Σύστημα Επεξεργασίας, Ανάλυσης και Ομαδοποίησης Εγγράφων Ειδήσεων του Διαδικτύου

Αφροδίτη Αλεβιζοπούλου

Τμήμα Μηχανικών Ηλεκτρονικών Υπολογιστών & Πληροφορικής
Πανεπιστήμιο Πατρών

alevizopou@ceid.upatras.gr

16 Νοεμβρίου 2017

Περιεχόμενα

Περιεχόμενα

Συστήματα Συστάσεων
Ανάγκη εξατομίκευσης

- Ραγδαία εξέλιξη Παγκόσμιου Ιστού/Μεγάλος όγκος πληροφοριών που κατακλύζει το διαδίκτυο (προσβάσιμο σε εκατομμύρια χρήστες, μειωμένο κόστος διανομής και πρόσβασης στις ειδήσεις, παγκόσμια αποστολή και κατανάλωση πληροφορίας, μικρός χρόνος για τη δημοσίευση ειδήσεων).
- Δυσκολία να ξεχωρίσουμε πληροφορίες σχετικές με τα ενδιαφέροντά μας.
- Ανάγκη για φιλτράρισμα (διαφορετικά ενδιαφέροντα χρηστών) και εξατομικευμένες προτάσεις.
  - Συστήματα Συστάσεων: Λαμβάνοντας ως είσοδο τις προτιμήσεις των χρηστών, υπολογίζουν το εκτιμώμενο ενδιαφέρον ως προς ένα αντικείμενο.

Συστήματα Συστάσεων
Ανάγκη εξατομίκευσης

- Ραγδαία εξέλιξη Παγκόσμιου Ιστού/Μεγάλος όγκος πληροφοριών που κατακλύζει το διαδίκτυο (προσβάσιμο σε εκατομμύρια χρήστες, μειωμένο κόστος διανομής και πρόσβασης στις ειδήσεις, παγκόσμια αποστολή και κατανάλωση πληροφορίας, μικρός χρόνος για τη δημοσίευση ειδήσεων).
- Δυσκολία να ξεχωρίσουμε πληροφορίες σχετικές με τα ενδιαφέροντά μας.
- Ανάγκη για φιλτράρισμα (διαφορετικά ενδιαφέροντα χρηστών) και εξατομικευμένες προτάσεις.
    - Συστήματα Συστάσεων: Λαμβάνοντας ως είσοδο τις προτιμήσεις των χρηστών, υπολογίζουν το εκτιμώμενο ενδιαφέρον ως προς ένα αντικείμενο.

Συστήματα Συστάσεων
Ανάγκη εξατομίκευσης

- Ραγδαία εξέλιξη Παγκόσμιου Ιστού/Μεγάλος όγκος πληροφοριών που κατακλύζει το διαδίκτυο (προσβάσιμο σε εκατομμύρια χρήστες, μειωμένο κόστος διανομής και πρόσβασης στις ειδήσεις, παγκόσμια αποστολή και κατανάλωση πληροφορίας, μικρός χρόνος για τη δημοσίευση ειδήσεων).
- Δυσκολία να ξεχωρίσουμε πληροφορίες σχετικές με τα ενδιαφέροντά μας.
- Ανάγκη για φιλτράρισμα (διαφορετικά ενδιαφέροντα χρηστών) και εξατομικευμένες προτάσεις.
    - Συστήματα Συστάσεων: Λαμβάνοντας ως είσοδο τις προτιμήσεις των χρηστών, υπολογίζουν το εκτιμώμενο ενδιαφέρον ως προς ένα αντικείμενο.

Συστήματα Συστάσεων
Ανάγκη εξατομίκευσης

- Ραγδαία εξέλιξη Παγκόσμιου Ιστού/Μεγάλος όγκος πληροφοριών που κατακλύζει το διαδίκτυο (προσβάσιμο σε εκατομμύρια χρήστες, μειωμένο κόστος διανομής και πρόσβασης στις ειδήσεις, παγκόσμια αποστολή και κατανάλωση πληροφορίας, μικρός χρόνος για τη δημοσίευση ειδήσεων).
- Δυσκολία να ξεχωρίσουμε πληροφορίες σχετικές με τα ενδιαφέροντά μας.
- Ανάγκη για φιλτράρισμα (διαφορετικά ενδιαφέροντα χρηστών) και εξατομικευμένες προτάσεις.
  - Συστήματα Συστάσεων: Λαμβάνοντας ως είσοδο τις προτιμήσεις των χρηστών, υπολογίζουν το εκτιμώμενο ενδιαφέρον ως προς ένα αντικείμενο.

Συστήματα Συστάσεων
Στόχος Διπλωματικής Εργασίας

- Σχεδιασμός και υλοποίηση συστήματος επεξεργασίας, ανάλυσης και ομαδοποίησης εγγράφων ειδήσεων του διαδικτύου με στόχο τις εξατομικευμένες συστάσεις άρθρων ειδήσεων σε χρήστες.
  - Επεξεργασία, ανάλυση και ομαδοποίηση άρθρων ειδήσεων του διαδικτύου και αποθήκευση πληροφορίας στη βάση δεδομένων.
  - Περιήγηση χρήστη μεταξύ των άρθρων του συστήματος.
  - Δημιουργία προφίλ/ιστορικού χρήστη.
  - Δημιουργία συστάσεων ειδησεογραφικών άρθρων που εκτιμάται ότι σχετίζονται με τα ενδιαφέροντά του.
  - Αξιολόγηση προτάσεων συστήματος από το χρήστη.

Συστήματα Συστάσεων
Στόχος Διπλωματικής Εργασίας

- Σχεδιασμός και υλοποίηση συστήματος επεξεργασίας, ανάλυσης και ομαδοποίησης εγγράφων ειδήσεων του διαδικτύου με στόχο τις εξατομικευμένες συστάσεις άρθρων ειδήσεων σε χρήστες.
  - Επεξεργασία, ανάλυση και ομαδοποίηση άρθρων ειδήσεων του διαδικτύου και αποθήκευση πληροφορίας στη βάση δεδομένων.
  - Περιήγηση χρήστη μεταξύ των άρθρων του συστήματος.
  - Δημιουργία προφίλ/ιστορικού χρήστη.
  - Δημιουργία συστάσεων ειδησεογραφικών άρθρων που εκτιμάται ότι σχετίζονται με τα ενδιαφέροντά του.
  - Αξιολόγηση προτάσεων συστήματος από το χρήστη.

Συστήματα Συστάσεων
Στόχος Διπλωματικής Εργασίας

- Σχεδιασμός και υλοποίηση συστήματος επεξεργασίας, ανάλυσης και ομαδοποίησης εγγράφων ειδήσεων του διαδικτύου με στόχο τις εξατομικευμένες συστάσεις άρθρων ειδήσεων σε χρήστες.
  - Επεξεργασία, ανάλυση και ομαδοποίηση άρθρων ειδήσεων του διαδικτύου και αποθήκευση πληροφορίας στη βάση δεδομένων.
  - Περιήγηση χρήστη μεταξύ των άρθρων του συστήματος.
  - Δημιουργία προφίλ/ιστορικού χρήστη.
  - Δημιουργία συστάσεων ειδησεογραφικών άρθρων που εκτιμάται ότι σχετίζονται με τα ενδιαφέροντά του.
  - Αξιολόγηση προτάσεων συστήματος από το χρήστη.

Συστήματα Συστάσεων
Στόχος Διπλωματικής Εργασίας

- Σχεδιασμός και υλοποίηση συστήματος επεξεργασίας, ανάλυσης και ομαδοποίησης εγγράφων ειδήσεων του διαδικτύου με στόχο τις εξατομικευμένες συστάσεις άρθρων ειδήσεων σε χρήστες.
  - Επεξεργασία, ανάλυση και ομαδοποίηση άρθρων ειδήσεων του διαδικτύου και αποθήκευση πληροφορίας στη βάση δεδομένων.
  - Περιήγηση χρήστη μεταξύ των άρθρων του συστήματος.
  - Δημιουργία προφίλ/ιστορικού χρήστη.
  - Δημιουργία συστάσεων ειδησεογραφικών άρθρων που εκτιμάται ότι σχετίζονται με τα ενδιαφέροντά του.
  - Αξιολόγηση προτάσεων συστήματος από το χρήστη.

