



ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΗΛΕΚΤΡΟΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ ΚΑΙ
ΠΛΗΡΟΦΟΡΙΚΗΣ

Σύστημα Επεξεργασίας,
Ανάλυσης και Ομαδοποίησης
Εγγράφων Ειδήσεων του Διαδικτύου

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ
της
ΑΦΡΟΔΙΤΗΣ ΑΛΕΒΙΖΟΠΟΥΛΟΥ

Επιβλέπων: Ιωάννης Χατζηλυγερούδης

Πάτρα, Νοέμβριος 2017



ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΗΛΕΚΤΡΟΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

Σύστημα Επεξεργασίας,
Ανάλυσης και Ομαδοποίησης
Εγγράφων Ειδήσεων του Διαδικτύου

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΑΦΡΟΔΙΤΗΣ ΑΛΕΒΙΖΟΠΟΥΛΟΥ

Επιβλέπων: Ιωάννης Χατζηλυγερούδης

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 16η Νοεμβρίου 2017.

(Υπογραφή)

(Υπογραφή)

(Υπογραφή)

.....

.....

.....

Ιωάννης Χατζηλυγερούδης

Καθηγητής

Πάτρα, Νοέμβριος 2017

Ευχαριστίες

Στο σημείο αυτό είναι απαραίτητο να ευχαριστήσω όλους όσους συνέβαλλαν στην εκπόνηση και συγγραφή της διπλωματικής μου εργασίας.

Ευχαριστώ θερμά τον Καθηγητή κ. Ιωάννη Χατζηλυγερούδη για την εμπιστοσύνη που μου έδειξε αναθέτοντάς μου την εκπόνηση της συγκεκριμένης διπλωματικής.

Θα ήθελα επίσης να ευχαριστήσω τον επιβλέποντα της διπλωματικής μου εργασίας, Διδάκτορα κ. Ισίδωρο Περίκο, για τη συνεχή καθοδήγησή του, τις χρήσιμες συμβουλές του, αλλά και την εξαιρετική συνεργασία που είχαμε.

Τέλος, ευχαριστώ την οικογένειά μου και τον αγαπημένο μου φίλο, Αλέξανδρο Αντωνόπουλο, για τη στήριξη και την καθοδήγηση που μου προσέφεραν καθ'όλη τη διάρκεια των σπουδών μου.

*Στη μνήμη της μητέρας μου, Χριστίνας-Λώρεν Αλεβιζοπούλου
και του αγαπημένου δασκάλου και φίλου, Νώντα Καρύμπαλη*

Περίληψη

Λόγω του μεγάλου όγκου πληροφοριών που κατακλύζει το διαδίκτυο, συχνά οι χρήστες δυσκολεύονται να ξεχωρίσουν τις πληροφορίες που πραγματικά σχετίζονται με τα ενδιαφέροντά τους. Επιπλέον, οι χρήστες έχουν πολύ διαφορετικά ενδιαφέροντα ή προτιμήσεις που μπορούν να ληφθούν υπόψη ώστε να φιλτραριστούν ή να ταξινομηθούν τα αποτελέσματα μιας ερώτησης με σκοπό το αποτέλεσμα να ικανοποιεί τις εξατομικευμένες ανάγκες κάθε χρήστη. Η κατηγορία αυτών των συστημάτων εξατομικεύσης ονομάζεται “Συστήματα Συστάσεων” (Recommender Systems). Τα Συστήματα Συστάσεων εκμεταλλεύονται τις ιδιαιτερότητες των χρηστών με σκοπό να τους διευκολύνουν στο να προσδιορίζουν ακριβέστερα τις πληροφορίες ή τις υπηρεσίες για τις οποίες ενδιαφέρονται περισσότερο ή σχετίζονται με τις ανάγκες τους, κάνοντας χρήση ειδικών αλγορίθμων. Οι αλγόριθμοι που χρησιμοποιούνται λαμβάνουν ως είσοδο τα χαρακτηριστικά και τις προτιμήσεις των χρηστών ή τις σχέσεις μεταξύ των χρηστών ή τα γνωρίσματα των προς σύσταση αντικειμένων και υπολογίζουν το εκτιμώμενο ενδιαφέρον του χρήστη για κάθε αντικείμενο. Στην συνέχεια ταξινομούν ή φιλτράρουν τα αντικείμενα με κριτήριο το εκτιμώμενο ενδιαφέρον.

Στο πλαίσιο της παρούσας διπλωματικής εργασίας σχεδιάστηκε και υλοποιήθηκε ένα σύστημα επεξεργασίας, ανάλυσης και ομαδοποίησης εγγράφων ειδήσεων του διαδικτύου, σχεδιασμένο ως εφαρμογή ειδησεογραφικού περιεχομένου που επιτρέπει στο χρήστη την περιήγηση μεταξύ των άρθρων μιας βάσης δεδομένων και λαμβάνοντας υπόψη τις επιλογές του, δηλαδή το προφίλ/ιστορικό του κάθε χρήστη, του προτείνει νέα άρθρα ειδήσεων που ταιριάζουν περισσότερο με τα ενδιαφέροντά του και δίνει τη δυνατότητα παρουσίασης των ειδήσεων οι οποίες σχετίζονται σημασιολογικά με αυτές που έχει ήδη αναγνώσει.

Λέξεις Κλειδιά

Μηχανική Εκμάθηση, Πολυεπίπεδο Σύστημα Συστάσεων Διαδικτυακών Εγγράφων Ειδήσεων, Προσωποποιημένη Ανάκτηση Περιεχομένου, Προσωποποιημένα Διαδικτυακά Έγγραφα Ειδήσεων, Εξατομίκευση, Προφίλ Χρήστη, Μοντέλα Θεμάτων

Περιεχόμενα

Ευχαριστίες	i
Περίληψη	v
Περιεχόμενα	viii
1 Εισαγωγή	1
1.1 Σκοπός και Διάρθρωση της Διπλωματικής Εργασίας	2
1.2 Ορισμός της Τεχνητής Νοημοσύνης	3
1.3 Ιστορία, Εξέλιξη και Τομείς εφαρμογών Τεχνητής Νοημοσύνης	6
1.4 Συμβολή της Τεχνητής Νοημοσύνης στον Ειδησεογραφικό Τομέα	9
1.5 Συστήματα Συστάσεων	10
1.5.1 Κατηγορίες Συστημάτων Συστάσεων	11
1.5.2 Προβλήματα Τεχνικών Συστάσεων	12
1.6 Επεξεργασία Φυσικής Γλώσσας	12
1.6.1 Πεδία έρευνας Επεξεργασίας Φυσικής Γλώσσας	13
1.6.2 Ανάλυση Φυσικής Γλώσσας με χρήση Θεματικών Μοντέλων	14
2 Ανάλυση και Σχεδίαση Συστήματος	17
2.1 Περιγραφή Συστήματος	17
2.2 Αρχιτεκτονική Συστήματος	18
2.3 Υποσυστήματα	20
2.3.1 Υποσύστημα Δημιουργίας Βάσης Δεδομένων και Καταχώρησης Πληροφορίας	21
2.3.2 Υποσύστημα Ομαδοποίησης Ειδησεογραφικών Άρθρων	21
2.3.3 Υποσύστημα Δημιουργίας Προφίλ Χρήστη	28
2.3.4 Υποσύστημα Προσωποποιημένων Συστάσεων	31
3 Τεχνολογίες Υλοποίησης	39
3.1 Η γλώσσα προγραμματισμού Python	39

3.2	MySQL	40
3.3	HTML	40
3.4	CSS	41
3.5	Apache HTTP Server	41
3.6	Flask Web Framework	42
3.7	Εργαλεία Επεξεργασίας Φυσικής Γλώσσας	42
3.7.1	Tree Tagger	42
3.7.2	NLTK	42
3.7.3	GATE for Named Entity Recognition	45
4	PELOMA: A Personalized News Recommendation System	47
5	Αξιολόγηση Συστήματος	55
6	Συμπεράσματα και Μελλοντικές Επεκτάσεις	63
6.1	Συμπεράσματα	63
6.2	Μελλοντικές Επεκτάσεις	65

Κεφάλαιο 1

Εισαγωγή

Η ραγδαία εξέλιξη του Παγκόσμιου Ιστού τα τελευταία χρόνια κατέστησε το διαδίκτυο προσβάσιμο σε εκατομμύρια χρήστες, επιτρέποντας με αυτό τον τρόπο όχι μόνο να έχουν πρόσβαση σε περισσότερη πληροφορία παγκόσμιων ειδήσεων, αλλά να μπορούν να ενημερώνονται και πιο γρήγορα. Υπάρχουν αρκετοί παράγοντες που συνετέλεσαν στην επιτυχία αυτή του διαδικτύου, όπως, για παράδειγμα, το μειωμένο κόστος της διανομής και της πρόσβασης στις ειδήσεις, η διαθεσιμότητα του διαδικτύου σε μια πληθώρα από πλατφόρμες περιγητών, η παγκόσμια αποστολή και κατανάλωση πληροφορίας, ο μικρός χρόνος για τη δημοσίευση ειδήσεων κλπ.

Λόγω της μεγάλης ποσότητας ειδήσεων που δημοσιεύονται κάθε μέρα, είναι δύσκολο να βρει κανείς νέες ειδήσεις που τον ενδιαφέρουν. Μια λύση σε αυτό το πρόβλημα είναι τα “Συστήματα Συστάσεων” (Recommender Systems), τα οποία χρησιμοποιούν αλγόριθμους συσχέτισης, οι οποίοι, με βάση διαφόρων ειδών πληροφορίες για τους χρήστες, επιλέγουν και προτείνουν σε αυτούς συγκεκριμένα προϊόντα ή υπηρεσίες, ανάλογα με τις προτιμήσεις τους.

1.1 Σκοπός και Διάρθρωση της Διπλωματικής Εργασίας

Η παρούσα διπλωματική εργασία επικεντρώνεται στο πεδίο των Συστημάτων Συστάσεων και στο πλαίσιο της σχεδιάστηκε και υλοποιήθηκε ένα σύστημα συστάσεων ειδησεογραφικού περιεχομένου που προτείνει άρθρα ειδήσεων σε μεμονωμένους χρήστες.

Η εργασία είναι οργανωμένη σε έξι κεφάλαια:

Στο κεφάλαιο 1 παρουσιάζονται κάποιες θεωρητικές έννοιες που σχετίζονται με την Τεχνητή Νοημοσύνη (TN), εισάγεται ο ορισμός της TN όπως έχει διαμορφωθεί μέχρι σήμερα και εξηγείται λεπτομερώς τι είναι η TN. Στη συνέχεια, δίνονται ιστορικά στοιχεία που αφορούν στη δημιουργία και εξέλιξη της μέσα στο χρόνο, καθώς, επίσης, αναλύονται τομείς στους οποίους αυτή χρησιμοποιείται, δίνοντας έμφαση στο πώς έχει συμβάλει η TN στον τομέα της ειδησεογραφικής ενημέρωσης. Επιπρόσθετα, παραθέτουμε τον ορισμό των συστημάτων συστάσεων, αναλύουμε τις κατηγορίες στις οποίες χωρίζονται και παραθέτουμε τα σημαντικότερα προβλήματα που αντιμετωπίζουν οι τεχνικές συστάσεων. Τέλος, μελετάμε τον ορισμό της Επεξεργασίας Φυσικής Γλώσσας, παρουσιάζουμε τα σημαντικότερα επιστημονικά πεδία έρευνας πάνω σε αυτή και δίνουμε το θεωρητικό υπόβαθρο της Ανάλυσης Φυσικής Γλώσσας με χρήση Μοντέλων Θεμάτων.

Στο κεφάλαιο 2 γίνεται παρουσίαση του Συστήματος Επεξεργασίας, Ανάλυσης και Ομαδοποίησης Εγγράφων Ειδήσεων του Διαδικτύου που αναπτύξαμε, σχεδιασμένο ως ένα σύστημα συστάσεων, αναλύεται ο σκοπός του, περιγράφεται η αρχιτεκτονική του, παρουσιάζοντας και αναλύοντας τα υποσυστήματά του. Τέλος, αναλύεται ο αλγόριθμος που εκτελεί για να παράγει αποτελέσματα.

Στο κεφάλαιο 3 αναφερόμαστε διεξοδικά στα προγραμματιστικά εργαλεία και τις τεχνολογίες που χρησιμοποιήθηκαν τόσο για την επεξεργασία φυσικής γλώσσας όσο και για την υλοποίηση της εφαρμογής.

Στο κεφάλαιο 4 παρουσιάζουμε την διεπαφή χρήστη καθώς και μερικά παραδείγματα χρήσης της εφαρμογής.

Στο κεφάλαιο 5 περιλαμβάνεται η πειραματική εφαρμογή του συστήματός μας. Παρουσιάζονται εκτενώς τα αποτελέσματα των πειραμάτων μας καθώς και η αξιολόγηση του συστήματος.

Τέλος, στο κεφάλαιο 6 δίνεται η συνεισφορά αυτής της διπλωματικής εργασίας μέσω των συμπερασμάτων που προκύπτουν, καθώς και κατευθύνσεις για πιθανές μελλοντικές επεκτάσεις.

