμαστε

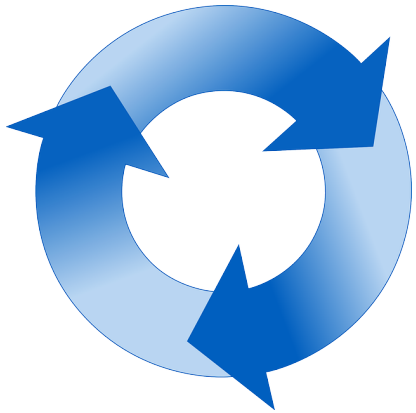
- NewSum: εξαγωγή περιλήψεων...
- ανεξάρτητη γλώσσας...
- ...και όχι μόνο για ειδήσεις!
- Μία επιστημονική ιδέα μπορεί να ταξιδέψει



Τι να θυμόμαστε

- NewSum: εξαγωγή περιλήψεων...
- ανεξάρτητη γλώσσας...
- ...και όχι μόνο για ειδήσεις!
- Μία επιστημονική ιδέα μπορεί να ταξιδέψει

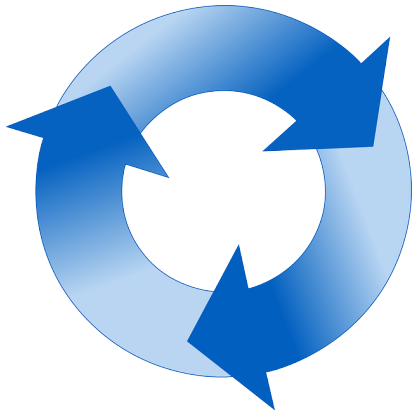
  αν υποστηρίζεται από αγάπη, σκληρή δουλειά, επιμονή



Τι να θυμόμαστε

- NewSum: εξαγωγή περιλήψεων...
- ανεξάρτητη γλώσσας...
- ...και όχι μόνο για ειδήσεις!
- Μία επιστημονική ιδέα μπορεί να ταξιδέψει

  αν υποστηρίζεται από αγάπη, σκληρή δουλειά, επιμονή
- Μια αναγνωρισμένη ιδέα, μπορεί να γίνει επιχείριση

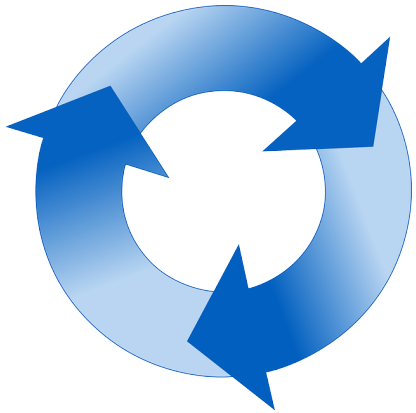


Τι να θυμόμαστε

- NewSum: εξαγωγή περιλήψεων...
- ανεξάρτητη γλώσσας...
- ...και όχι μόνο για ειδήσεις!
- Μία επιστημονική ιδέα μπορεί να ταξιδέψει

  αν υποστηρίζεται από αγάπη, σκληρή δουλειά, επιμονή
- Μια αναγνωρισμένη ιδέα, μπορεί να γίνει επιχείριση

  αν υποστηρίζεται από συνεργασία, επιμονή και γερό στομάχι

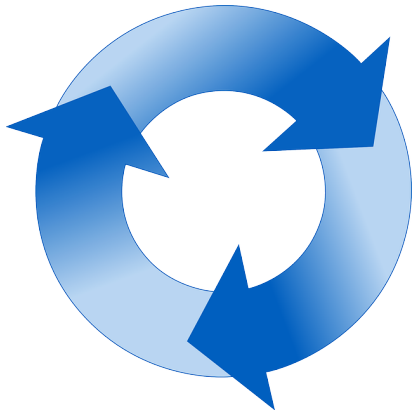


Τι να θυμόμαστε

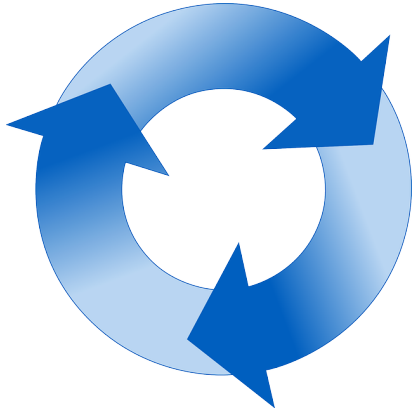
- NewSum: εξαγωγή περιλήψεων...
- ανεξάρτητη γλώσσας...
- ...και όχι μόνο για ειδήσεις!
- Μία επιστημονική ιδέα μπορεί να ταξιδέψει

  αν υποστηρίζεται από αγάπη, σκληρή δουλειά, επιμονή
- Μια αναγνωρισμένη ιδέα, μπορεί να γίνει επιχείριση

  αν υποστηρίζεται από συνεργασία, επιμονή και γερό στομάχι
- **8 χρόνια** από την ιδέα στην 1η αναγνώριση



Τι να θυμόμαστε



- NewSum: εξαγωγή περιλήψεων...
- ανεξάρτητη γλώσσας...
- ...και όχι μόνο για ειδήσεις!
- Μία επιστημονική ιδέα μπορεί να ταξιδέψει

  αν υποστηρίζεται από αγάπη, σκληρή δουλειά, επιμονή
- Μια αναγνωρισμένη ιδέα, μπορεί να γίνει επιχείριση

  αν υποστηρίζεται από συνεργασία, επιμονή και γερό στομάχι
- **8 χρόνια** από την ιδέα στην 1η αναγνώριση
- **10 χρόνια** από την 1η αναγνώριση

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 ΑΜΚΕ

2018 – ggianna@iit.demokritos.gr