Συστήματα Συστάσεων
Στόχος Διπλωματικής Εργασίας

- Σχεδιασμός και υλοποίηση συστήματος επεξεργασίας, ανάλυσης και ομαδοποίησης εγγράφων ειδήσεων του διαδικτύου με στόχο τις εξατομικευμένες συστάσεις άρθρων ειδήσεων σε χρήστες.
  - Επεξεργασία, ανάλυση και ομαδοποίηση άρθρων ειδήσεων του διαδικτύου και αποθήκευση πληροφορίας στη βάση δεδομένων.
  - Περιήγηση χρήστη μεταξύ των άρθρων του συστήματος.
  - Δημιουργία προφίλ/ιστορικού χρήστη.
  - Δημιουργία συστάσεων ειδησεογραφικών άρθρων που εκτιμάται ότι σχετίζονται με τα ενδιαφέροντά του.
  - Αξιολόγηση προτάσεων συστήματος από το χρήστη.

Περιεχόμενα

Επεξεργασία Φυσικής Γλώσσας

- Διεπιστημονικός κλάδος της επιστήμης της Πληροφορικής, της Τεχνητής Νοημοσύνης και της Υπολογιστικής Γλωσσολογίας.
- Αλληλεπίδραση μεταξύ των υπολογιστών και της ανθρώπινης (φυσικής) γλώσσας.
- Προσπάθεια να καταστούν ικανοί οι υπολογιστές να κατανοήσουν/εξάγουν νοήματα από γλωσσικά δεδομένα.

Επεξεργασία Φυσικής Γλώσσας

Πεδία Έρευνας Επεξεργασίας Φυσικής Γλώσσας

- Ανάλυση λόγου
- Αυτόματη αναγνώριση ομιλίας
- Αυτόματη περίληψη
- Εξόρυξη πληροφοριών
- Μηχανική μετάφραση
- Οπτική αναγνώριση χαρακτήρων
- Σύνθεση ομιλίας
- Συντακτική ανάλυση

Επεξεργασία Φυσικής Γλώσσας

Ανάλυση Φυσικής Γλώσσας με χρήση Θεματικών Μοντέλων

- **Θεματικό Μοντέλο (topic model):** Ένας τύπος στατιστικού μοντέλου για την ανακάλυψη θεμάτων που υπάρχουν σε μία συλλογή κειμένων.
- Βασική ιδέα: Τα κείμενα αντιπροσωπεύονται από τυχαίες προσμείξεις κρυφών θεμάτων, όπου κάθε θέμα χαρακτηρίζεται από μία κατανομή ως προς τις λέξεις.

Επεξεργασία Φυσικής Γλώσσας

Ανάλυση Φυσικής Γλώσσας με χρήση Θεματικών Μοντέλων

Επεξεργασία Φυσικής Γλώσσας

Ανάλυση Φυσικής Γλώσσας με χρήση Θεματικών Μοντέλων

Στόχος:

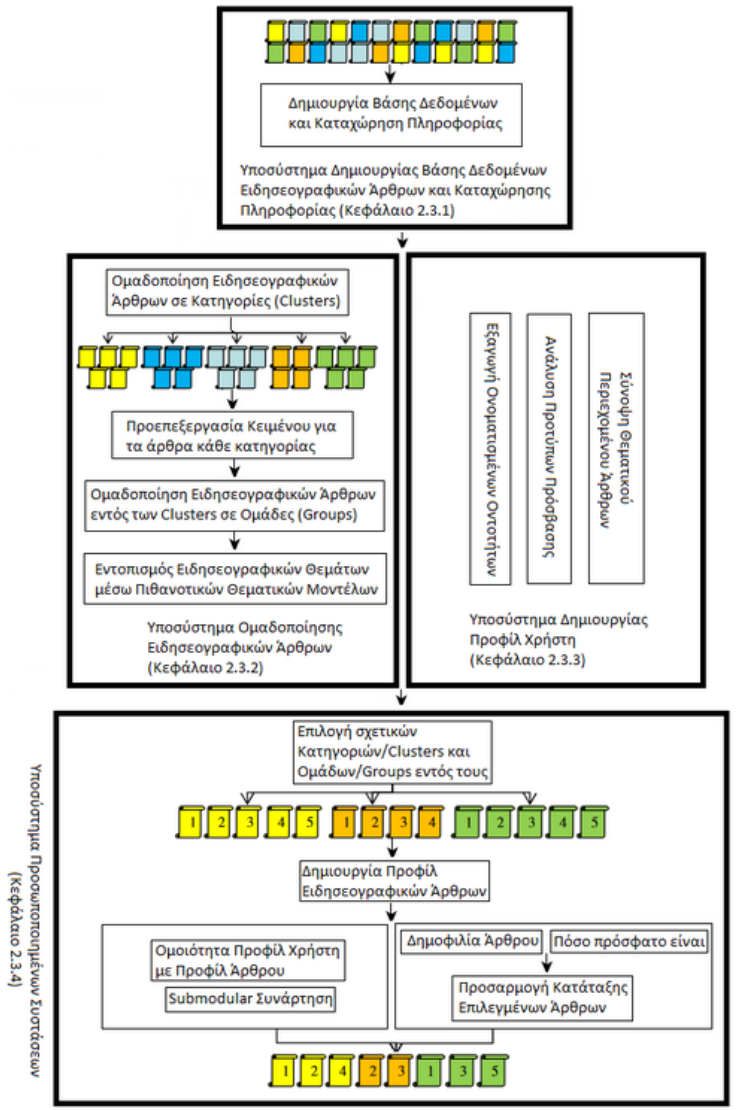
Να βρούμε σύντομες περιγραφές των μελών της συλλογής που επιτρέπουν μία αποτελεσματική επεξεργασία μεγάλων συλλογών, διατηρώντας τις απαραίτητες στατιστικές σχέσεις που είναι χρήσιμες για βασικές διεργασίες, όπως η περίληψη κειμένου.

Επεξεργασία Φυσικής Γλώσσας

Ανάλυση Φυσικής Γλώσσας με χρήση Θεματικών Μοντέλων

- Στο σύστημά μας εφαρμόζουμε το **Latent Dirichlet Allocation (LDA)** ως το μοντέλο για την ανακάλυψη κρυμμένων θεμάτων.
- Αναπαράσταση της κατανομής θεμάτων ως ένα διάνυσμα κάθε εγγραφή του οποίου δηλώνει το βάρος της αντίστοιχης λέξης.
- Εφαρμογή μέσω της βιβλιοθήκης gensim της Python.

Περιεχόμενα


Δημιουργία Βάσης Δεδομένων και Καταχώρηση Πληροφορίας
Υποσύστημα Δημιουργίας Βάσης Δεδομένων Ειδησεογραφικών Άρθρων και Καταχώρησης Πληροφορίας (Κεφάλαιο 2.3.1)
Ομαδοποίηση Ειδησεογραφικών Άρθρων σε Κατηγορίες (Clusters)
Προεπεξεργασία Κειμένου για τα άρθρα κάθε κατηγορίας
Ομαδοποίηση Ειδησεογραφικών Άρθρων εντός των Clusters σε Ομάδες (Groups)
Εντοπισμός Ειδησεογραφικών Θεμάτων μέσω Πιθανοτικών Θεματικών Μοντέλων
Υποσύστημα Ομαδοποίησης Ειδησεογραφικών Άρθρων (Κεφάλαιο 2.3.2)
Εξαγωγή Ονοματισμένων Οντοτήτων
Ανάλυση Προτύπων Πρόσβασης
Σύνοψη Θεματικού Περιεχομένου Άρθρων
Υποσύστημα Δημιουργίας Προφίλ Χρήστη (Κεφάλαιο 2.3.3)
Υποσύστημα Προσωποποιημένων Συστάσεων (Κεφάλαιο 2.3.4)
Επιλογή σχετικών Κατηγοριών/Clusters και Ομάδων/Groups εντός τους
1 2 3 4 5 1 2 3 4 1 2 3 4 5
Δημιουργία Προφίλ Ειδησεογραφικών Άρθρων
Ομοιότητα Προφίλ Χρήστη με Προφίλ Άρθρου
Submodular Συνάρτηση
Δημοφιλία Άρθρου
Πόσο πρόσφατο είναι
Προσαρμογή Κατάταξης Επιλεγμένων Άρθρων
1 2 4 2 3 1 3 5

Περιεχόμενα

Υποσυστήματα μηχανισμού

1. **Υποσύστημα Δημιουργίας Βάσης Δεδομένων και Καταχώρησης Πληροφορίας**
2. **Υποσύστημα Ομαδοποίησης Ειδησεογραφικών Άρθρων**
3. **Υποσύστημα Δημιουργίας Προφίλ Χρήστη**
4. **Υποσύστημα Προσωποποιημένων Συστάσεων**

Υποσυστήματα μηχανισμού

1. Υποσύστημα Δημιουργίας ΒΔ και Καταχώρησης Πληροφορίας

- **Συλλογή Ειδησεογραφικών Άρθρων:**

  The Guardian, New York Times, Washington Post, Fox News, Independent, Reuters και συλλογή άρθρων Reuters του NLTK.