1.2 Ορισμός της Τεχνητής Νοημοσύνης

‘Ονομάζουμε το είδος μας *homo sapiens* - άνθρωπος ο σοφός - επειδή οι νοητικές μας ικανότητες είναι πολύ σημαντικές για μας. Για χιλιάδες χρόνια προσπαθούμε να κατανοήσουμε το πώς σκεπτόμαστε, δηλαδή, πώς μια χούφτα ύλης μπορεί να αντιλαμβάνεται, να κατανοεί, να προβλέπει και να χειρίζεται έναν κόσμο πολύ μεγαλύτερο και πολύ πιο πολύπλοκο από τον εαυτό της. Το πεδίο της Τεχνητής Νοημοσύνης πηγαίνει ακόμα πιο πέρα: Επιχειρεί όχι μόνο να κατανοήσει αλλά και να κατασκευάσει νοήμονες οντότητες.’

— Russell and Norvig, *Artificial Intelligence: A modern approach* [19]

Για να μπορέσουμε να αντιληφθούμε καλύτερα το επιστημονικό πεδίο της Τεχνητής Νοημοσύνης, θα ήταν χρήσιμο να προσεγγίσουμε αρχικά την έννοια της ανθρώπινης νοημοσύνης.

Ο Howard Gardner στο βιβλίο του “Frames of mind: The theory of multiple Intelligences” (1983), διακρίνει σε κάθε άνθρωπο οχτώ διαφορετικούς τύπους νοημοσύνης (Γλωσσική, Λογική/Μαθηματική, Μουσική, Χωρική, Σωματική, Διαπροσωπική, Ενδοπροσωπική, Φυσιοκρατική), όπου αν και είναι ευδιάκριτοι μέσα στον ανθρώπινο εγκέφαλο, στην πραγματικότητα χρησιμοποιείται ένα μίγμα από αυτούς.

Επίσης, ο Douglas Hofstadter προτείνει ότι νοημοσύνη είναι να ανταποκρίνεται σε καταστάσεις με ελαστικότητα (αποφυγή μηχανικής συμπεριφοράς), να μπορείς να κατανοείς τα ασαφή ή αντιφατικά μηνύματα από τα συμφραζόμενα, να μπορείς να αναγνωρίζεις και να ιεραρχείς τις καταστάσεις με βάση τη σπουδαιότητά τους, ενώ υποστηρίζει ότι η ανθρώπινη νοημοσύνη θα πρέπει να χαρακτηρίζεται από την ικανότητα να βρίσκεις ομοιότητες μεταξύ καταστάσεων που μοιάζουν διαφορετικές, αλλά και να μπορείς να βρίσκεις διαφορές σε καταστάσεις που φαίνονται παρόμοιες.

Η δοκιμασία Turing (Turing test), η οποία προτάθηκε από τον Alan Turing (1950), σχεδιάστηκε για να παρέχει έναν ικανοποιητικό λειτουργικό ορισμό της νοημοσύνης. Αντί να προτείνει μια εκτεταμένη και ενδεχομένως αντιφατική λίστα γνωρισμάτων που απαιτούνται για τη νοημοσύνη, ο Turing πρότεινε μια δοκιμασία που βασιζόταν στην αδυναμία να γίνει διάκριση από τις αναμφίβολα νοήμονες οντότητες — τους ανθρώπους.

Ο φημισμένος Άγγλος μαθηματικός έθετε το ερώτημα: «Μπορούν οι μηχανές να σκεφτούν;». Ο υπολογιστής περνά τη δοκιμασία αν ένας άνθρωπος εξεταστής, αφού θέσει μερικές γραπτές ερωτήσεις, δεν μπορεί να συμπεράνει αν οι γραπτές απαντήσεις προέρχονται από άνθρωπο ή όχι. Τότε η μηχανή χαρακτηρίζεται ως ευφυής. [19]

Μετά από αρκετούς ορισμούς και φιλοσοφικές προσεγγίσεις της Τεχνητής Νοημοσύνης, οι επιστήμονες φαίνεται να έχουν κατασταλάξει στον παρακάτω γενικό ορισμό:

“Τεχνητή Νοημοσύνη (TN) είναι ο τομέας της επιστήμης των υπολογιστών που ασχολείται με τη σχεδίαση και την υλοποίηση ευφύων (νοημόνων) υπολογιστικών συστημάτων τα οποία είναι ικανά να μιμηθούν τις ανθρώπινες γνωστικές ικανότητες, εμφανίζοντας χαρακτηριστικά που αποδίδουμε συνήθως σε ανθρώπινη συμπεριφορά, όπως η επίλυση προβλημάτων, η αντίληψη μέσω της όρασης, η μάθηση, η εξαγωγή συμπερασμάτων, η κατανόηση της φυσικής γλώσσας κλπ”.

Η TN αποτελεί σημείο τομής μεταξύ πολλαπλών επιστημών όπως της πληροφορικής, της ψυχολογίας, της φιλοσοφίας, της νευρολογίας, της γλωσσολογίας και της επιστήμης μηχανικών, με στόχο τη σύνθεση ευφυούς συμπεριφοράς, με στοιχεία συλλογιστικής, μάθησης, προσαρμογής στο περιβάλλον, εξαγωγής συμπερασμάτων, κατανόησης από συμφραζόμενα και επίλυσης προβλημάτων, ενώ συνήθως εφαρμόζεται σε μηχανές ή υπολογιστές ειδικής κατασκευής.

Παρακάτω παρουσιάζονται οι τέσσερις μεγάλες κατηγορίες στις οποίες ταξινομούνται οι ορισμοί της TN. Ιστορικά, έχουν ακολουθηθεί και οι τέσσερις προσεγγίσεις. Όπως θα περίμενε κανείς, υπάρχει κάποια διένεξη ανάμεσα στις προσεγγίσεις που εστιάζονται στον άνθρωπο και τις προσεγγίσεις που εστιάζονται στην ορθολογικότητα. Μια ανθρωποκεντρική προσέγγιση θα πρέπει να είναι εμπειρική επιστήμη, με υποθέσεις και με πειραματική επιβεβαίωση. Μια ορθολογιστική (rationalist) προσέγγιση περιλαμβάνει ένα συνδυασμό μαθηματικών και τεχνολογίας. [19]

ΣΥΣΤΗΜΑΤΑ ΠΟΥ ΣΚΕΦΤΟΝΤΑΙ ΣΑΝ ΤΟΝ ΑΝΘΡΩΠΟ	ΣΥΣΤΗΜΑΤΑ ΠΟΥ ΣΚΕ- ΦΤΟΝΤΑΙ ΟΡΘΟΛΟΓΙΚΑ
<p>“Η συναρπαστική νέα προσπάθεια για να κάνουμε τους υπολογιστές να σκέπτονται... μηχανές με νόηση, με την πλήρη και κυριολεκτική έννοια.” (Haugeland, 1985)</p> <p>“Η αυτοματοποίηση των δραστηριοτήτων που συσχετίζουμε με την ανθρώπινη σκέψη, όπως η λήψη αποφάσεων, η επίλυση προβλημάτων, η μάθηση...” (Bellman, 1978)</p>	<p>“Η μελέτη των νοητικών ικανοτήτων με τη χρήση υπολογιστικών μοντέλων” (Craniak και McDermott, 1985)</p> <p>“Η μελέτη των υπολογιστικών εργασιών που μας δίνουν τη δυνατότητα να αντιλαμβανόμαστε, να συλλογίζομαστε και να ενεργούμε” (Winston, 1992)</p>
ΣΥΣΤΗΜΑΤΑ ΠΟΥ ΕΝΕΡΓΟΥΝ ΣΑΝ ΤΟΝ ΑΝΘΡΩΠΟ	ΣΥΣΤΗΜΑΤΑ ΠΟΥ ΕΝΕΡ- ΓΟΥΝ ΟΡΘΟΛΟΓΙΚΑ
<p>“Η τέχνη της δημιουργίας μηχανών που πραγματοποιούν λειτουργίες οι οποίες απαιτούν νοημοσύνη όταν πραγματοποιούνται από ανθρώπους” (Kuzweil, 1990)</p> <p>“Η μελέτη του πώς μπορούμε να κάνουμε τους υπολογιστές να κάνουν πράγματα στα οποία, προς το παρόν, οι άνθρωποι είναι καλύτεροι” (Rich και Knight, 1991)</p>	<p>“Υπολογιστική νοημοσύνη είναι η μελέτη της σχεδίασης ευφυών πρακτόρων” (Poole, 1998)</p> <p>“Η τεχνητή νοημοσύνη ασχολείται με την ευφυή συμπεριφορά των τεχνουργημάτων” (Nilsson, 1998)</p>

Πίνακας 1.1: Ορισμοί Τεχνητής Νοημοσύνης

1.3 Ιστορία, Εξέλιξη και Τομείς εφαρμογών Τεχνητής Νοημοσύνης

Η Τεχνητή Νοημοσύνη (TN) έχει ήδη συμπληρώσει περισσότερο από μισό αιώνα ύπαρξης. Τυπικά ξεκίνησε το 1956 με τη συνάντηση επιφανών επιστημόνων, όπως οι John McCarthy, Marvin Minsky και Claude Shannon. Η ιστορία και η εξέλιξη της TN φαίνεται να έχει επηρεαστεί από διάφορους επιστημονικούς κλάδους, όπως τα Μαθηματικά, η Φιλοσοφία και η Ιατρική/Νευροεπιστήμες.

Οι αρχές της TN τέθηκαν στην Αρχαία Ελλάδα όταν ο Αριστοτέλης διατύπωσε πρώτος ένα ακριβές σύνολο νόμων που διέπουν το ορθολογικό μέρος της νόησης. Σε μία προσπάθεια να οριοθετήσει την “ορθή σκέψη”, δηλαδή τη διαδικασία συλλογισμού, ανέπτυξε ένα άτυπο σύστημα συλλογισμών για τη σωστή συλλογιστική, οι οποίοι θεωρητικά επέτρεπαν να παράγει κανείς συμπεράσματα μηχανικά με δεδομένες κάποιες αρχικές υποθέσεις.

Αν και η ιδέα της τυπικής λογικής συναντάται από τον καιρό των αρχαίων Ελλήνων φιλοσόφων, η μαθηματική της ανάπτυξη ξεκίνησε ουσιαστικά τον 19ο αιώνα με την εργασία του George Boole, ο οποίος επεξεργάστηκε τις λεπτομέρειες της προτασιακής λογικής ή λογικής Boole. Το 1879, ο Gottlob Frege επέκτεινε τη λογική του Boole ώστε να συμπεριλάβει αντικείμενα και σχέσεις, θέτοντας τις βάσεις του κατηγορηματικού λογισμού. Η μαθηματική επιστήμη προκαλείται να δώσει λύσεις για έννοιες όπως η λογική, ο υπολογισμός και η πιθανότητα. Ο Thomas Bayes πρότεινε ένα κανόνα για τον υπολογισμό της πιθανότητας, ο οποίος ονομάστηκε “ανάλυση Bayes” και αποτελεί μέχρι σήμερα την βάση των συστημάτων αβέβαιης λογικής.

Μια σημαντική πρώτη εργασία πάνω στο θέμα δημιουργήθηκε από τους Warren McCulloch και Walter Pitts. Αυτή είχε σαν στόχο να συσχετίζει βιολογικούς νευρώνες του εγκεφάλου με απλά υπολογιστικά στοιχεία. Το συμπέρασμα αυτού ήταν μια πρόταση για επικοινωνία μεταξύ βιολογικών νευρώνων και υπολογιστικών στοιχείων που συνέθεταν ένα νευρωνικό δίκτυο με την ικανότητα να μαθαίνει και να κάνει υπολογισμούς. Το πρώτο νευρωνικό δίκτυο δημιουργήθηκε το 1951 από δυο φοιτητές του μαθηματικού τμήματος του Princeton, Marvin Minsky και Dean Edmonds. Αυτό ονομάστηκε SNARC και αποτελούνταν από 40 νευρώνες, 3000 λυχνίες και άλλα ηλεκτρονικά εξαρτήματα.

Στο συνέδριο που διοργανώθηκε το 1956 στο Dartmouth της Μασαχουσέτης από τους John McCarthy, Marvin Minsky, Claude Shannon και Nathaniel Rochester πα-

ρουσιάστηκε από τους ερευνητές Allen Newell και Herbert Simon το Logic Theorist, ένα πρόγραμμα συλλογισμού το οποίο μπορούσε να αποδεικνύει τα περισσότερα από τα θεωρήματα των Russell και Whitehead. Στο ίδιο συνέδριο προτάθηκε και από τον McCarthy το όνομα Τεχνητή Νοημοσύνη, το οποίο έγινε αποδεκτό.

Το 1958 δημιουργήθηκε η συναρτησιακή γλώσσα Lisp. Η γλώσσα αυτή έγινε ταυτόσημο της TN για αρκετά μεγάλο χρονικό διάστημα. Την ίδια χρονιά ο McCarthy πρότεινε το Advice Taker, ένα πρόγραμμα το οποίο χρησιμοποιούσε γνώση για την επίλυση καθημερινών προβλημάτων. Η υλοποίησή του έγινε το 1963 από τον ίδιο. Λίγο αργότερα, στο Stanford Research Institute δημιουργήθηκε το πρώτο ρομπότ με το όνομα Shakey.

Η επόμενη δεκαετία δεν ήταν τόσο ελπιδοφόρα όσο η πρώτη. Η TN κατηγορήθηκε ως ένα μέσο επίλυσης απλών προβλημάτων (toy problems). Έτσι, το 1973 η Βρετανική κυβέρνηση διέκοψε την υποστήριξη της έρευνας στην TN.

Την δεκαετία που ακολούθησε αναπτύχθηκαν “έμπειρα συστήματα”, τα οποία περιείχαν αρκετή γνώση του προβλήματος που αντιμετώπιζαν. Την ίδια δεκαετία αναπτύχθηκε η γλώσσα προγραμματισμού Prolog, η οποία βασιζόταν στην λογική.