- **Δημιουργία ΒΔ και Καταχώρηση Πληροφορίας:**
  - Πληροφορίες που αποθηκεύονται για κάθε άρθρο: τίτλος, συγγραφέας, ημερομηνία δημοσίευσης, κείμενο άρθρου, γενική κατηγορία στην οποία ανήκει.
  - Δημιουργία χρηστών και αναγνωστικού ιστορικού (αποθηκευμένων) χρηστών.

Υποσυστήματα μηχανισμού

2. Υποσύστημα Ομαδοποίησης Ειδησεογραφικών Άρθρων

- Ομαδοποίηση Ειδησεογραφικών Άρθρων σε Κατηγορίες (Clusters)
- Προεπεξεργασία Κειμένου (Text Preprocessing)
- Ομαδοποίηση Ειδησεογραφικών Άρθρων εντός των Clusters σε Ομάδες (Groups)
- Εντοπισμός Ειδησεογραφικών Θεμάτων μέσω Πιθανοτικών Θεματικών Μοντέλων

Υποσυστήματα μηχανισμού

2. Υποσύστημα Ομαδοποίησης Ειδησεογραφικών Άρθρων

- **Ομαδοποίηση Ειδησεογραφικών Άρθρων σε Κατηγορίες (Clusters):**
  - Άρθρα από εφτά διαφορετικές κατηγορίες (Science/Technology, Politics, Sports, Life & Style, Sugar, Coffee, Housing).
  - Κάθε άρθρο ανήκει αποκλειστικά σε ένα και μοναδικό cluster.
  - Ομαδοποίηση βάσει προεπιλεγμένης κατηγορίας.

Υποσυστήματα μηχανισμού

2. Υποσύστημα Ομαδοποίησης Ειδησεογραφικών Άρθρων

- **Προεπεξεργασία Κειμένου (Text Preprocessing):**
  - Λεξική Ανάλυση: Διαμερισμός άρθρων στα συστατικά στοιχεία του κειμένου τους (tokenization).
  - Αφαίρεση τετριμμένων λέξεων και τερματικών όρων: Καθαρισμός από άρθρα, συνδέσμους, αντωνυμίες και συχνά χρησιμοποιούμενες λέξεις χωρίς ιδιαίτερη σημασιολογική πληροφορία.
  - Κανονικοποίηση των λέξεων: Αναγνώριση ριζών λέξεων (lemmatization) και αποκατάληξη (stemming).
  - Επιλογή των αντιπροσωπευτικών όρων: Εφαρμογή tf-idf (Η tf-idf έχει μεγάλη τιμή για έναν όρο και επομένως, είναι σημαντικός για ένα κείμενο όταν ο όρος εμφανίζεται συχνά σε ένα κείμενο της συλλογής και σπάνια στα υπόλοιπα κείμενα).

Υποσυστήματα μηχανισμού

2. Υποσύστημα Ομαδοποίησης Ειδησεογραφικών Άρθρων

- **Ομαδοποίηση Ειδησεογραφικών Άρθρων εντός των Clusters σε Ομάδες (Groups):**
  - Εφαρμογή αλγορίθμου ομαδοποίησης **k-means** (βιβλιοθήκη scikit-learn της Python). Διάσπαση δεδομένων σε k διαφορετικές ομάδες μέσω μίας επαναληπτικής διαδικασίας.
  - Κριτήριο σύγκλισης αλγορίθμου, τρόπος μέτρησης απόστασης δεδομένων από τα κέντρα των ομάδων, τρόπος ανάδειξης αρχικών κέντρων: ορίζονται από το χρήστη.
  - 1ο επίπεδο σύστασης άρθρων του συστήματος: Τα ήδη υπάρχοντα clusters μαζί με τα παραγόμενα groups που περιέχονται σε αυτά.

Υποσυστήματα μηχανισμού

2. Υποσύστημα Ομαδοποίησης Ειδησεογραφικών Άρθρων

- **Εντοπισμός Ειδησεογραφικών Θεμάτων μέσω Πιθανοτικών Θεματικών Μοντέλων:**
  - Ανακάλυψη κρυμμένων θεμάτων μίας συλλογής κειμένων.
  - Λίστα αντιπροσωπευτικών λέξεων από την αρχική συλλογή κειμένων μαζί με το αντίστοιχο βάρος για κάθε λέξη (μοντελοποίηση μίας συλλογής ως ένα πεπερασμένο μείγμα από ένα σύνολο θεματικών πιθανοτήτων).
  - Εξερεύνηση των συσχετισμών μεταξύ των clusters (ή των groups) άρθρων και του προφίλ του δοθέντος χρήστη.

Υποσυστήματα μηχανισμού

2. Υποσύστημα Ομαδοποίησης Ειδησεογραφικών Άρθρων

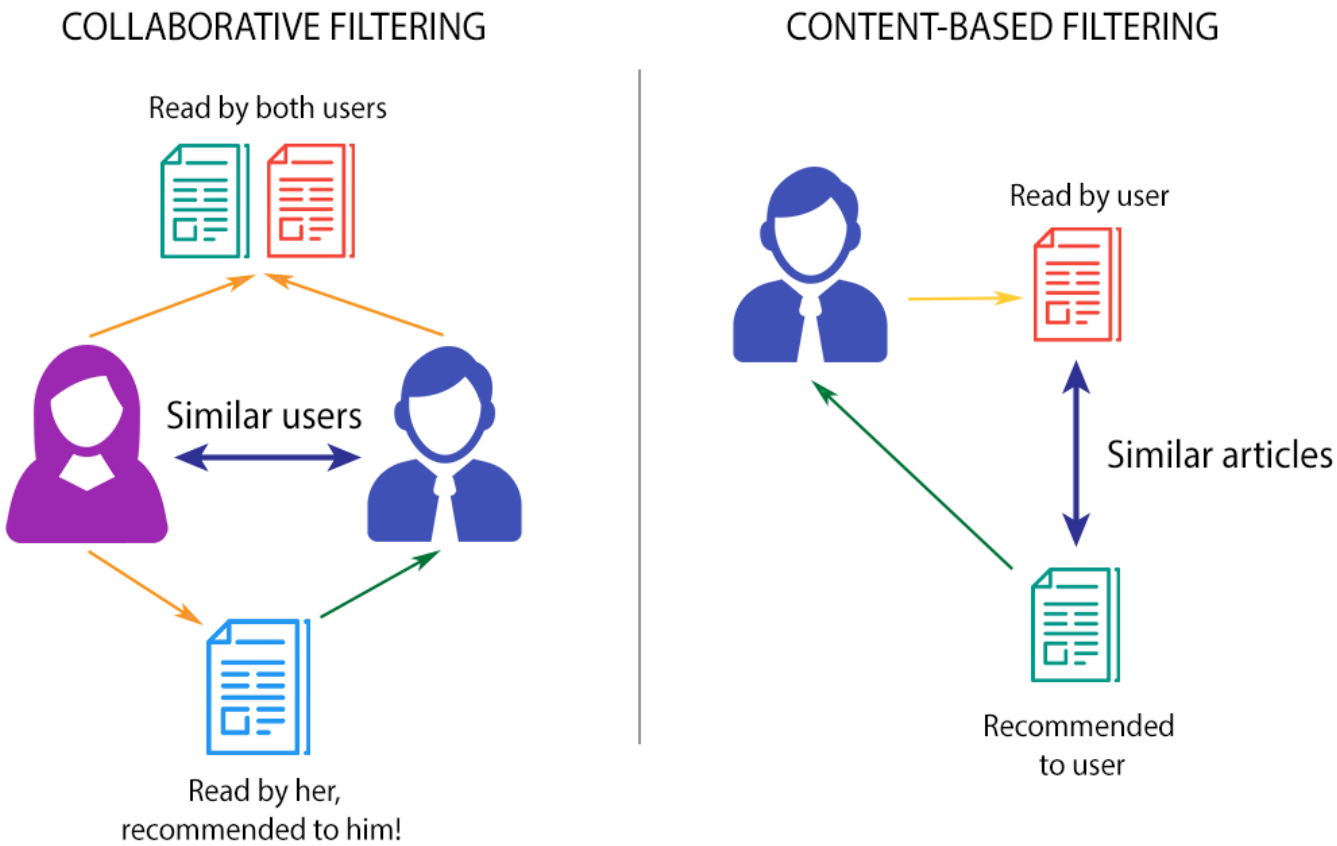
- **Εντοπισμός Ειδησεογραφικών Θεμάτων μέσω Πιθανοτικών Θεματικών Μοντέλων:**
  - Εφαρμογή του **LDA** σε κάθε cluster (κατηγορία) άρθρων, σε κάθε group εντός των clusters, καθώς και σε κάθε μεμονωμένο άρθρο.
  - Διάνυσμα κατανομής θεμάτων τόσο για τα άρθρα, όσο και για τα clusters και τα groups που έχουν δημιουργηθεί.
  - Κάθε καταχώρηση ενός τέτοιου διανύσματος θεμάτων αποτελείται από μία αντιπροσωπευτική λέξη και το αντίστοιχο βάρος. Επιλογή αριθμού αντιπροσωπευτικών λέξεων απ'τις οποίες θα αποτελείται ένα τέτοιο διάνυσμα κατά την εφαρμογή του αλγορίθμου.