Ένα από τα συστήματα που αναπτύχθηκαν κατά την πορεία της TN μέσα στις δεκαετίες ήταν και το DENDRAL. Αυτό περιείχε σημαντική ποσότητα γνώσης η οποία εκφραζόταν με την μορφή κανόνων. Ένα άλλο παρόμοιο σύστημα ήταν το MYCIN, το οποίο περιελάμβανε 450 κανόνες και ο σκοπός του ήταν η διάγνωση μολύνσεων στο αίμα. Η γνώση του συγκεκριμένου συστήματος δεν προέκυψε από κάποιο μοντέλο όπως στο DENDRAL, αλλά από συνεντεύξεις σε γιατρούς. Αξίζει να σημειωθεί ότι το σύστημα αυτό εισήγαγε την έννοια της αβεβαιότητας. Ένα ακόμα αντιπροσωπευτικό σύστημα ήταν και το PROSPECTOR, στον τομέα της γεωλογίας, αφού έδινε πληροφορίες για τοποθεσίες εξόρυξης κοιτασμάτων. Στον τομέα της κατανόησης της φυσικής γλώσσας ήταν το SHRDLU, το οποίο περιοριζόταν σε προβλήματα μετακίνησης αντικειμένων και το LUNAR, το οποίο χρησιμοποιήθηκε σε πραγματικές εφαρμογές και δεχόταν ερωτήσεις για τα πετρώματα που έφερναν τα διαστημόπλοια APOLLO στην γη από τη σελήνη. Το πρώτο επιτυχημένο εμπορικό σύστημα ήταν το R1/XCON, του οποίου η χρήση ήταν η διαμόρφωση των παραγγελιών της εταιρίας Digital Equipments Corporation με βάση τις ανάγκες των πελατών και τα διαθέσιμα αποθέματα εξαρτημάτων. Κατά τα μέσα της δεκαετίας του '80, τα νευρωνικά δίκτυα ήρθαν πάλι στο προσκήνιο.

Αν τώρα θελήσουμε να δούμε την εξέλιξη της TN χρονολογικά, θα δούμε ότι πολλοί

συγγραφείς διακρίνουν τέσσερις περιόδους: την προϊστορική, την κλασική, τη ρομαντική και την μοντέρνα. Στην προϊστορική περίοδο, η TN συναντάται σαν αντικείμενο μόνο σε διηγήματα επιστημονικής φαντασίας. Την κλασική περίοδο, στην οποία τα συστήματα που αναπτύχθηκαν έλυναν γρίφους και έπαιζαν παιχνίδια. Κατά την ρομαντική περίοδο, η TN επικεντρώνεται κυρίως στην ανάπτυξη συστημάτων για κατανόηση ιστοριών και διαλόγων σε φυσική γλώσσα. Η μοντέρνα περίοδος βασίζεται στην δημιουργία συστημάτων που έχουν ως σκοπό την εμπορική εκμετάλλευση των αποτελεσμάτων της έρευνας της TN.

Αυτή την εποχή βιώνουμε τη μετα-μοντέρνα περίοδο, όπου η TN καλείται να παίξει σημαντικό ρόλο σε ένα πληροφοριακό περιβάλλον όπου τα κύρια χαρακτηριστικά του είναι η εξάπλωση του διαδικτύου και η διείσδυση των υπολογιστικών συστημάτων σε κάθε είδους συσκευές ευρείας και καθημερινής χρήσης.

Τα τελευταία χρόνια, η ανάπτυξη της TN είναι αλματώδης και έτσι σήμερα υπάρχουν εφαρμογές της όπως η ρομποτική (συστήματα ρομποτικής χειρουργικής), η μηχανική όραση και η μηχανική μάθηση. Υπάρχουν πολλά ευφυή προγράμματα τα οποία βοηθούν τον χρήστη να αναζητά πληροφορίες στο διαδίκτυο, να στέλνει email και άλλα. Άλλη εφαρμογή είναι τα συστήματα αναγνώρισης φωνής, όπως το σύστημα PEGASUS που κάνει αεροπορικές κρατήσεις και προτείνει τις βέλτιστες πτήσεις για κάθε πελάτη. Υπάρχουν, επίσης, έμπειρα συστήματα πραγματικού χρόνου τα οποία επεξεργάζονται δεδομένα τα οποία στέλνονται από διαστημόπλοια. Τέλος, υπάρχουν ευφυή συστήματα τα οποία οδηγούν οχήματα σε πραγματικές συνθήκες χρησιμοποιώντας κάμερες και αποστασιόμετρα.

Οι νευροεπιστήμες που ασχολούνται με την μελέτη του νευρικού συστήματος και ιδιαίτερα του εγκεφάλου, καθώς και η ιατρική, αποτελούν επιστήμες που τα τελευταία χρόνια χρησιμοποιούν συστήματα TN. Οι περιοχές όπου βρίσκουν εφαρμογή τέτοια συστήματα είναι τα πεδία της πρόληψης, της διάγνωσης και της θεραπείας διαφόρων ασθενειών. Χρησιμοποιώντας κυρίως ασαφή συστήματα, η TN προκαλείται να απαντήσει σε σύγχρονα προβλήματα, τα οποία σχετίζονται με την υγεία, το καλύτερο βιωτικό επίπεδο και της βέλτιστη θεραπευτική επιλογή.

Άλλες επιστήμες στις οποίες η TN βρίσκει εφαρμογή αλλά και επηρεάστηκε από αυτές είναι η ψυχολογία, η πληροφορική και η γλωσσολογία.

Τα υπολογιστικά συστήματα πλέον εξελίσσονται με αλματώδη ρυθμό. Έτσι, λοιπόν, δημιουργούνται νέες απαιτήσεις για την επίλυση προβλημάτων. Η TN ανεβάζει διαρκώς τον πήχη και στοχεύει στη δημιουργία συστημάτων τα οποία να εξαρτώνται ελάχιστα

από τον κατασκευαστή τους και περισσότερο από την ικανότητά τους να μαθαίνουν πώς να συμπεριφέρονται σε αλληλεπίδραση με το περιβάλλον τους.

1.4 Συμβολή της Τεχνητής Νοημοσύνης στον Ειδησεογραφικό Τομέα

Το διαδίκτυο είναι μία από τις δημοφιλέστερες πλατφόρμες για τη διανομή και την ανάγνωση ειδήσεων. Υπάρχουν αρκετοί παράγοντες που συνετέλεσαν στην επιτυχία αυτή του διαδικτύου, όπως το μειωμένο κόστος της διανομής και της πρόσβασης στις ειδήσεις, η διαθεσιμότητα του διαδικτύου σε μια πληθώρα από πλατφόρμες περιηγητών, η παγκόσμια αποστολή και κατανάλωση πληροφορίας, ο μικρός χρόνος για τη δημοσίευση ειδήσεων κλπ.

Δυστυχώς, η επιτυχία του διαδικτύου είναι και ένα από τα μεγαλύτερα προβλήματά του: η μεγάλη ποσότητα καθημερινά δημοσιευμένων ειδήσεων κάνει δύσκολη τη διαδικασία εύρεσης αυτών που αντιστοιχούν σε συγκεκριμένα ενδιαφέροντα. Διαδικτυακές εφημερίδες παρουσιάζουν τα τελευταία νέα στους ιστοτόπους τους σε πραγματικό χρόνο, και οι χρήστες μπορούν να λαμβάνουν αυτόματες ειδοποιήσεις για αυτά μέσω τροφοδοσιών RSS. Το πρότυπο RSS προέρχεται από το αγγλικό Really Simple Syndication και είναι βασισμένο στη γλώσσα XML για την διανομή των ενημερώσεων του περιεχομένου του ιστού. Τα RSS feeds προσφέρουν τη δυνατότητα στους χρήστες να λαμβάνουν νέες πληροφορίες από διάφορες ιστοσελίδες τη στιγμή που δημοσιεύονται χωρίς να χρειάζεται να τις επισκεφθούν. Το RSS είναι, δηλαδή, ένας νέος τρόπος ενημέρωσης για νέα, εξελίξεις και γεγονότα.

Είναι γεγονός πως το διαδίκτυο αποτελείται πλέον από δισεκατομμύρια σελίδες οι οποίες περιέχουν τέτοιο πλούτο πληροφοριών που είναι σχεδόν αδύνατο για τον οποιονδήποτε να μπορεί να παρακολουθεί διαρκώς ό,τι νεότερο συμβαίνει στον κόσμο ή στο αντικείμενο που τον ενδιαφέρει. Στο πρόβλημα αυτό ήρθαν να δώσουν τη λύση τα RSS feeds. Με τα feeds ο χρήστης μπορεί να βλέπει τότε ανανεώθηκε το περιεχόμενο των δικτυακών τόπων που τον ενδιαφέρουν, λαμβάνοντας κατευθείαν στον υπολογιστή του τους τίτλους των τελευταίων ειδήσεων και των άρθρων (ή ακόμα και εικόνων ή βίντεο) αμέσως μόλις αυτά γίνουν διαθέσιμα χωρίς να είναι απαραίτητο να επισκέπτεται καθημερινά τους αντίστοιχους δικτυακούς τόπους. Χρησιμοποιώντας αυτό το φορμάτ, οι υπεύθυνοι του ιστού δημιουργούν επικεφαλίδες και καινούργιο περιεχόμενο με συντεταγμένο τρόπο(τροφοδοσία) και οι χρήστες μπορούν να χρησιμοποιήσουν αναγνώστες feeds και συνανθροιστές ειδήσεων για να συλλέξουν και να παρακολουθήσουν τις αγαπημένες τους τροφοδοσίες από ένα κεντρικό σημείο. Τα feeds γίνονται όλο και πιο δημοφιλή. Όμως, η διαρκής ροή ειδήσεων από διάφορες πηγές ενημέρωσης δεν εξυπηρετεί το χρήστη, καθώς καθιστά αδύνατη την παρακολούθησή τους και κρίνεται

επιβεβλημένη η προσωποποίηση του αποτελέσματος. Μια πιθανή λύση στο πρόβλημα της υπερφόρτωσης πληροφοριών με ειδήσεις είναι η χρήση συστημάτων συστάσεων, ο σκοπός των οποίων είναι να προτείνουν αντικείμενα που προηγουμένως ήταν άγνωστα, στην περίπτωση μας ειδήσεις, και θα ενδιέφεραν ένα συγκεκριμένο χρήστη. Τυπικά, τέτοια συστήματα χρησιμοποιούν προφίλ χρηστών και έχουν στόχο την σύσταση ειδήσεων που ταιριάζουν καλύτερα στο προφίλ αυτό.

1.5 Συστήματα Συστάσεων

Τα συστήματα συστάσεων είναι εφαρμογές λογισμικού που παρέχουν εξατομικευμένες προστάσεις στους χρήστες σχετικά με προϊόντα ή υπηρεσίες που μπορεί να τους ενδιαφέρουν. Προτείνουν στοιχεία σχετικά με τα ενδιαφέροντα των χρηστών βάσει των προτιμήσεών τους, οι οποίες εκφράζονται είτε άμεσα (explicitly), είτε έμμεσα (implicitly). Τα συστήματα αυτά με τη βοήθεια ειδικών αλγορίθμων επιχειρούν να προβλέψουν ποιες υπηρεσίες είναι πιθανόν να ενδιαφέρουν περισσότερο τον χρήστη. Οι τεχνικές που χρησιμοποιούνται λαμβάνουν ως είσοδο τα χαρακτηριστικά και τις προτιμήσεις του χρήστη (προσωπικά στοιχεία, ιστορικό περιήγησης), τις σχέσεις μεταξύ των χρηστών και τα γνωρίσματα των προς σύσταση αντικειμένων, και υπολογίζουν το εκτιμώμενο ενδιαφέρον του χρήστη για κάθε αντικείμενο. Στη συνέχεια, φιλτράρουν ή ταξινομούν τα αντικείμενα με κριτήριο το εκτιμώμενο ενδιαφέρον. Τυπικά, τέτοια συστήματα χρησιμοποιούν προφίλ χρηστών και έχουν σαν στόχο τη σύσταση στοιχείων που ταιριάζουν περισσότερο στο προφίλ αυτό. Η δημιουργία προφίλ αποτελεί το πλέον σημαντικό στοιχείο ενός συστήματος συστάσεων. Τα συστήματα συστάσεων εκμεταλλεύονται τις ιδιαιτερότητες των χρηστών με σκοπό να διευκολύνουν στο να προσδιορίζουν ακριβέστερα τις πληροφορίες ή τα προϊόντα για τα οποία ενδιαφέρονται περισσότερο ή σχετίζονται με τις ανάγκες τους. Πιο συγκεκριμένα, ένα σύστημα συστάσεων μπορεί να κρατάει ιστορικό από τα άρθρα που έχει διαβάσει κάποιος χρήστης, οπότε την επόμενη φορά που θα επισκεφθεί τον ιστότοπο, το σύστημα θα του προτείνει νέα άρθρα σύμφωνα με την θεματολογία αυτών που είχε διαβάσει στο παρελθόν και πιθανόν να τον ενδιέφεραν.

Τα συστήματα συστάσεων αποτελούν ένα σημαντικό πεδίο με μεγάλο ενδιαφέρον σε ερευνητικό επίπεδο, από την εμφάνιση των πρώτων δημοσιεύσεων για το συνεργατικό φιλτράρισμα στα μέσα της δεκαετίας του '90. Υπάρχει μεγάλη δραστηριότητα τόσο στη βιομηχανία όσο και στην ακαδημαϊκή κοινότητα για την ανάπτυξη νέων προσεγγίσεων κατά την τελευταία δεκαετία.