Υποσυστήματα μηχανισμού

3. Υποσύστημα Δημιουργίας Προφίλ Χρήστη

Το **Προφίλ Χρήστη** παραμετροποιείται μέσω μίας τριπλέτας U = <T, P, E>, όπου:

- **T**: κατανομή θεμάτων άρθρων του ιστορικού ανάγνωσης του χρήστη **(Σύνοψη Θεματικού Περιεχομένου Άρθρων)**.
- **P**: Λίστα χρηστών οι οποίοι έχουν παρόμοιες αναγνωστικές προτιμήσεις **(Ανάλυση Προτύπων Πρόσβασης)**.
- **E**: Λίστα από ονοματισμένες οντότητες εξαγόμενες από το ιστορικό αναγνωσμένων άρθρων του χρήστη, συσχετισμένες με τον αντίστοιχο τύπο οντότητας **(Εξαγωγή Ονοματισμένων Οντοτήτων)**.


COLLABORATIVE FILTERING
Read by both users
Similar users
Read by her,
recommended to him!
CONTENT-BASED FILTERING
Read by user
Similar articles
Recommended
to user

Υποσυστήματα μηχανισμού

3. Υποσύστημα Δημιουργίας Προφίλ Χρήστη

- **Σύνοψη Θεματικού Περιεχομένου Άρθρων:**
  - Ίδια στρατηγική που εφαρμόσαμε και για τον εντοπισμό ειδησεογραφικών θεμάτων μέσω πιθανοτικών θεματικών μοντέλων στα clusters άρθρων.
  - Ίδια αναπαράσταση άρθρων (Διάνυσμα κατανομής θεμάτων κάθε καταχώρηση του οποίου αποτελείται από μία αντιπροσωπευτική λέξη και το αντίστοιχο βάρος).

Υποσυστήματα μηχανισμού

3. Υποσύστημα Δημιουργίας Προφίλ Χρήστη

- **Ανάλυση Προτύπων Πρόσβασης:**
  - Το προφίλ ενός χρήστη μπορεί να εμπλουτιστεί αναλύοντας τις αναγνωστικές προτιμήσεις άλλων χρηστών παρόμοιων με το δεδομένο χρήστη και ενσωματώνοντάς τες σε αυτό.
  - Ανάλυση του ιστορικού αναγνωσμένων άρθρων όλων των χρηστών του συστήματος.
  - Υπολογισμός των ανά ζεύγος ομοιοτήτων Jaccard του ιστορικού αναγνωσμένων άρθρων μεταξύ των χρηστών του συστήματος. (Ομοιότητα Jaccard για δύο σύνολα είναι το μέγεθος της τομής προς το μέγεθος της ένωσής τους).
  - Εξαγωγή λίστας με τα ονόματα παρόμοιων χρηστών για κάθε χρήστη.
  - Αποθήκευση λίστας στη ΒΔ και ενημέρωση καθ'όλη τη διάρκεια περιήγησης του χρήστη στο σύστημα.

Υποσυστήματα μηχανισμού

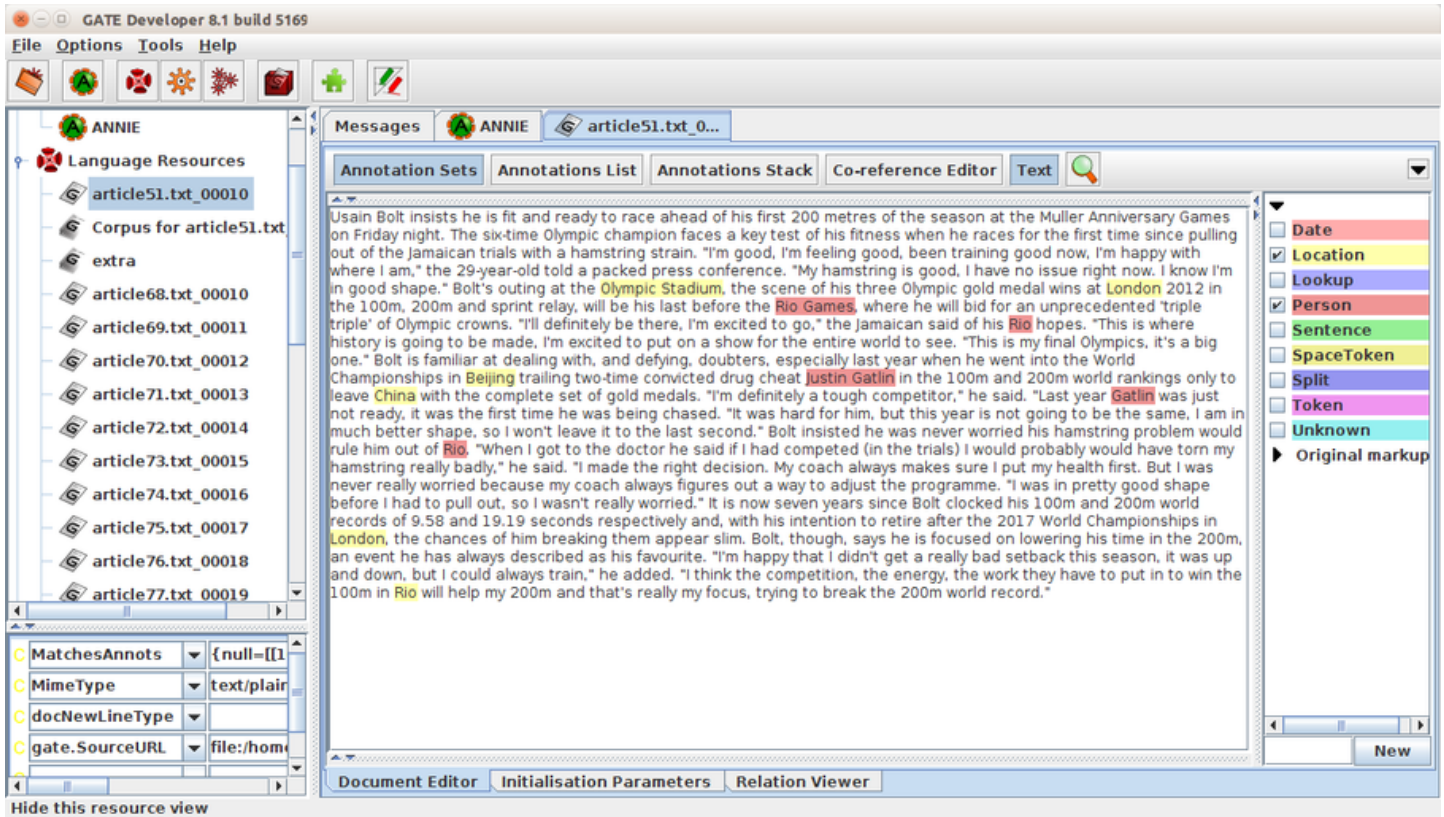
3. Υποσύστημα Δημιουργίας Προφίλ Χρήστη

- **Εξαγωγή Ονοματισμένων Οντοτήτων:**
  - Προτίμηση αναγνωστών σε λέξεις/φράσεις όπως 'πότε, πού, τι συνέβη, ποιος εμπλέκεται'
  - Χρήση εργαλείου επεξεργασίας Φυσικής Γλώσσας GATE (General Architecture for Text Engineering)
  - Εξαγωγή πληροφορίας για οντότητες τύπου "Organization", "Person" και "Location".
  - Εξαγωγή λίστας με ονοματισμένες οντότητες και τον αντίστοιχο τύπο τους για κάθε άρθρο.

Υποσυστήματα μηχανισμού

3. Υποσύστημα Δημιουργίας Προφίλ Χρήστη

Εξαγωγή ονοματισμένων οντοτήτων με χρήση του GATE:



Υποσυστήματα μηχανισμού

4. Υποσύστημα Προσωποποιημένων Συστάσεων

- Αντιστοίχιση Αναγνωστικών Προτιμήσεων για την Αναπαράσταση 1ου Επιπέδου.
- Αντιστοίχιση Αναγνωστικών Προτιμήσεων για την Αναπαράσταση 2ου Επιπέδου.
  - Δημιουργία Προφίλ Ειδησεογραφικών Άρθρων.
  - Μοντέλο Συστάσεων.
  - Προσαρμογή Κατάταξης Ειδησεογραφικών Άρθρων.

Υποσυστήματα μηχανισμού

4. Υποσύστημα Προσωποποιημένων Συστάσεων

- **Αντιστοίχιση Αναγνωστικών Προτιμήσεων για την Αναπαράσταση 1ου Επιπέδου:**
  - Δεδομένα: Ιεραρχία των ειδησεογραφικών άρθρων (συσταδοποίηση σε clusters και groups εντός αυτών), καθώς και το προφίλ του χρήστη.
  - Διαδοχική αντιστοίχιση του προφίλ του χρήστη στην ιεραρχία ειδήσεων (Σύγκριση βαθμού ομοιότητας κατανομής θεμάτων κάθε ομάδας με αυτή του προφίλ χρήστη, μέσω της ομοιότητας συνημιτόνου).
  - Επιλογή των clusters με σκορ ομοιότητας μεγαλύτερο από ένα δυναμικό κατώφλι (Κάθε cluster αντιστοιχίζεται σε μία κατηγορία θεμάτων).
  - Ομοίως, εισχωρούμε σε κάθε επιλεγμένο cluster και επιλέγουμε το group νέων που είναι πιο κοντά σε ομοιότητα με τις προτιμήσεις του χρήστη.

Υποσυστήματα μηχανισμού

4. Υποσύστημα Προσωποποιημένων Συστάσεων

- **Αντιστοίχιση Αναγνωστικών Προτιμήσεων για την Αναπαράσταση 2ου Επιπέδου:**
  - Δεδομένα: Clusters και groups που πιθανότατα ενδιαφέρουν το χρήστη.
  - Στόχος: Επιλογή συγκεκριμένων άρθρων προς αναπαράσταση.
  - Δημιουργία προφίλ ειδησεογραφικών άρθρων.

Υποσυστήματα μηχανισμού

4. Υποσύστημα Προσωποποιημένων Συστάσεων

Δημιουργία Προφίλ Ειδησεογραφικών Άρθρων:
Το Προφίλ Άρθρου παραμετροποιείται μέσω μίας τριπλέτας F = <T, P, E>, όπου:

- **T**: κατανομή θεμάτων άρθρου (Σύνοψη Θεματικού Περιεχομένου Άρθρου).
- **P**: Λίστα χρηστών οι οποίοι έχουν αναγνώσει το άρθρο (Ανάλυση Προτύπων Πρόσβασης).
- **E**: Λίστα από ονοματισμένες οντότητες εξαγόμενες από το άρθρο (Εξαγωγή Ονοματισμένων Οντοτήτων).

Στατικά χαρακτηριστικά άρθρου: Κατανομή θεμάτων, λίστα χρηστών, ονοματισμένες οντότητες.
Δυναμικά χαρακτηριστικά άρθρου: Χρήστες που το διάβασαν, δημοφιλία, πόσο 'φρέσκο' είναι από τη σκοπιά του πόσο πρόσφατα δημοσιεύθηκε.

Υποσυστήματα μηχανισμού

4. Υποσύστημα Προσωποποιημένων Συστάσεων

Μοντέλο Συστάσεων:

- Δεδομένα: Λίστες άρθρων, μία για κάθε group άρθρων που επιλέχθηκε βάσει ομοιότητας.
- Αφαιρούμε από κάθε επιλεγμένο group άρθρων τα άρθρα τα οποία βρίσκονται ήδη στο ιστορικό ανάγνωσης του χρήστη.
- Καταλήγουμε με μία λίστα από λίστες, κάθε μία εκ των οποίων περιλαμβάνει τα εναπομείναντα υποψήφια προς πρόταση άρθρα από κάθε group.
- Υπολογίζουμε το budget B για κάθε λίστα, δηλαδή το μέγιστο αριθμό από προτεινόμενα άρθρα μέσα σε κάθε group.

Υποσυστήματα μηχανισμού

4. Υποσύστημα Προσωποποιημένων Συστάσεων

Υπολογισμός budget B βάσει ομοιότητας χρήστη με την εν λόγω κατηγορία:

- Όσο μεγαλύτερη είναι η τιμή που προκύπτει από τη σύγκριση μεταξύ των κατανομών θεμάτων χρήστη και του εκάστοτε cluster, τόσο μεγαλύτερος είναι ο αριθμός άρθρων που επιλέγονται να προταθούν από την εν λόγω κατηγορία άρθρων.

Υποσυστήματα μηχανισμού

4. Υποσύστημα Προσωποποιημένων Συστάσεων

Μοντέλο Συστάσεων:

- Υπολογισμός για κάθε group όλων των δυνατών συνδυασμών άρθρων, μεγέθους ίσου με το αντίστοιχο budget.
- Για παράδειγμα, αν το budget για ένα group είναι ίσο με την τιμή v, τότε υπολογίζουμε όλες τις πιθανές v-άδες άρθρων.
- Υλοποίηση συνάρτησης αξιολόγησης κάθε v-άδας με στόχο την επιλογή της v-άδας που προσφέρει τη μέγιστη αύξηση ωφέλειας για το χρήστη (αποδείχθηκε υπερβολικά χρονοβόρα διαδικασία για τα δεδομένα ενός συστήματος συστάσεων - Αναφορά σε φιλοσοφία submodular συναρτήσεων).
- Εφαρμογή random συνάρτησης για επιλογή v-άδας άρθρων από κάθε group.

Υποσυστήματα μηχανισμού

4. Υποσύστημα Προσωποποιημένων Συστάσεων

Προσαρμογή Κατάταξης Ειδησεογραφικών Άρθρων:

- Δεδομένα: Λίστα με ειδησεογραφικά άρθρα από κάθε επιλεγμένη θεματική κατηγορία.
- Δημοφιλία άρθρων και πόσο πρόσφατα δημοσιευμένα είναι.
- Επιλέγουμε διαδοχικά δύο γειτονικά άρθρα από την κορυφή της λίστας προς τα κάτω και τα συγκρίνουμε ως προς το δυναμικό σκορ (συνδυασμός των παραπάνω δύο χαρακτηριστικών).
- Όσο πιο πρόσφατα δημοσιεύθηκε ένα άρθρο, τόσο υψηλότερη θέση παίρνει στην τελική κατάταξη.
- Η παραγόμενη κατάταξη δίνει έμφαση στα πιο δημοφιλή και 'φρέσκα' ειδησεογραφικά άρθρα.