Παρά το μεγάλο ενδιαφέρον των εταιρειών και το σημαντικό όγκο ερευνητικής δραστηριότητας για τα συστήματα συστάσεων, απαιτούνται περαιτέρω βελτιώσεις, οι οποίες περιλαμβάνουν καλύτερες μεθόδους αναπαράστασης του προφίλ των χρηστών και των

στοιχειών που προτείνονται, πιο εξελιγμένες μεθόδους δημιουργίας συστάσεων, ενσωμάτωση των διαφορών, βασισμένων στα συμφραζόμενα, πληροφοριών στη διαδικασία συστάσεων, και ανάπτυξη πιο ευέλικτων μεθόδων, οι οποίες θα στηρίζονται σε μέτρα που καθορίζουν αποτελεσματικότερα την παραγωγή συστάσεων.

1.5.1 Κατηγορίες Συστημάτων Συστάσεων

Τα συστήματα συστάσεων ταξινομούνται στις παρακάτω τρεις βασικές κατηγορίες ανάλογα με την προσέγγιση που εφαρμόζουν:

1. Φιλτράρισμα βασισμένο στο περιεχόμενο (Content - Based Filtering):

Τα συστήματα συστάσεων που βασίζονται στο περιεχόμενο συγκρίνουν τα ενδιαφέροντα του χρήστη, που έχουν συλλεχθεί έμμεσα ή άμεσα, με τα χαρακτηριστικά των αντικειμένων. Το βασικό χαρακτηριστικό της μεθόδου αυτής είναι το μέτρο ομοιότητας που δηλώνει πόσο σχετίζεται ένα αντικείμενο με κάποιον χρήστη. Τα συστήματα προτάσεων περιεχομένου που βασίζονται σε μοντέλα συνήθως «βλέπουν» τη δημιουργία σύστασης σαν ένα πρόβλημα κατηγοριοποίησης διαφορετικό για κάθε χρήστη, και μαθαίνουν έναν ταξινομητή για αυτά που αρέσουν και αυτά που δεν αρέσουν στον χρήστη με βάση τα χαρακτηριστικά του αντικειμένου.

2. Συνεργατικό φιλτράρισμα (Collaborative Filtering):

Χρησιμοποιεί δεδομένα σχετικά με τις προτιμήσεις ενός συνόλου χρηστών για να προτείνει περιεχόμενο σε έναν χρήστη-στόχο με παρόμοια ενδιαφέροντα. Συνήθως, αυτές οι μέθοδοι δεν χρησιμοποιούν πληροφορίες που έχουν να κάνουν με το περιεχόμενο αυτό καθ' αυτό, αλλά βασίζονται στις γνώμες των χρηστών (συνήθως, αξιολογήσεις που έχουν συλλεχθεί άμεσα). Τα συνεργατικά συστήματα προτάσεων βάσει μνήμης χρησιμοποιούν συνήθως ευρετικές τεχνικές, όπως ανάλυση συσχέτισης και ομοιότητα διανυσμάτων και μπορούν να χωριστούν σε δύο διαφορετικούς τύπους, ανάλογα με τη βάση της ομοιότητας: σε συστήματα βάσει χρήστη, όταν ο αλγόριθμος συνίσταται στην εύρεση παρόμοιων χρηστών με τον ενεργό και σε συστήματα βάσει αντικειμένου, όταν ο αλγόριθμος έχει να κάνει με την εύρεση αντικειμένων παρόμοιων με αυτά που αρέσουν στον ενεργό χρήστη. Τα συνεργατικά συστήματα προτάσεων βάσει μνήμης συνήθως χρησιμοποιούν πιθανοτικούς ταξινομητές όπως Μπεϋζιανά δίκτυα καθώς και μοντέλα συσταδοποίησης.

3. Υβριδικές τεχνικές συστάσεων (Hybrid Recommendation Methods):

Υβριδικά συστήματα συστάσεων είναι τα συστήματα που συνδυάζουν δύο ή περισσότερες τεχνικές συστάσεων. Σκοπός της δημιουργίας τους είναι ότι μπορούν να ξεπεράσουν τα προβλήματα που παρουσιάζουν οι υπάρχουσες τεχνικές συστάσεων.

1.5.2 Προβλήματα Τεχνικών Συστάσεων

Στην ενότητα αυτή παρουσιάζουμε τα σημαντικότερα προβλήματα που αντιμετωπίζουν οι τεχνικές συστάσεων:

1. **Το πρόβλημα Ψυχρής Εκκίνησης (Cold-Start Problem):** Το πρόβλημα ψυχρής εκκίνησης για έναν χρήστη δημιουργείται όταν ο χρήστης είναι καινούριος και το σύστημα δε γνωρίζει τις προτιμήσεις του, μιας και δεν έχει αξιολογήσει κάποιο αντικείμενο.
2. **Υπερ-Εξειδίκευση (Over-Specialization):** Το πρόβλημα αυτό παρουσιάζεται όταν το σύστημα προτείνει στον χρήστη συνεχώς αναμενόμενα αντικείμενα.
3. **Περιορισμένη Ανάλυση Περιεχομένου:** Τα Content - Based Filtering συστήματα έχουν έναν περιορισμό στον αριθμό και στον τύπο των χαρακτηριστικών των αντικειμένων. Για ακριβείς προβλέψεις, οι Content - Based Filtering τεχνικές χρειάζονται αρκετή πληροφορία για να διακρίνουν τα αντικείμενα που αρέσουν στον εκάστοτε χρήστη από αυτά που δεν του αρέσουν. Για παράδειγμα, μερικές προσεγγίσεις μπορεί να λαμβάνουν υπόψη μόνο μερικά από τα χαρακτηριστικά του περιεχομένου, ενώ απαιτούνται και τα υπόλοιπα για να προσφέρουν μία πιο ακριβή σύσταση.

1.6 Επεξεργασία Φυσικής Γλώσσας

Η Επεξεργασία της Φυσικής Γλώσσας (ΕΦΓ) (Natural Language Processing - NLP) είναι ένας διεπιστημονικός κλάδος της επιστήμης της Πληροφορικής, της Τεχνητής Νοημοσύνης και της Υπολογιστικής Γλωσσολογίας και ασχολείται με τις αλληλεπιδράσεις μεταξύ των υπολογιστών και της ανθρώπινης (φυσικής) γλώσσας. Η ανάπτυξη του πεδίου αυτού ξεκίνησε σαν ένα μέρος της ΤΝ. Προκλήσεις στην ΕΦΓ περιλαμβάνουν την κατανόηση φυσικής γλώσσας, δηλαδή την προσπάθεια να καταστούν ικανοί οι υπολογιστές να εξάγουν νοήματα από ανθρώπινα ή γλωσσικά δεδομένα, αλλά και την παραγωγή φυσικής γλώσσας.

Οι βασικές τεχνικές επεξεργασίας φυσικού κειμένου βασίζονται στις γενικές γνώσεις σχετικά με τη φυσική γλώσσα. Χρησιμοποιούν ορισμένους απλούς ευρετικούς κανόνες οι οποίοι στηρίζονται στη συντακτική και σημασιολογική προσέγγιση και ανάλυση του κειμένου. Ορισμένες τεχνικές που αφορούν σε όλα τα πεδία εφαρμογής είναι: ο διαμερισμός στα συστατικά στοιχεία του κειμένου (tokenization), η χρήση της διάταξης του κειμένου (structural data mining), η απαλοιφή λέξεων που δεν φέρουν ουσιαστική πληροφορία (elimination of insignificant words), η γραμματική δεικτοδότηση (PoS tagging), η μορφολογική ανάλυση και η συντακτική ανάλυση.

1.6.1 Πεδία έρευνας Επεξεργασίας Φυσικής Γλώσσας

Στην υποενότητα αυτή παρουσιάζονται τα κυριότερα πεδία στα οποία γίνεται εκτεταμένη έρευνα της Επεξεργασίας Φυσικής Γλώσσας. Το κριτήριο διαχωρισμού των πεδίων αυτών είναι το γεγονός ότι για το καθένα από αυτά υπάρχει ένας επίσημα ορισμένος χώρος μελέτης και επίλυσης ζητημάτων, ένα καθιερωμένο μετρικό σύστημα για την αξιολόγηση των ερευνών που προκύπτουν από το πεδίο, κάποια δεδομένα σύνολα κειμένων πάνω στα οποία κάθε πεδίο αξιολογείται και διαγωνισμοί αφιερωμένοι στο κάθε πεδίο.

- **Ανάλυση λόγου:** Αναγνώριση της δομής του λόγου εντός των αναλυόμενων κειμένων, π.χ. την φύση των σχέσεων του λόγου μεταξύ δύο προτάσεων. Επίσης, αναφέρεται στην αναγνώριση και την κατηγοριοποίηση των γλωσσικών πράξεων σε ένα μέρος του κειμένου.
- **Αυτόματη αναγνώριση ομιλίας:** Η αυτόματη μετατροπή του ανθρώπινου λόγου όπως προφέρεται σε κείμενο από τους υπολογιστές.
- **Αυτόματη ερωταπόκριση:** Η αναζήτηση της σωστής απάντησης σε μία δεδομένη ερώτηση, όπως διαμορφώνεται από την ανθρώπινη γλώσσα.
- **Αυτόματη μορφολογική τεμαχιοποίηση:** Η κατάτμηση των λέξεων στα μορφήματά τους, καθώς και η αναγνώριση και κατηγοριοποίηση αυτών των μορφημάτων. Η δυσκολία του συγκεκριμένου πεδίου μελέτης εξαρτάται σε μεγάλο βαθμό από την περιπλοκότητα της μορφολογίας της εκάστοτε γλώσσας υπό εξέταση.
- **Αυτόματη περίληψη:** Η παραγωγή μίας αναγνώσιμης (από τον άνθρωπο) περίληψης ενός κειμένου. Συχνά χρησιμοποιείται για να παρέχει περιλήψεις σε κείμενα γνωστής διάταξης, όπως οικονομικά ή πολιτικά ειδησεογραφικά άρθρα.
- **Εξόρυξη πληροφοριών:** Η ανάκτηση πληροφοριών από μη δομημένα ή ημιδομημένα δεδομένα (τυπικά κείμενα γραμμένα σε φυσική γλώσσα, ιστοσελίδες κ.α.)

- Επίλυση σχέσεων συναναφοράς: Η αναζήτηση των λέξεων (αναφορές) οι οποίες αναφέρονται στα ίδια υποκείμενα (οντότητες) σε μία δεδομένη πρόταση ή μεγαλύτερο τμήμα κειμένου. Η επίλυση σχέσεων αναφοράς είναι ένα συγκεκριμένο παράδειγμα αυτού του πεδίου και αναφέρεται συγκεκριμένα στην σύνδεση των αντωνυμιών με τα ουσιαστικά ή τα ονόματα στα οποία αναφέρονται.
- Επισήμανση των μερών του λόγου: Ο αυτόματος καθορισμός των μερών του λόγου σε μία δεδομένη πρόταση και η επίλυση της συντακτικής αμφισημίας.
- Κατανόηση του φυσικού λόγου: Η μετατροπή κομματιών κειμένου σε πιο τυπικές αναπαραστάσεις όπως σε δομές λογικής πρώτου βαθμού, οι οποίες μπορούν να μεταχειριστούν ευκολότερα από τους υπολογιστές.
- Μηχανική μετάφραση: Η αυτόματη μετάφραση ενός κειμένου από μία ανθρώπινη γλώσσα σε μία άλλη.
- Οπτική αναγνώριση χαρακτήρων (Optical Character Recognition - OCR): Ο προσδιορισμός του αντίστοιχου κειμένου από μία δεδομένη εικόνα που αναπαριστά κάποιο τυπογραφημένο κείμενο.
- Παραγωγή φυσικού λόγου: Η μετατροπή των πληροφοριών από υπολογιστικές βάσεις δεδομένων σε αναγνώσιμο φυσικό λόγο.
- Σύνθεση ομιλίας: Η αυτόματη, τεχνητή παραγωγή του ανθρώπινου λόγου από υπολογιστές.
- Συντακτική ανάλυση: Ο αυτόματος καθορισμός της σύνταξης μίας δεδομένης πρότασης και η επίλυση των οποιοδήποτε συντακτικών αμφισημιών. Εξαιτίας των αμφισημιών που πιθανόν να φέρει μία πρόταση, είναι δυνατόν η εν λόγω πρόταση να αναλυθεί σε παραπάνω από ένα συντακτικά δέντρα.

1.6.2 Ανάλυση Φυσικής Γλώσσας με χρήση Θεματικών Μοντέλων

Στη Μηχανική Εκμάθηση και στην Επεξεργασία Φυσικής Γλώσσας ένα θεματικό μοντέλο (topic model) είναι ένας τύπος στατιστικού μοντέλου για την ανακάλυψη θεμάτων που υπάρχουν σε μία συλλογή κειμένων. Τα μοντέλα θεμάτων βασίζονται στην ιδέα ότι τα κείμενα είναι μείγματα θεμάτων όπου ένα θέμα είναι μια πιθανότητα κατανομής λέξεων. Στόχος μας είναι να βρούμε σύντομες περιγραφές των μελών της συλλογής που επιτρέπουν μια αποτελεσματική επεξεργασία μεγάλων συλλογών, διατηρώντας τις απαραίτητες στατιστικές σχέσεις που είναι χρήσιμες για βασικές

διεργασίες όπως η περίληψη κειμένου.