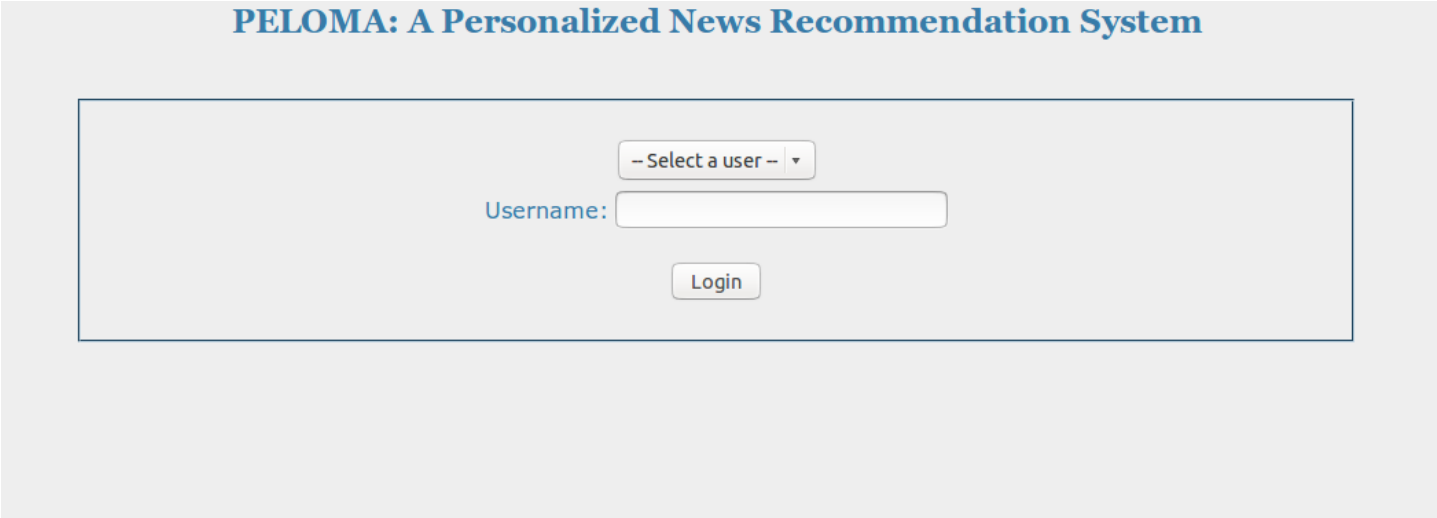
Εργαλεία & Τεχνολογίες Υλοποίησης

- Γλώσσα προγραμματισμού Python, έκδοση 3.4.0
- MySQL: Σύστημα διαχείρισης σχεσιακών ΒΔ
- Web Technologies: Flask Web Framework, H™L, CSS, Apache HTTP Server
- Εργαλεία Επεξεργασίας Φυσικής Γλώσσας:
  - NLTK: Sentence Tokenization, Word Tokenization, Removing Stopwords
  - GATE for Named Entity Recognition: Αναγνώριση διαφορετικών τύπων οντοτήτων, όπως ονόματα, τοποθεσίες και οργανισμοί, απαντώντας σε ερωτήσεις όπως 'Τι συνέβη, ποιος εμπλέκεται, πότε συνέβη'

Περιεχόμενα

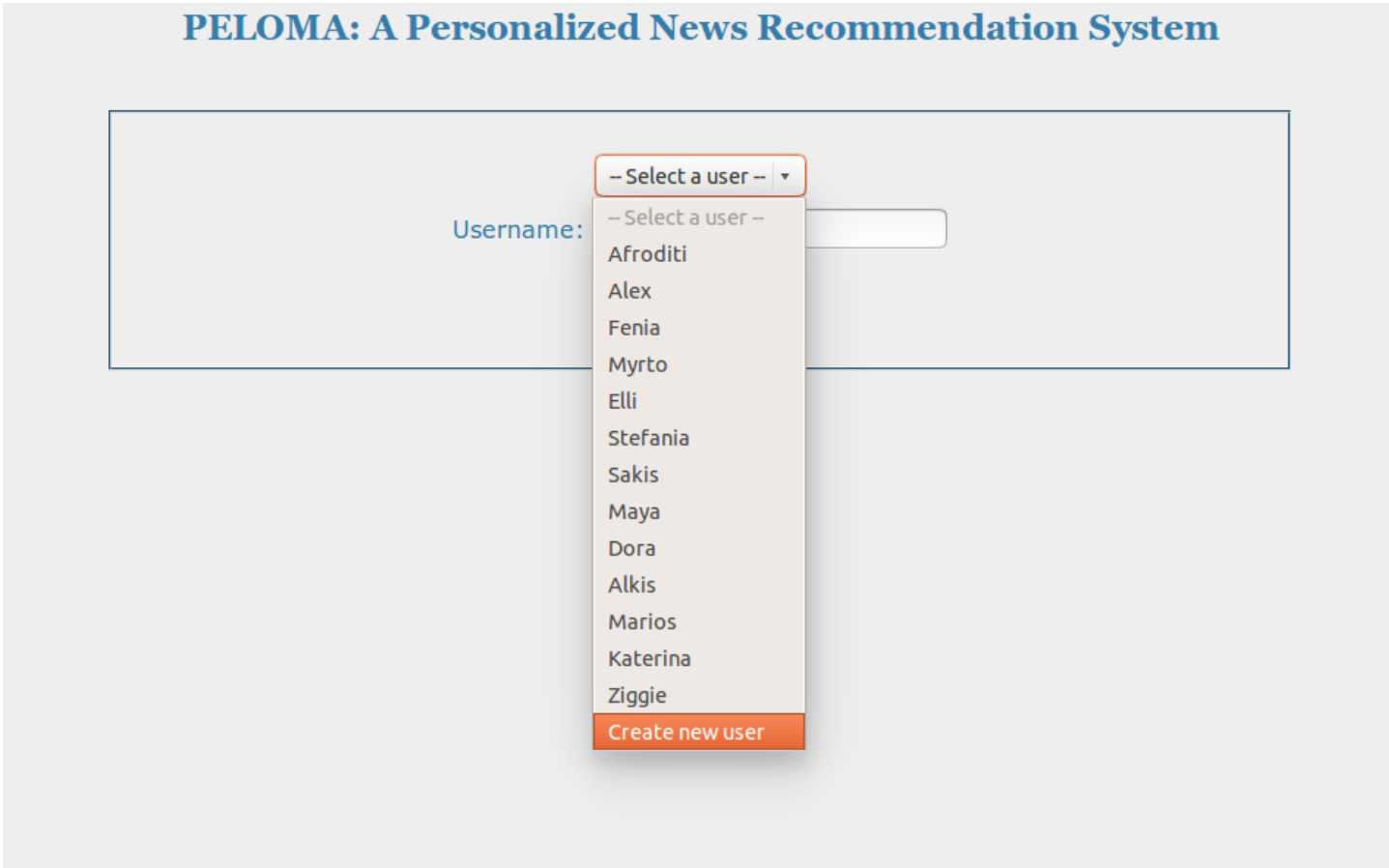
Σύστημα Συστάσεων

Είσοδος χρήστη στο σύστημα



Σύστημα Συστάσεων

Είσοδος χρήστη στο σύστημα



Σύστημα Συστάσεων

Προσπάθεια λήψης συστάσεων πριν την ανάγνωση άρθρων

Σύστημα Συστάσεων

Προβολή λίστας άρθρων Βάσης Δεδομένων

PELOMA: A Personalized News Recommendation System

Afroditi's reading history

Recommendations for Afroditi

Category	id	Articles
1	1	How we all could benefit from synaesthesia
	2	Children of older men at greater risk of mental illness, study suggests
	3	The mysteries of 'lucid' dreaming
	4	Taller people more likely to get cancer, say researchers
	5	Universe recreated in massive computer simulation
	6	Ketamine may help treat depression, UK study finds
	7	Can stress really make us sick?
	8	Loss of vision strengthens sense of hearing, study finds
	9	Scientists say 'runner's high' is like a marijuana high
	10	Study of Holocaust survivors finds trauma passed on to children's genes
	11	Mysteries of computer from 65BC are solved
	12	How did the Enigma machine work?
	13	How will the internet of things impact data security?
	14	Google introduces 'time machine' feature in Street View
	15	European Court Lets Users Erase Records on Web
	16	Wikipedia's view of the world is written by the west
	17	Reading, Writing, Arithmetic, and Lately, Coding
	18	Independent booksellers bolstered in fight against Amazon
	19	If a robot rocks my son to sleep, am I still his parent?
	20	The future of shopping: drones, digital mannequins and leaving without paying

Σύστημα Συστάσεων

Προβολή πλήρους κειμένου ενός άρθρου

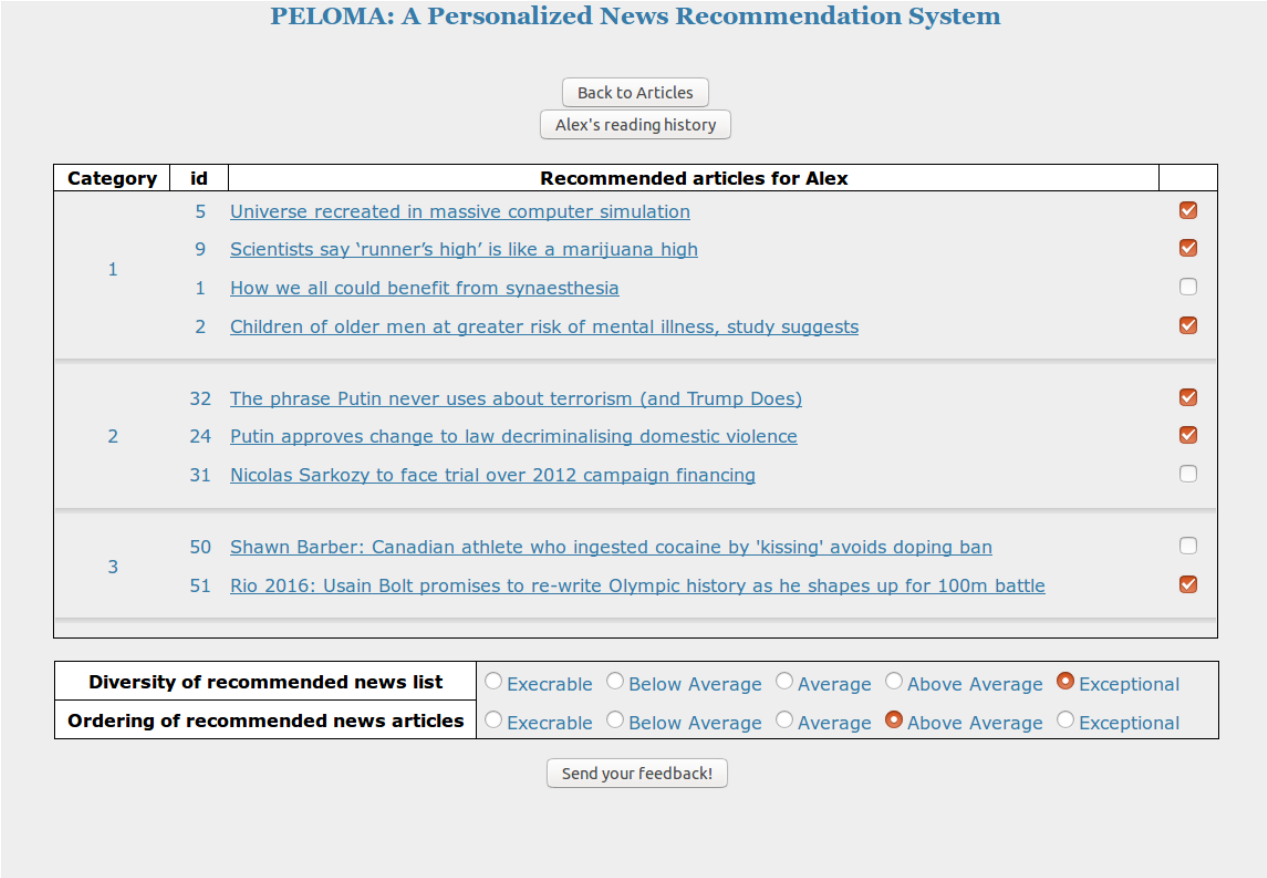
PELOMA: A Personalized News Recommendation System

Back to Articles

Article #12
Title: How did the Enigma machine work?
Like all the best cryptography, the Enigma machine is simple to describe, but infuriating to break. Straddling the border between mechanical and electrical, Enigma looked from the outside like an oversize typewriter. Enter the first letter of your message on the keyboard and a letter lights up showing what it has replaced within the encrypted message. At the other end, the process is the same: type in the "ciphertext" and the letters which light are the decoded missive. Inside the box, the system is built around three physical rotors. Each takes in a letter and outputs it as a different one. That letter passes through all three rotors, bounces off a "reflector" at the end, and passes back through all three rotors in the other direction. The board lights up to show the encrypted output, and the first of the three rotors clicks round one position – changing the output even if the second letter input is the same as the first one. When the first rotor has turned through all 26 positions, the second rotor clicks round, and when that's made it round all the way, the third does the same, leading to more than 17,000 different combinations before the encryption process repeats itself. Adding to the scrambling was a plugboard, sitting between the main rotors and the input and output, which swapped pairs of letters. In the earliest machines, up to six pairs could be swapped in that way; later models pushed it to 10, and added a fourth rotor.Despite the complexity, all the operators needed was information about the starting position, and order, of the three rotors, plus the positions of the plugs in the board. From there, decoding is as simple as typing the cyphertext back into the machine. Thanks to the reflector, decoding was the same as encoding the text, but in reverse.But that reflector also led to the flaw in Enigma, and the basis on which all codebreaking efforts were founded: no letter would ever be encoded as itself. With that knowledge, as well as an educated guess at what might be encrypted in some of the messages (common phrases included "Keine besonderen Ereignisse", or "nothing to report" and "An die Gruppe", or "to the group"), it was possible to eliminate thousands of potential rotor positions. Eventually, the team at Bletchley Park built a machine, the Bombe, which could handle that logical analysis. But the final steps were always performed manually: the job of the Bombe was merely to reduce the number of combinations that the cryptanalysts had to examine. Even as the Allied code-breaking team were working on Enigma, the Axis was improving its machines, adding more and different rotors, and minimising operator error. Eventually, the Enigma was superseded by the Lorenz. These required yet more codebreaking in Britain, and more automation to do it – leading to the production of Colossus, the world's first digital programmable computer.

Σύστημα Συστάσεων

Προσωποποιημένες Συστάσεις Άρθρων



PELOMA: A Personalized News Recommendation System

Back to Articles

Alex's reading history

Category	id	Recommended articles for Alex	
1	5	Universe recreated in massive computer simulation	☑
	9	Scientists say 'runner's high' is like a marijuana high	☑
	1	How we all could benefit from synaesthesia	☐
	2	Children of older men at greater risk of mental illness, study suggests	☑
2	32	The phrase Putin never uses about terrorism (and Trump Does)	☑
	24	Putin approves change to law decriminalising domestic violence	☑
	31	Nicolas Sarkozy to face trial over 2012 campaign financing	☐
3	50	Shawn Barber: Canadian athlete who ingested cocaine by 'kissing' avoids doping ban	☐
	51	Rio 2016: Usain Bolt promises to re-write Olympic history as he shapes up for 100m battle	☑

Diversity of recommended news list	○ Execrable	○ Below Average	○ Average	○ Above Average	◉ Exceptional
Ordering of recommended news articles	○ Execrable	○ Below Average	○ Average	◉ Above Average	○ Exceptional

Send your feedback!

Σύστημα Συστάσεων

Αξιολόγηση Συστήματος από τη χρήστη

Αξιολόγηση Συστήματος Συστάσεων

Κριτήρια υπολογισμού ικανοποίησης χρήστη

Αξιολόγηση του συστήματος από 50 χρήστες ως προς τα παρακάτω κριτήρια:

- Συσχετισμός των προτεινόμενων άρθρων με τα πραγματικά τους ενδιαφέροντα (Preference)
- Ποικιλία της λίστας συστάσεων (Diversity)
- Κατάταξη των άρθρων της λίστας συστάσεων (Ordering)

Αξιολόγηση Συστήματος Συστάσεων

Κριτήρια υπολογισμού ικανοποίησης χρήστη

Αξιολόγηση του συστήματος από 50 χρήστες ως προς τα παρακάτω κριτήρια:

- Συσχετισμός των προτεινόμενων άρθρων με τα πραγματικά τους ενδιαφέροντα (Preference)
- Ποικιλία της λίστας συστάσεων (Diversity)
- Κατάταξη των άρθρων της λίστας συστάσεων (Ordering)