Σημαντική επεξεργασία έχει γίνει πάνω σε αυτό το πρόβλημα από ερευνητές στο αντικείμενο του πεδίου ανάκτησης πληροφοριών (IR) (Baeza-Yates και Ribeiro-Neto, 1999). Η βασική μεθοδολογία που προτάθηκε από τους IR ερευνητές για συλλογές κειμένων – μια μεθοδολογία η οποία εφαρμόστηκε με επιτυχία στις μοντέρνες μηχανές αναζήτησης του Διαδικτύου – μετατρέπει κάθε κείμενο σε ένα διάνυσμα πραγματικών αριθμών, καθένα από τα οποία αντιπροσωπεύει μία αναλογία μετρήσεων. Στο δημοφιλές **tf-idf** σχήμα (Salton και McGill, 1983) επιλέγεται το βασικό λεξιλόγιο των “όρων” και για κάθε κείμενο της συλλογής σχηματίζεται μία μέτρηση από τον αριθμό των φορών που έχει παρουσιαστεί μια λέξη. Μετά από κατάλληλη κανονικοποίηση, ο δείκτης συχνότητας συγκρίνεται με το αντίστροφο δείκτη συχνότητας κειμένων, που μετράει τον αριθμό που έχει βρεθεί μια λέξη σε όλη την συλλογή (γενικά σε λογαριθμική κλίμακα και μετά μοντελοποιείται κατάλληλα). Το τελικό αποτέλεσμα είναι ένας πίνακας όρων ανά κειμένων X , του οποίου οι στήλες περιέχουν τις **tf-idf** αξίες για κάθε κείμενό της. Έτσι, το **tf-idf** σχήμα μειώνει τα κείμενα αυθαίρετου μήκους σε φορμαρισμένου μήκους λίστες αριθμών.

Καθώς η **tf-idf** μείωση έχει κάποιες δελεαστικές ιδιότητες, η προσέγγιση παρέχει επίσης μια σχετικά μικρή μείωση του μήκους περιγραφής και φανερώνει λίγα για την στατιστική μορφή των κειμένων. Για την αντιμετώπιση των παραπάνω προβλημάτων οι IR ερευνητές πρότειναν πολλές άλλες τεχνικές μείωσης του πλήθους διαστάσεων, η πιο αξιοσημείωτη εκ των οποίων είναι η **Latent Semantic Indexing (LSI)** (Deerwester et al., 1990)). Η LSI εφαρμόζει αποσύνθεση τιμών στον X πίνακα για να ορίσει έναν γραμμικό υποχώρο στον χώρο των **tf-idf** ιδιοτήτων, που πιάνει ένα μεγάλο μέρος της διακύμανσης της συλλογής. Αυτή η προσέγγιση μπορεί να πετύχει μια σημαντική ελαχιστοποίηση μεγάλων συλλογών. Επιπλέον, ο Deerwester υποστηρίζει ότι τα παραγόμενα χαρακτηριστικά του LSI, που είναι γραμμικοί συνδυασμοί των αυθεντικών **tf-idf** χαρακτηριστικών, μπορούν να πιάσουν κάποιες πτυχές των βασικών γλωσσολογικών ιδεών όπως η συνωνυμία και η πολυσημία.

Ένα σημαντικό βήμα ήταν η παρουσίαση του πιθανοτικού LSI (**probabilistic LSI**) μοντέλου, γνωστού επίσης και σαν Μοντέλο Συμπερασμάτων. Η **pLSI** προσέγγιση μοντελοποιεί κάθε λέξη ενός κειμένου ως ένα δείγμα ενός mixture μοντέλου, όπου τα αναμεμιγμένα στοιχεία είναι τυχαίες πολυωνυμικές μεταβλητές που μπορούν να κατανοηθούν ως παρουσιάσεις θεμάτων. Έτσι, κάθε λέξη δημιουργείται από ένα απλό θέμα, και διαφορετικές λέξεις σε ένα κείμενο μπορεί να δημιουργηθούν από διαφορετικά θέματα. Κάθε κείμενο απεικονίζεται ως μία λίστα αναμεμιγμένων

αναλογιών για κάθε mixture συστατικό και έτσι μειώνεται σε μια πιθανοτική κατανομή φορμαρισμένων συνόλων θεμάτων. Αυτή η κατανομή είναι μια “ελαχιστοποιημένη περιγραφή” συσχετισμένη με ένα κείμενο.

Αν και η δουλειά του Hoffman είναι ένα χρήσιμο βήμα πάνω στην θεματική μοντελοποίηση κειμένων, δεν είναι πλήρης, γιατί δεν παρέχει κανένα πιθανοτικό μοντέλο για την επεξεργασία σε επίπεδο κειμένων. Στο pLSI κάθε κείμενο παρουσιάζεται ως λίστα αριθμών και δεν υπάρχει κανένα γενετικό πιθανοτικό μοντέλο για αυτούς τους αριθμούς. Αυτό επιφέρει αρκετά προβλήματα: (1) ο αριθμός των παραμέτρων στο μοντέλο μεγαλώνει γραμμικά με το μέγεθος της συλλογής, κάτι που οδηγεί σε σοβαρά προβλήματα σχετικά με το overfitting και (2) δεν είναι ξεκάθαρο πως διανέμεται η κατανομή σε ένα κείμενο εκτός του εκπαιδευτικού συνόλου.

Για να καταλάβετε πώς θα προχωρήσουμε πέρα από το LSI, ας θεωρήσουμε τις βασικές πιθανοτικές υποθέσεις υπογραμμίζοντας την κλάση των μεθόδων μείωσης διαστάσεων που περιέχει το LSI και το pLSI. Όλες αυτές οι μέθοδοι βασίζονται στην *bag-of-words* υπόθεση – ότι η σειρά των λέξεων μπορεί να αγνοηθεί. Στη γλώσσα της πιθανοτικής θεωρίας αυτή είναι η υπόθεση της ανταλλαξιμότητας των λέξεων σε ένα κείμενο. Αυτές οι μέθοδοι υποθέτουν επίσης ότι τα κείμενα είναι ανταλλάξιμα: η σειρά των κειμένων σε μια συλλογή μπορεί να αλλάξει.

Ένα κλασσικό θεώρημα του de Finetti (1990) τεκμηριώνει ότι κάθε συλλογή από ανταλλάξιμες τυχαίες μεταβλητές έχει μια απεικόνιση σαν mixture κατανομή. Έτσι, αν επιθυμούμε να θεωρήσουμε ανταλλάξιμες απεικονίσεις κειμένων για κείμενα και λέξεις, πρέπει να θεωρήσουμε mixture μοντέλα που λαμβάνουν υπόψη τους την ανταλλαξιμότητα και των λέξεων αλλά και των κειμένων. Αυτό οδήγησε στην δημιουργία του μοντέλου **Latent Dirichlet Allocation (LDA)**. Το LDA είναι ένα γενετικό πιθανοτικό μοντέλο ενός σώματος. Η βασική ιδέα είναι ότι τα κείμενα αντιπροσωπεύονται από τυχαίες προσμειζεις κρυφών θεμάτων, όπου κάθε θέμα χαρακτηρίζεται από μία κατανομή ως προς τις λέξεις. Τα pLSI και LDA μοντέλα είναι παρόμοια, με το LDA να αποτελεί την *Bayesian* εκδοχή του pLSI. Το μοντέλο αυτό θα αναλυθεί περαιτέρω στο Κεφάλαιο 2 όπου και θα χρησιμοποιηθεί.

Κεφάλαιο 2

Ανάλυση και Σχεδίαση Συστήματος

2.1 Περιγραφή Συστήματος

Το αντικείμενο του κεφαλαίου αυτού είναι μια λεπτομερής περιγραφή του συστήματος που αναπτύξαμε. Σκοπός της εργασίας είναι να παρουσιάσει ένα σύστημα το οποίο επιτρέπει την εξαγωγή συμπερασμάτων για την υποστήριξη εξατομικευμένων συστάσεων για άρθρα ειδήσεων.

Όπως έχουμε ήδη αναφέρει και σε προηγούμενο κεφάλαιο, το διαδίκτυο αυξάνεται με γρήγορους ρυθμούς και παράλληλα αυξάνεται και η πληροφορία που περιέχεται σε αυτό, και συγκεκριμένα η ενημερωτική πληροφορία. Όλο και περισσότεροι χρήστες χρησιμοποιούν το διαδίκτυο για να ενημερώνονται. Λόγω του μεγάλου όγκου των πληροφοριών που κατακλύζουν το διαδίκτυο, οι χρήστες δυσκολεύονται να ξεχωρίσουν τις πληροφορίες που σχετίζονται πραγματικά με τα ενδιαφέροντά τους. Η παραπάνω κατάσταση δημιουργεί ένα σημαντικό πρόβλημα για τους χρήστες του διαδικτύου σε καθημερινή βάση.

Η διαρκής ροή ειδήσεων από διάφορες πηγές ενημέρωσης δεν εξυπηρετεί τον χρήστη, καθώς καθιστά αδύνατη την παρακολούθησή τους και κρίνεται επιβεβλημένη η προσωποποίηση του αποτελέσματος. Σε ένα τέτοιο σενάριο τα συστήματα συστάσεων μπορούν να αντιμετωπίσουν το πρόβλημα, προτείνοντας αντικείμενα που προηγουμένως ήταν άγνωστα, στην περίπτωση μας άρθρα ειδήσεων, και θα ενδιέφεραν έναν συγκεκριμένο χρήστη. Τυπικά, τέτοια συστήματα χρησιμοποιούν προφίλ χρηστών και έχουν σαν στόχο τη σύσταση ειδήσεων που ταιριάζουν καλύτερα στο προφίλ αυτό.

Στο σύστημά μας τα άρθρα ειδήσεων λαμβάνονται από το διαδίκτυο από αρκετές υπηρεσίες ειδήσεων, καθώς και από τη συλλογή άρθρων *Reuters* του NLTK (Natural Language Toolkit). Τα άρθρα αποθηκεύονται στη βάση δεδομένων του συστήματος

και το σύστημα επεξεργάζεται το περιεχόμενο κάθε άρθρου που είναι αποθηκευμένο στη βάση.

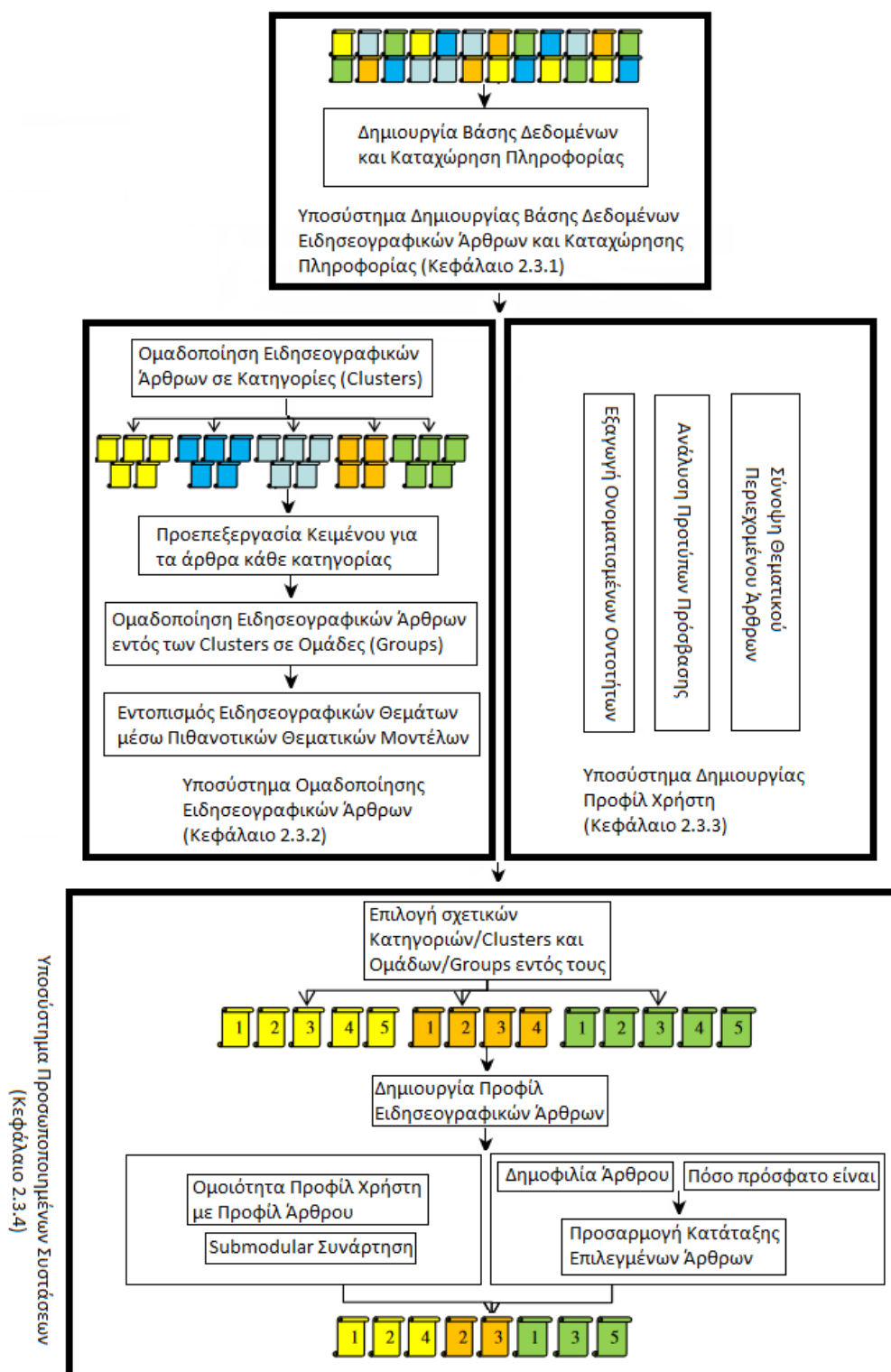
Στη συνέχεια, το λειτουργικό μέρος το οποίο πραγματοποιεί την δημιουργία του προφίλ δέχεται σαν είσοδο τις προτιμήσεις του χρήστη. Ο μηχανισμός προτάσεων του συστήματος παράγει ένα σύνολο εξατομικευμένων προτάσεων που παρουσιάζει στον χρήστη.