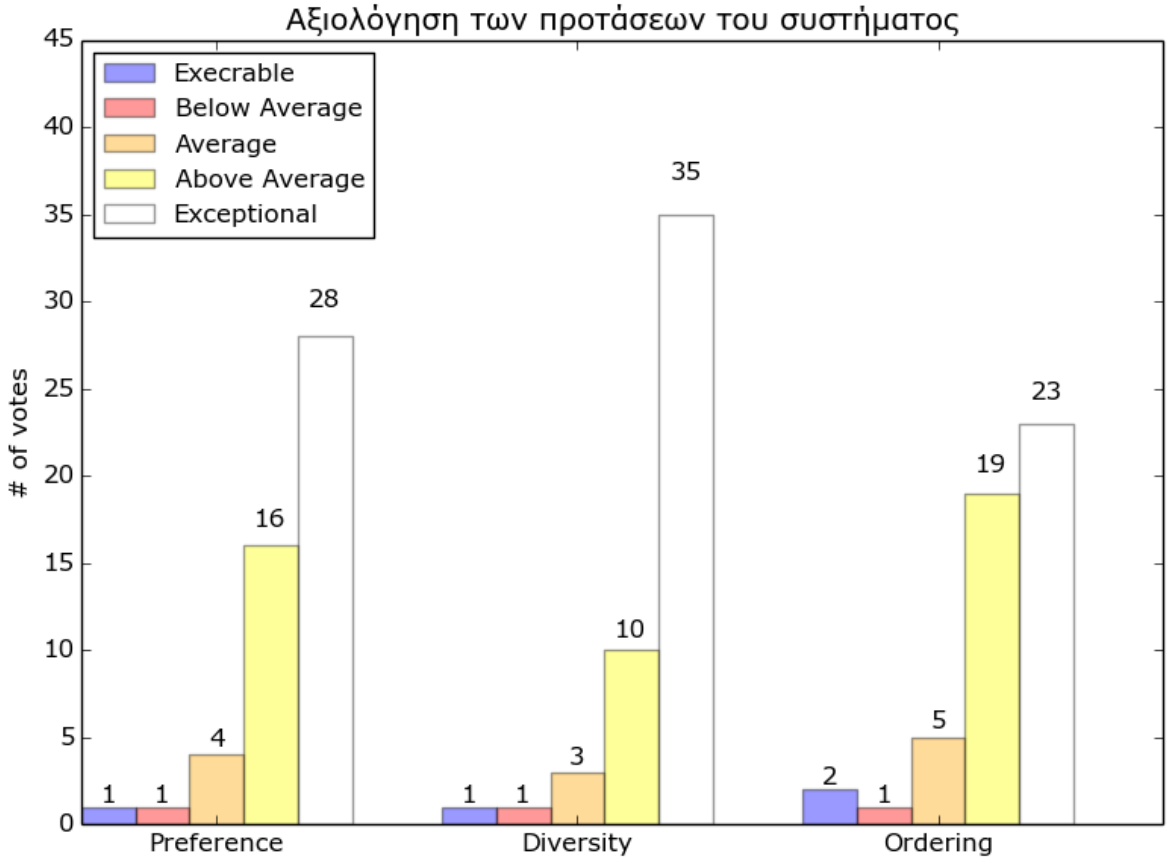
Αξιολόγηση Συστήματος Συστάσεων

Κριτήρια υπολογισμού ικανοποίησης χρήστη

Αξιολόγηση του συστήματος από 50 χρήστες ως προς τα παρακάτω κριτήρια:

- Συσχετισμός των προτεινόμενων άρθρων με τα πραγματικά τους ενδιαφέροντα (Preference)
- Ποικιλία της λίστας συστάσεων (Diversity)
- Κατάταξη των άρθρων της λίστας συστάσεων (Ordering)

Αποτελέσματα πειραματικής αξιολόγησης



Αποτελέσματα πειραματικής αξιολόγησης

Ιστορικό ανάγνωσης και 1η φάση συστάσεων

Διαμόρφωση συστάσεων προς έναν χρήστη κατά το χρόνο παραμονής του στο σύστημα, όπου δέχεται συστάσεις και συνεχίζει την ανάγνωση άρθρων:

1st Recommendation	# of articles in user's history	Similarity with Category	# of recommended articles
Category #1	4	0.2607	0
Category #2	**6**	**0.3544**	**3**
Category #3	**7**	**0.3638**	**4**
Category #4	4	0.2469	0
Category #5	4	0.3427	0
Category #6	**4**	**0.3466**	**2**
Category #7	2	0.26788	0

Αποτελέσματα πειραματικής αξιολόγησης

Ιστορικό ανάγνωσης και 2η φάση συστάσεων

2nd Recommendation	# of articles in user's history	Similarity with Category	# of recommended articles
Category #1	4	0.2846	0
Category #2	9	0.3273	0
Category #3	**7**	**0.3566**	**2**
Category #4	4	0.2513	0
Category #5	**5**	**0.3677**	**3**
Category #6	**5**	**0.3919**	**4**
Category #7	2	0.2696	0

Αποτελέσματα πειραματικής αξιολόγησης

Ιστορικό ανάγνωσης και 3η φάση συστάσεων

3rd Recommendation	# of articles in user's history	Similarity with Category	# of recommended articles
Category #1	4	0.2853	0
Category #2	**12**	**0.3837**	**2**
Category #3	8	0.3578	0
Category #4	4	0.2779	0
Category #5	**6**	**0.4013**	**3**
Category #6	**7**	**0.4026**	**4**
Category #7	2	0.3012	0

Αποτελέσματα πειραματικής αξιολόγησης

Λήψη Συστάσεων

Σενάριο χρήσης του συστήματος όπου ένας χρήστης με κενό αναγνωστικό ιστορικό συνδέεται στο σύστημα και επιλέγει προς ανάγνωση άρθρα μόνο από μία κατηγορία, την Κατηγορία #1:

```
--- Cluster 1 Group 0: [5, 17, 19]
--- Cluster 1 Group 1: [1, 2, 4, 6, 7, 8, 9, 10]
--- Cluster 1 Group 2: [3, 11, 12, 13, 14, 15, 16, 18, 20]
```

Σχήμα : Groups άρθρων της Κατηγορίας #1.

Αποτελέσματα πειραματικής αξιολόγησης

Λήψη Συστάσεων

Category	id	Recommended articles for Antonis	Preference
1	19	If a robot rocks my son to sleep, am I still his parent?	☐
	17	Reading, Writing, Arithmetic, and Lately, Coding	☐
	5	Universe recreated in massive computer simulation	☐

Category	id	Recommended articles for Antonis	Preference
1	7	Can stress really make us sick?	☐
	2	Children of older men at greater risk of mental illness, study suggests	☐
	4	Taller people more likely to get cancer, say researchers	☐
	8	Loss of vision strengthens sense of hearing, study finds	☐

Category	id	Recommended articles for Antonis	Preference
1	9	Scientists say 'runner's high' is like a marijuana high	☐
	6	Ketamine may help treat depression, UK study finds	☐
	1	How we all could benefit from synaesthesia	☐
	10	Study of Holocaust survivors finds trauma passed on to children's genes	☐

Αποτελέσματα πειραματικής αξιολόγησης
Λήψη Συστάσεων

Category	id	Recommended articles for Antonis	Preference
1	15	European Court Lets Users Erase Records on Web	☐
	14	Google introduces 'time machine' feature in Street View	☐
	13	How will the internet of things impact data security?	☐
	20	The future of shopping: drones, digital mannequins and leaving without paying	☐

Category	id	Recommended articles for Antonis	Preference
1	11	Mysteries of computer from 65BC are solved	☐
	16	Wikipedia's view of the world is written by the west	☐
	3	The mysteries of 'lucid' dreaming	☐
	18	Independent booksellers bolstered in fight against Amazon	☐

Συμπεράσματα

- Το σύστημα είναι σε θέση να καλύψει επαρκώς τις αναγνωστικές ανάγκες του χρήστη με βάση τα δεδομένα που του έχουν δοθεί ως είσοδος.
- Ιδιαίτερα ικανοποιητικά είναι τα αποτελέσματα σχετικά με την ποικιλία της λίστας συστάσεων.
- Καλή απόδοση σε διαφορετικούς τύπους κειμένων και ιδιαίτερα, όταν τα κείμενα είναι μικρά σε μέγεθος.

Συμπεράσματα

- Το σύστημα είναι σε θέση να καλύψει επαρκώς τις αναγνωστικές ανάγκες του χρήστη με βάση τα δεδομένα που του έχουν δοθεί ως είσοδος.
- Ιδιαίτερα ικανοποιητικά είναι τα αποτελέσματα σχετικά με την ποικιλία της λίστας συστάσεων.
- Καλή απόδοση σε διαφορετικούς τύπους κειμένων και ιδιαίτερα, όταν τα κείμενα είναι μικρά σε μέγεθος.

Συμπεράσματα

- Το σύστημα είναι σε θέση να καλύψει επαρκώς τις αναγνωστικές ανάγκες του χρήστη με βάση τα δεδομένα που του έχουν δοθεί ως είσοδος.
- Ιδιαίτερα ικανοποιητικά είναι τα αποτελέσματα σχετικά με την ποικιλία της λίστας συστάσεων.
- Καλή απόδοση σε διαφορετικούς τύπους κειμένων και ιδιαίτερα, όταν τα κείμενα είναι μικρά σε μέγεθος.

Μελλοντικές Επεκτάσεις

- **Συστάσεις σε ομάδες χρηστών:** Συνδυασμός πολλαπλών προφίλ (διανυσμάτων) για τη δημιουργία ενός κοινού προφίλ για μία ομάδα χρηστών.
- **Ανατροφοδότηση/Εκπαίδευση συστήματος με την αξιολόγηση του χρήστη:**
  - Αποθήκευση προτεινόμενων άρθρων που βρέθηκαν ενδιαφέροντα στη ΒΔ.
  - Είσοδος αξιολόγησης στο σύστημα.
  - Δημιουργία σύναψης μεταξύ συγκεκριμένων άρθρων μέσα σε κάθε cluster και θεμάτων που διαφαίνονται από τις επιλογές ενός χρήστη, προκειμένου να λάβουμε βελτιωμένα αποτελέσματα σε επόμενη σύσταση.
- **Εφαρμογή συνάρτησης αξιολόγησης** για την επιλογή ν-άδων σύστασης που προσφέρουν τη μέγιστη αύξηση ωφέλειας για το χρήστη / Εφαρμογή submodular μοντέλου συστάσεων.

Ευχαριστώ!