Παρακάτω, παρουσιάζεται η μελέτη που έγινε για την υλοποίηση του συστήματος συστάσεων. Αρχικά, παρουσιάζεται η αρχιτεκτονική του συστήματος και γίνεται ο διαχωρισμός του στα επιμέρους υποσυστήματα και εν συνεχεία, αναλύεται λεπτομερώς ο τρόπος λειτουργίας των υποσυστημάτων του.

2.2 Αρχιτεκτονική Συστήματος

Στην ενότητα αυτή παρουσιάζεται ο χωρισμός του συστήματος συστάσεων σε υποσυστήματα όσον αφορά την αρχιτεκτονική.

Η αρχιτεκτονική που χρησιμοποιήθηκε για το σύστημά μας παρουσιάζεται στο παρακάτω σχήμα:



Σχήμα 2.1: Αρχιτεκτονική Συστήματος

2.3 Υποσυστήματα

Στη συνέχεια παρουσιάζουμε τα διάφορα υποσυστήματα από τα οποία αποτελείται ο μηχανισμός προκειμένου να γίνει κατανοητή η λειτουργία του.

1. Υποσύστημα Δημιουργίας Βάσης Δεδομένων και Καταχώρησης Πληροφορίας
 - Συλλογή Ειδησεογραφικών Άρθρων
 - Δημιουργία Βάσης Δεδομένων και Καταχώρηση Πληροφορίας
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
3. Υποσύστημα Δημιουργίας Προφίλ Χρήστη
 - Σύνοψη Θεματικού Περιεχομένου Άρθρων
 - Ανάλυση Προτύπων Πρόσβασης
 - Εξαγωγή Ονοματισμένων Οντοτήτων
4. Υποσύστημα Προσωποποιημένων Συστάσεων
 - Αντιστοίχιση Αναγνωστικών Προτιμήσεων για την Αναπαράσταση 1ου Επιπέδου
 - Αντιστοίχιση Αναγνωστικών Προτιμήσεων για την Αναπαράσταση 2ου Επιπέδου
 - (α') Δημιουργία Προφίλ Ειδησεογραφικών Άρθρων
 - (β') Εισαγωγή στις Submodular Συναρτήσεις
 - (γ') Μοντέλο Συστάσεων
 - (δ') Προσαρμογή Κατάταξης Ειδησεογραφικών Άρθρων

Παρακάτω δίνεται λεπτομερής περιγραφή για καθένα από τα υποσυστήματα που αναφέραμε.

2.3.1 Υποσύστημα Δημιουργίας Βάσης Δεδομένων και Καταχώρησης Πληροφορίας

Η συλλογή των άρθρων ανακτήθηκε κατά βάση από το διαδίκτυο από αρκετές διαδικτυακές υπηρεσίες ειδήσεων, όπως οι εξής: *The Guardian*, *New York Times*, *Washington Post*, *Fox News*, *Independent*, *Reuters*, *Sky News*. Ένα μέρος της συλλογής προήλθε από τη συλλογή άρθρων *Reuters* του NLTK.

Μετά την ολοκλήρωση της διαδικασίας συλλογής, δημιουργούμε τη βάση δεδομένων και αποθηκεύουμε τα άρθρα στους αντίστοιχους πίνακες της βάσης του συστήματος. Οι πληροφορίες που αποθηκεύονται για κάθε άρθρο είναι: τίτλος, συγγραφέας, ημερομηνία δημοσίευσης, κείμενο άρθρου, γενική κατηγορία (cluster) στην οποία ανήκει το κάθε άρθρο.

Επιπρόσθετα, στο αρχικό αυτό στάδιο, δημιουργούμε και καταχωρούμε μία σειρά από χρήστες του συστήματος. Συγκεκριμένα, κατά τη δημιουργία της βάσης δεδομένων παράγεται αυτόματα ένα αναγνωστικό ιστορικό για κάθε αποθηκευμένο χρήστη, τόσο ελεγχόμενα όσο και τυχαία και αποθηκεύεται στον αντίστοιχο πίνακα της βάσης. Ως προς το ελεγχόμενο μέρος επιχειρούμε να ταιριάζουμε κάθε αποθηκευμένο χρήστη με μία συγκεκριμένη κατηγορία άρθρων, φορτώνοντας στο αναγνωστικό ιστορικό του μέσω ενός αρχείου μεγαλύτερο αριθμό άρθρων από κάποια συγκεκριμένη κατηγορία. Μέσω του ελεγχόμενου τρόπου καταφέρνουμε να διασφαλίσουμε ότι κάθε άρθρο της βάσης δεδομένων θα έχει αναγνωσθεί τουλάχιστον μία φορά από κάποιον τυχαίο χρήστη. Ως προς το τυχαίο μέρος επιλέγουμε μέσω της συνάρτησης *rand* και έναν τυχαίο (εντός ορίων) αριθμό άρθρων από οποιαδήποτε κατηγορία για να εμπλουτίσουμε το περιεχόμενο του αναγνωστικού ιστορικού του χρήστη.

Πλέον, έχουμε συλλέξει στη βάση δεδομένων όλη την απαραίτητη πληροφορία για το επόμενο υποσύστημα.

2.3.2 Υποσύστημα Ομαδοποίησης Ειδησεογραφικών Άρθρων

Τυπικά, δεδομένου ενός σετ ειδησεογραφικών άρθρων $N = \{n_1, n_2, \dots, n_M\}$, όπου $|N| = M$, στόχος μας είναι μια στιβαρή ομαδοποίηση (hard clustering) $C = \{C_1, C_2, \dots, C_K\}$ στο N , όπου το K είναι ένας προκαθορισμένος αριθμός από clusters. Κάθε cluster C_i αποτελείται από μία λίστα από groups ειδησεογραφικών άρθρων $G = \{G_1, G_2, \dots\}$ και κάθε group G_j περιέχει t_j ειδησεογραφικά άρθρα. Κάθε cluster καθώς και τα groups άρθρων που περιλαμβάνει είναι αντιστοίχως συσχετισμένα με μία κατανομή θεμάτων T , η οποία περιγράφει τα θέματα που “κρύβονται” μέσα στα άρθρα. Στόχος της συγκεκριμένης αναπαράστασης είναι ο εξής: Οι προσωποποιημένες

συστάσεις άρθρων απαιτούν γρήγορη απόκριση για να παρουσιάσουν άμεσα αποτελέσματα στους χρήστες. Η συγκεκριμένη αναπαράσταση άρθρων μπορεί να βοηθήσει στη γρήγορη πλοήγηση προς άρθρα που ενδιαφέρουν το χρήστη.

Ομαδοποίηση Ειδησεογραφικών Άρθρων σε Κατηγορίες (Clusters)

Δεδομένου του τρόπου συλλογής των άρθρων που βρίσκονται αποθηκευμένα στη βάση δεδομένων του συστήματος, γνωρίζουμε εκ των προτέρων την κατηγορία στην οποία ανήκουν το κάθε ένα από αυτά. Έτσι, είτε πρόκειται για άρθρα που προέρχονται από διαδικτυακές υπηρεσίες ειδήσεων, είτε για άρθρα από τη συλλογή Reuters του NLTK, κάθε ένα είναι εξ αρχής συσχετισμένο με μια συγκεκριμένη κατηγορία από το σύνολο των διαφορετικών κατηγοριών που έχουμε επιλέξει.

Στο σύστημά μας εμφανίζονται άρθρα από επτά διαφορετικές κατηγορίες. Κάθε κατηγορία σχετίζεται με ένα από τα παρακάτω θέματα: *Science/Technology, Politics, Sports, Life & Style, Sugar, Coffee, Housing*. Οι τέσσερις πρώτες κατηγορίες περιλαμβάνουν άρθρα που συλλέχθηκαν από το διαδίκτυο, ενώ οι υπόλοιπες τρεις κατηγορίες αποτελούνται από άρθρα της συλλογής Reuters.

Σε αυτό το στάδιο της προεπεξεργασίας της συλλογής κειμένων μπορούμε να θεωρήσουμε με βεβαιότητα ότι αυτή η πρώτη ομαδοποίηση που πραγματοποιήθηκε πάνω στη συλλογή κειμένων είναι και η πιο αποτελεσματική, καθώς έχουμε εξασφαλίσει ότι τα ειδησεογραφικά άρθρα με κοινό θεματικό περιεχόμενο ανήκουν στο ίδιο cluster.

Τέλος, είναι σαφές ότι κάθε άρθρο ανήκει αποκλειστικά σε ένα και μοναδικό cluster.

Παρακάτω προχωρούμε σε περαιτέρω ομαδοποίηση των άρθρων, αυτή τη φορά εντός των clusters που δημιουργήσαμε, με σκοπό τη δημιουργία της βάσης του πρώτου επιπέδου σύστασης ειδησεογραφικών άρθρων του συστήματός μας.

Προεπεξεργασία Κειμένου (Text Preprocessing)

Πρωτού προβούμε σε εφαρμογή του αλγορίθμου ομαδοποίησης των άρθρων εντός των clusters σε ομάδες (groups), κρίνεται απαραίτητη η προεπεξεργασία των κειμένων των άρθρων κάνοντας χρήση της βιβλιοθήκης *scikit-learn* της Python. Η πληροφορία που δίνεται σαν είσοδος στο μηχανισμό προέρχεται από τη βάση δεδομένων του συστήματος και αποτελείται από τα κείμενα των άρθρων κάθε cluster. Πρωταρχικός σκοπός μας στο στάδιο αυτό είναι η απομάκρυνση ανεπιθύμητων συμβολοσειρών μέσω φιλτραρίσματος των λέξεων που δε φέρουν ουσιαστική πληροφορία και η ανάλυση του κειμένου για την εξαγωγή λέξεων-κλειδιών από το κείμενο του κάθε άρθρου. Βασικά σημεία προεπεξεργασίας:

- **Λεξική Ανάλυση:** Το στάδιο αυτό αφορά τον διαμερισμό κάθε άρθρου στα

συστατικά στοιχεία του κειμένου του (tokenization), μετατρέποντας το κείμενο σε ακολουθία λέξεων.

- **Αφαίρεση τετριμμένων λέξεων και τερματικών όρων (stopwords):** Για κάθε άρθρο της συλλογής πραγματοποιείται καθαρισμός από τετριμμένες λέξεις (elimination of insignificant words), όπως είναι τα άρθρα, οι σύνδεσμοι, οι αντωνυμίες, καθώς και οι συχνά χρησιμοποιούμενες λέξεις που δεν ανήκουν στις ανωτέρω κατηγορίες, αλλά δεν φέρουν καμία ιδιαίτερη σημασιολογική πληροφορία, όπως είναι ορισμένα επιρρήματα. Ένα *stopword* είναι μια συχνά χρησιμοποιούμενη λέξη, όπως οι “the”, “and”, “to”, “also”, όπου μία μηχανή αναζήτησης έχει σχεδιαστεί να αγνοεί, τόσο κατά την ευρετηρίαση των καταχωρήσεων προς αναζήτηση, όσο και κατά την ανάκτησή τους ως αποτελέσματα μιας αναζήτησης. Δεν επιθυμούμε αυτές οι λέξεις να καταλαμβάνουν χώρο στη βάση δεδομένων ή να κατασπαταλούν πολύτιμο υπολογιστικό χρόνο, καθότι φέρουν ελάχιστον λεκτικό περιεχόμενο και η παρουσία τους σε ένα κείμενο δε διευκολύνει τη διάκριση του κειμένου από άλλα κείμενα. Το NLTK περιέχει μία λίστα από *stopwords* αποθηκευμένα σε δεκαέξι διαφορετικές γλώσσες. Στο σημείο αυτό απομακρύνουμε, επίσης, τα αριθμητικά δεδομένα και τα σημεία στίξης.
- **Κανονικοποίηση των λέξεων:** Το στάδιο αυτό αφορά την αναγνώριση των ριζών των λέξεων, γνωστή ως λημματοποίηση (lemmatization) και την αποκατάληξη (stemming), ώστε να μην επηρεάζεται η εξαγωγή χαρακτηριστικών γνωρισμάτων των κειμένων από την πτώση ή το χρόνο που κλίνουνται οι λέξεις. Έτσι, καταλήγουμε σε αναγωγή όλων των μορφολογικών τύπων μίας λέξης σε μία ενιαία αναπαράσταση.
- **Επιλογή των αντιπροσωπευτικών όρων:** Στη συνέχεια, η διαδικασία βασίζεται στην εξαγωγή χαρακτηριστικών γνωρισμάτων (λέξεων-κλειδιών) των κειμένων, η οποία αφορά την εφαρμογή μέτρων ποιότητας για την επιλογή διατήρησης ορισμένων όρων για κάθε κείμενο. Η εξαγωγή των λέξεων-κλειδιών των κειμένων γίνεται μέσω του εργαλείου *TfidfVectorizer* [29] της βιβλιοθήκης *scikit-learn* της Python. Η εξαγωγή των σωστών λέξεων-κλειδιών είναι πολύ σημαντική.
Η μέθοδος **TF-IDF** [26] είναι μια μετρική που δηλώνει πόσο σημαντικός είναι ένας όρος σε ένα έγγραφο από μια συλλογή εγγράφων. Η μέθοδος **TF-IDF** στοχεύει στο να σταθμίσει όλους τους όρους μιας συλλογής κειμένων. Με λίγα λόγια, στόχος της είναι να αποδώσει το αντίστοιχο βάρος σε κάθε όρο και, κατά επέκταση, σε κάθε διάσταση του πολυδιάστατου αυτού χώρου. Αυτό συμβαίνει γιατί η απλή αρίθμηση ενός όρου σε ένα κείμενο δεν αρκεί για να μας πληροφορήσει για τη σημαντικότητα του όρου αυτού και τη βαρύτητα της

πληροφορίας που περιέχει.

Η μέθοδος αυτή αποτελείται από τις ποσότητες TF και IDF. Η ποσότητα TF (συχνότητα όρου) υποδηλώνει το πόσες φορές εμφανίζεται ένας όρος σε ένα κείμενο. Από την άλλη, η ποσότητα IDF υποδηλώνει το πόσο ένας όρος είναι διαδεδομένος σε ένα κείμενο αλλά και σε ολόκληρη τη συλλογή κειμένων. Τελικά, το βάρος ενός όρου προκύπτει από τον πολλαπλασιασμό των ποσοτήτων TF και IDF.

Στόχος της μεθόδου αυτής μέσω του βάρους είναι η επιλογή εκείνων των όρων που αποτυπώνουν καλύτερα το περιεχόμενο ενός κειμένου. Για τον προσδιορισμό του βάρους ενός όρου είναι εξίσου σημαντικές και οι δύο ποσότητες TF και IDF. Αυτό επισημαίνεται διότι αν χρησιμοποιούσαμε μόνο τη συχνότητα εμφάνισης ενός όρου (TF) ως βάρος, αυτό θα είχε ως συνέπεια οι συχνότερα εμφανιζόμενοι όροι να θεωρούνται ως οι πιο σημαντικοί. Αυτή η υπόθεση θα μπορούσε να μας οδηγήσει σε λανθασμένη επιλογή όρων οι οποίοι εμφανίζονται σε πολλά κείμενα και δεν προσφέρουν κάποια ιδιαίτερη πληροφορία σε ένα κείμενο. Για παράδειγμα, η λέξη «εξόρυξη» σε μία συλλογή κειμένων με θέμα «Τεχνικές Εξόρυξης Κειμένων» θα εμφανίζεται με μεγάλη συχνότητα σε όλα τα κείμενα της συλλογής. Επομένως, μέσα από αυτό το παράδειγμα καταλαβαίνουμε πως ένας τέτοιος όρος, παρότι θα μπορούσε να εμφανίζεται αρκετές φορές σε ένα κείμενο, δε θα μπορούσε να θεωρηθεί ως ένας σημαντικός όρος, γιατί δεν προσφέρει ένα ιδιαίτερο χαρακτηριστικό στο κείμενο σε σχέση με τα υπόλοιπα κείμενα της συλλογής. Εδώ, λοιπόν, καταλαβαίνουμε τη σημαντικότητα της ποσότητας IDF στον υπολογισμό του βάρους ενός όρου. Όταν ένας όρος εμφανίζεται σε πολλά κείμενα της συλλογής, η τιμή της ποσότητας IDF είναι μικρή, ενώ όταν ένας όρος εμφανίζεται σε λίγα κείμενα της συλλογής, η τιμή της ποσότητας IDF είναι μεγάλη.

Επομένως, μεγάλο βάρος για έναν όρο προκύπτει όταν ο όρος αυτός εμφανίζεται πολλές φορές σε ένα κείμενο και λιγότερες φορές στο σύνολο των κειμένων.

Χρησιμοποιώντας τη διαδικασία που περιγράψαμε παραπάνω, ο μηχανισμός διαβάζει καινούργια άρθρα από τη βάση δεδομένων του συστήματος ανά κατηγορία (cluster), εξάγει τις λέξεις-κλειδιά για κάθε άρθρο της κατηγορίας και τις συσχετίζει με το άρθρο και το αντίστοιχο βάρος. Έχοντας ολοκληρώσει την προεπεξεργασία των κειμένων της συλλογής, τα δεδομένα μας έχουν πλέον την κατάλληλη μορφή ώστε να προχωρήσουμε στη διαδικασία ομαδοποίησης των άρθρων εντός της κάθε κατηγορίας (cluster).

Ομαδοποίηση Ειδησεογραφικών Άρθρων εντός των Clusters σε Ομάδες (Groups)

Τα ειδησεογραφικά άρθρα που περιέχονται σε κάθε cluster επικεντρώνονται σε παρόμοιες πτυχές ενός γενικού θέματος, εκείνου που αντιπροσωπεύεται από τον τίτλο της εκάστοτε κατηγορίας. Τυπικά, ένας χρήστης ενδιαφέρεται για συγκεκριμένες πτυχές ενός θέματος και όχι για όλες. Βασισμένοι σε αυτή την παραδοχή και με στόχο την γρήγορη πλοήγηση, κατά το στάδιο σύστασης, προς άρθρα που ενδιαφέρουν το χρήστη, προχωρούμε σε περαιτέρω ομαδοποίηση των άρθρων εντός κάθε κατηγορίας με σκοπό τη δημιουργία ομάδων (groups) εντός κάθε cluster.

Η ομαδοποίηση εντός κάθε cluster πραγματοποιείται με εφαρμογή του αλγορίθμου k-means με βάση κάποιο μέτρο ομοιότητας, με στόχο όλα τα άρθρα που ανήκουν στην ίδια ομάδα να είναι παρόμοια μεταξύ τους. Για την εφαρμογή του αλγορίθμου γίνεται χρήση της βιβλιοθήκης scikit-learn της Python.

Ο αλγόριθμος **k-means** [2] είναι ο πιο διαδεδομένος αλγόριθμος ομαδοποίησης. Ο k-means διασπάει τα δεδομένα σε k διαφορετικές ομάδες μέσω μίας επαναληπτικής διαδικασίας, όπου k είναι ένας προκαθορισμένος ακέραιος αριθμός και παρέχεται από εμάς. Η επαναληπτική διαδικασία τερματίζει τη στιγμή που θα ικανοποιηθεί ένα συγκεκριμένο κριτήριο. Τα κύρια βήματα του αλγορίθμου περιγράφονται παρακάτω:

1. Αρχικά, γίνεται η επιλογή k σημείων στο πεδίο των δεδομένων έτσι ώστε τα σημεία αυτά να αποτελούν τα κέντρα των αρχικών ομάδων.
2. Ανάθεσε κάθε δεδομένο σε μία ομάδα για το οποίο η απόστασή του από το κέντρο της ομάδας να είναι η μικρότερη από κάθε άλλη εκ των κέντρων των υπολοίπων ομάδων.
3. Όταν όλα τα δεδομένα έχουν ανατεθεί στις ομάδες, υπολόγισε ξανά τα κέντρα των ομάδων, παίρνοντας το μέσο όρο των δεδομένων κάθε ομάδας.
4. Επανάλαβε τα βήματα 2, 3 μέχρι να επαληθευτεί κάποιο κριτήριο σύγκλισης.

Το κριτήριο σύγκλισης του αλγορίθμου, ο τρόπος μέτρησης της απόστασης των δεδομένων από τα κέντρα των ομάδων, καθώς και ο τρόπος ανάδειξης των αρχικών κέντρων τους καθορίζουν σε μεγάλο βαθμό την τελική ομαδοποίηση των δεδομένων και ορίζονται από το χρήστη.

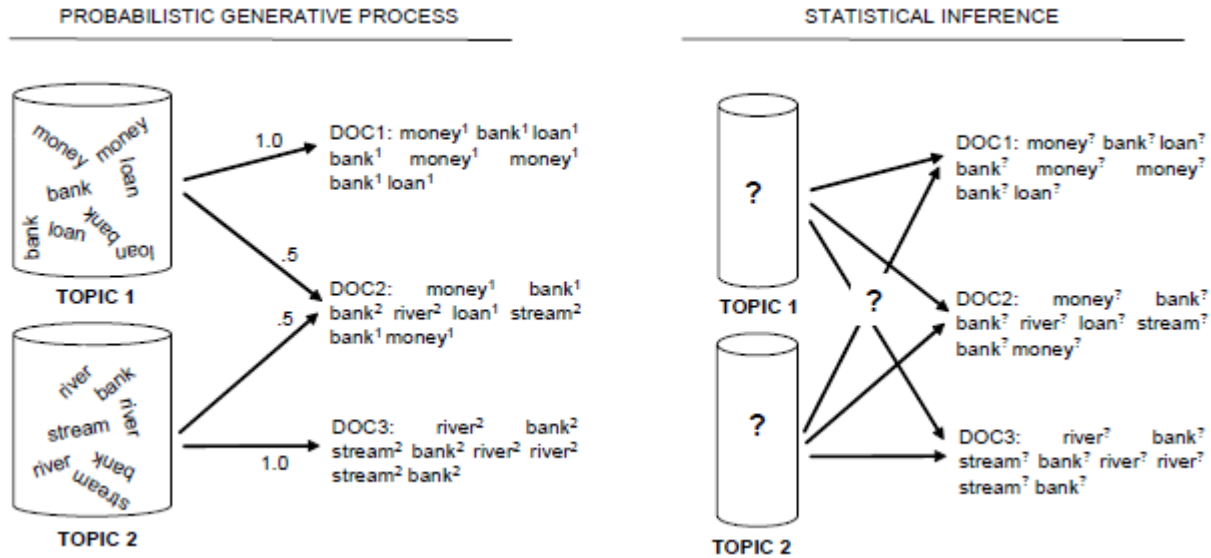
Τα ήδη υπάρχοντα clusters μαζί με τα παραγόμενα groups που περιέχονται σε αυτά αποτελούν τη βάση του πρώτου επιπέδου σύστασης ειδησεογραφικών άρθρων του συστήματός μας.

Εντοπισμός Ειδησεογραφικών Θεμάτων μέσω Πιθανοτικών Θεματικών Μοντέλων

Ένας φυσικός τρόπος να εξερευνήσουμε τους συσχετισμούς μεταξύ των clusters (ή των groups) άρθρων και του προφίλ του δοθέντος χρήστη είναι να συγκρίνουμε την ομοιότητα των θεμάτων που “κρύβονται” μέσα στα άρθρα τους. Γενικά, η ανακάλυψη κρυμμένων θεμάτων που υπάρχουν σε μια συλλογή κειμένων πραγματοποιείται χρησιμοποιώντας πιθανοτικά μοντέλα θεμάτων, όπως τα PLSI και LDA [12], εξάγοντας μια λίστα αντιπροσωπευτικών λέξεων από την αρχική συλλογή κειμένων μαζί με το αντίστοιχο βάρος για κάθε λέξη. Τα μοντέλα θεμάτων μοντελοποιούν κάθε αντικείμενο μιας συλλογής ως ένα πεπερασμένο μείγμα από ένα σύνολο θεματικών πιθανοτήτων.

Ένα γενετικό (generative) μοντέλο θεμάτων βασίζεται στους απλούς πιθανοτικούς κανόνες επιλογής (sampling) που περιγράφουν πως οι λέξεις μέσα στα κείμενα μπορεί να δημιουργηθούν μέσω τυχαίων κρυφών μεταβλητών. Όταν προσαρμόζουμε ένα γενετικό μοντέλο, ο στόχος είναι να βρεθεί το καταλληλότερο σύνολο κρυφών μεταβλητών που μπορεί να εξηγήσει τα παρατηρούμενα δεδομένα υποθέτοντας ότι το μοντέλο δημιουργήθηκε από τα δεδομένα. Το Σχήμα 2.2 δείχνει την προσέγγιση μοντέλων θεμάτων με δύο τρόπους: σαν γενετικό μοντέλο και σαν πρόβλημα μεταβολικής συμπερασματολογίας. Στα αριστερά, η γενετική διαδικασία παρουσιάζεται με δύο θέματα (topics). Τα Topic 1 και 2 είναι θεματικά συσχετισμένα με το ‘money’ και τα ‘rivers’ και απεικονίζονται σαν τσάντες που περιέχουν διαφορετικές κατανομές ως προς τις λέξεις. Διαφορετικά κείμενα μπορούν να παραχθούν επιλέγοντας λέξεις από ένα θέμα δεδομένου του βάρους που έχει. Για παράδειγμα, τα κείμενα Doc1 και Doc3 έχουν δημιουργηθεί με την διαδικασία *sampling* μόνο από το topic 1 και το topic 2 αντίστοιχα, ενώ το κείμενο Doc2 δημιουργήθηκε από την ίση μείξη των δύο θεμάτων. Με τον τρόπο όπου το μοντέλο ορίστηκε, δεν υπάρχει κάποια αμοιβαία αποκλειστικότητα που να περιορίζει τις λέξεις να είναι μέρος ενός μόνο θέματος. Αυτό επιτρέπει στο μοντέλο να καταλαβαίνει την πολυσημία, δηλαδή τις πολλές σημασίες όπου η ίδια η λέξη μπορεί να έχει ταυτόχρονα. Για παράδειγμα, και τα δύο θέματα ‘money’ και ‘rivers’ δίνουν υψηλή πιθανότητα στην λέξη *bank*, που είναι λογικό, δεδομένης της πολυσημικής φύσης της λέξης.

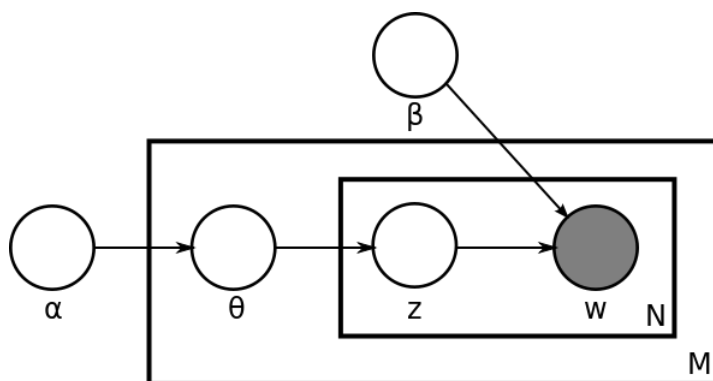
Στο σύστημά μας εφαρμόζουμε το **LDA (Latent Dirichlet Allocation)** ως το μοντέλο για την ανακάλυψη κρυμμένων θεμάτων και αναπαριστούμε την κατανομή θεμάτων ως ένα διάνυσμα κάθε εγγραφή του οποίου δηλώνει το βάρος της αντίστοιχης λέξης. Για την εφαρμογή του LDA μοντέλου γίνεται χρήση της βιβλιοθήκης *gensim* [7] της Python.



Σχήμα 2.2: Απεικόνιση γενετικής διαδικασίας και του προβλήματος στατιστικής συμπερασματολογίας που λύνεται με μοντέλα θεμάτων.

Ο αλγόριθμος ανάλυσης κειμένων LDA αποτελεί ένα πιθανοτικό μοντέλο που επιτρέπει την ερμηνεία κάποιων δεδομένων μέσα από κάποια σύνολα παρατηρήσεων. Για παράδειγμα, αν οι παρατηρήσεις είναι λέξεις που έχουν συλλεχθεί από κάποια κείμενα, ο αλγόριθμος θεωρεί ότι τα κείμενα αντιπροσωπεύονται από τυχαίες προσμεϊξεις κρυφών θεμάτων, όπου κάθε θέμα χαρακτηρίζεται από μία κατανομή ως προς τις λέξεις. Ο LDA είναι ένα παράδειγμα μοντέλου θεμάτων και παρουσιάστηκε για πρώτη φορά ως γενετικό μοντέλο ανακάλυψης θεμάτων από τους David Blei, Andrew Ng και Michael Jordan το 2002. Κύριο πλεονέκτημα του LDA, καθώς και των παραλλαγών του, είναι ο μεγάλος βαθμός προσαρμογής τους. Έτσι, μπορούν να εφαρμοστούν και σε διάφορα άλλα προβλήματα που αντικείμενα απασχόλησης δεν είναι κείμενα λέξεων. Για παράδειγμα, ο αλγόριθμος έχει χρησιμοποιηθεί στο πεδίο της μηχανική όρασης για ανάλυση εικόνας, στο πεδίο της βιοπληροφορικής για ανάλυση και ανάκτηση γενετικού κώδικα σε δεδομένα ερευνών κ.α. Στόχος μας είναι να βρούμε σύντομες περιγραφές των μελών της συλλογής που επιτρέπουν μια αποτελεσματική επεξεργασία μεγάλων συλλογών, διατηρώντας τις απαραίτητες στατιστικές σχέσεις που είναι χρήσιμες για βασικές διεργασίες όπως η περίληψη κειμένου.

Το LDA μοντέλο απεικονίζεται ως πιθανοτικό γραφικό μοντέλο στην εικόνα 2.3. Όπως γίνεται αντιληπτό από την εικόνα, υπάρχουν τρία επίπεδα στο LDA. Οι παράμετροι α και β είναι επιπέδου σώματος παράμετροι, που έχουν συλλεχθεί από την διαδικασία δημιουργίας του σώματος. Οι μεταβλητές θ είναι επιπέδου κειμένου μεταβλητές, που έχουν ληφθεί από ένα κείμενο. Τέλος, οι μεταβλητές z και w είναι επιπέδου λέξεων μεταβλητές και έχουν συλλεχθεί για κάθε λέξη κάθε κειμένου. Οι μεταβλητές M , N



Σχήμα 2.3: Γραφικό μοντέλο απεικόνισης του LDA. Τα πλαίσια απεικονίζουν τις επαναλήψεις. Το εξωτερικό πλαίσιο απεικονίζει τα κείμενα, ενώ το εσωτερικό την επαναληπτική επιλογή των θεμάτων και των λέξεων μέσα σε ένα κείμενο.

δηλώνουν αντίστοιχα τον αριθμό των κειμένων και τον αριθμό των λέξεων μέσα σε ένα κείμενο.

Είναι απαραίτητο να ξεχωρίσουμε το LDA από το απλό Dirichlet πολυωνυμικό μοντέλο ομαδοποίησης. Ένα κλασσικό μοντέλο ομαδοποίησης θα ήταν ένα δύο-επιπέδων μοντέλο, όπου η Dirichlet θα υπολογιζόταν μία φορά για ένα σώμα, μια πολυωνυμική μεταβλητή ομαδοποίησης θα επιλεγόταν για κάθε κείμενο του σώματος και ένα σύνολο λέξεων θα επιλεγόταν για ένα κείμενο της μεταβλητής ομάδας. Όπως και με πολλά αλλά μοντέλα ομαδοποίησης, ένα τέτοιο μοντέλο περιορίζει ένα κείμενο στο να συσχετίζεται μόνο με ένα θέμα. Αντιθέτως, το LDA είναι μοντέλο τριών επιπέδων και ο θεματικός κόμβος επιλέγεται επαναλαμβανόμενα από κάθε κείμενο. Με αυτό το μοντέλο τα κείμενα μπορούν να συσχετιστούν με πολλά θέματα.

Στο σημείο αυτό, έχοντας αναλύσει το θεωρητικό υπόβαθρο γύρω από τα πιθανοτικά θεματικά μοντέλα και συγκεκριμένα το μοντέλο LDA, προχωράμε στην εφαρμογή του LDA σε κάθε cluster (κατηγορία) άρθρων, σε κάθε group εντός των clusters, καθώς και σε κάθε μεμονωμένο άρθρο. Έτσι, καταλήγουμε με το διάνυσμα της 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. Κατά την εφαρμογή του αλγορίθμου μπορούμε να επιλέξουμε τον αριθμό αντιπροσωπευτικών λέξεων απ' τις οποίες θα αποτελείται ένα τέτοιο διάνυσμα.

2.3.3 Υποσύστημα Δημιουργίας Προφίλ Χρήστη

Ένα από τα ιδιαίτερα χαρακτηριστικά ενός συστήματος δημιουργίας εξατομικευμένων συστάσεων είναι το Προφίλ Χρήστη (στο πλαίσιο των συστημάτων συστάσεων). Το

Προφίλ Χρήστη είναι μια αναπαράσταση των πληροφοριών που υπάρχουν για ένα χρήστη, οι οποίες είναι απαραίτητες για ένα προσαρμοζόμενο σύστημα ώστε αυτό να προσφέρει εξατομίκευση, δηλαδή να συμπεριφέρεται διαφορετικά για διαφορετικούς χρήστες.

Παραδοσιακά, το προφίλ ενός χρήστη μπορεί να καθοριστεί παρακολουθώντας τα άρθρα τα οποία έχουν διαβαστεί από το χρήστη μέχρι στιγμής (ιστορικό περιήγησης), με βάση το περιεχόμενο των άρθρων. Ο τρόπος αναπαράστασης του προφίλ εξαρτάται από τη μέθοδο που χρησιμοποιεί το σύστημα πρόσβασης. Ουσιαστικά, το προφίλ αποτελεί ένα σύνολο των πιθανών ενδιαφερόντων του χρήστη. Παρ'όλ'αυτά, η προσέγγιση αυτή δεν αποτυπώνει αποτελεσματικά τις ακριβείς αναγνωστικές προτιμήσεις του χρήστη, καθώς το ενδιαφέρον του ενδέχεται να επηρεάζεται και από τις προτιμήσεις άλλων χρηστών. Επιπρόσθετα, η ανεπάρκεια της προσέγγισης στηριζόμενης αποκλειστικά στο περιεχόμενο των άρθρων οφείλεται και στο γεγονός ότι πολλοί χρήστες τείνουν να επιλέγουν άρθρα επηρεασμένοι από φράσεις/ονοματισμένες οντότητες όπως οι εξής: Τι συνέβη, ποιος εμπλέκεται, πότε συνέβη κ.λπ.

Στηριζόμενοι στην παραπάνω ανάλυση, το Προφίλ Χρήστη παραμετροποιείται μέσω μιας τριπλέτας $U = \langle T, P, E \rangle$ αποτελούμενης από τα εξής χαρακτηριστικά:

- Το T αναπαριστά την κατανομή των θεμάτων των ειδησεογραφικών άρθρων τα οποία ο χρήστης έχει διαβάσει στο παρελθόν, με μορφή διανύσματος θεμάτων $\{\langle t1, w1 \rangle, \langle t2, w2 \rangle, \dots\}$, όπου κάθε καταχώρηση του διανύσματος αποτελείται από μια αντιπροσωπευτική λέξη και το αντίστοιχο βάρος.
- Το P δηλώνει μια λίστα χρηστών $\{\langle u1, u2, \dots \rangle\}$ οι οποίοι έχουν παρόμοια πρότυπα πρόσβασης.
- Το E είναι μια λίστα από ονοματισμένες οντότητες $\{\langle e1, e2, \dots \rangle\}$ εξαγόμενες από το ιστορικό αναγνωσμένων άρθρων του χρήστη, συσχετισμένες με τον αντίστοιχο τύπο οντότητας.

Τα παραπάνω χαρακτηριστικά προφανώς αλληλεπιδρούν μεταξύ τους. Η κατανομή θεμάτων των ειδησεογραφικών άρθρων που εξάγονται απ'το ιστορικό ανάγνωσης του χρήστη είναι πολύ πιθανό να συνδέεται με τη λίστα ονοματισμένων οντοτήτων του προφίλ, ενώ αυτά τα δύο χαρακτηριστικά ενδέχεται να συνεισφέρουν και στα παρόμοια πρότυπα πρόσβασης δύο χρηστών.

Κάθε ενέργεια που πραγματοποιεί ο χρήστης στο δικτυακό τόπο καταγράφεται στο ιστορικό του, η ανάλυση του οποίου οδηγεί στην εξαγωγή συμπερασμάτων

προκειμένου το σύστημα να μπορεί να διαμορφώνει αυτόματα το Προφίλ Χρήστη, ώστε να εφαρμοστούν με μεγαλύτερη ακρίβεια οι τεχνικές εξατομίκευσης και κατ'επέκταση οι προτεινόμενες συστάσεις.

Παρακάτω αναλύουμε λεπτομερώς τις τεχνικές που υιοθετήθηκαν για την κατασκευή διαφορετικών πτυχών του προφίλ του χρήστη:

- **Σύνοψη Θεματικού Περιεχομένου Αναγνωσμένων Άρθρων:**

Στο σύστημα συστάσεων που υλοποιήσαμε συνοψίζουμε τα ειδησεογραφικά άρθρα του ιστορικού του χρήστη ως κατανομή θεμάτων, χρησιμοποιώντας την ίδια στρατηγική που εφαρμόσαμε και για τον εντοπισμό ειδησεογραφικών θεμάτων μέσω πιθανοτικών θεματικών μοντέλων στα clusters άρθρων της βάσης δεδομένων του συστήματος. Ουσιαστικά, χρησιμοποιούμε την ίδια αναπαράσταση του ιστορικού αναγνωσμένων άρθρων του χρήστη με αυτή των clusters άρθρων.

- **Ανάλυση Προτύπων Πρόσβασης:**

Στην πραγματικότητα πολλοί αναγνώστες διαδικτυακών άρθρων επιδεικνύουν παρόμοιες αναγνωστικές προτιμήσεις. Το προφίλ ενός χρήστη μπορεί να εμπλουτιστεί αναλύοντας τις αναγνωστικές προτιμήσεις άλλων χρηστών παρόμοιων με το δεδομένο χρήστη και ενσωματώνοντάς τις σε αυτό. Συγκεκριμένα, αναλύουμε το ιστορικό αναγνωσμένων άρθρων όλων των χρηστών του συστήματος. Υποθέτοντας ότι οι συλλογές άρθρων που αναγνώστηκαν από τους χρήστες A και B είναι οι N_A και N_B , η ανά ζεύγος ομοιότητα των προτύπων πρόσβασης ορίζεται ως η Jaccard ομοιότητα μεταξύ των συνόλων N_A και N_B . Η ομοιότητα Jaccard για δύο σύνολα είναι το μέγεθος της τομής προς το μέγεθος της ένωσής τους [24].

Υπολογίζοντας τις ανά ζεύγος ομοιότητες μεταξύ των χρηστών του συστήματος, δημιουργούμε ένα διάγραμμα ομοιοτήτων, κάθε εγγραφή του οποίου αποτελεί την Jaccard ομοιότητα μεταξύ των προτύπων πρόσβασης δύο χρηστών. Δεδομένου ενός χρήστη u , κάθε άλλος χρήστης μπορεί να θεωρηθεί όμοιος του αν το ανά ζεύγος σκορ ομοιότητάς τους ξεπερνά ένα προκαθορισμένο κατώφλι t_u^2 .

Η ανάλυση προτύπων πρόσβασης οδηγεί στην εξαγωγή μιας λίστας με τα ονόματα παρόμοιων χρηστών για κάθε χρήστη. Η πληροφορία αυτή αποθηκεύεται στη βάση δεδομένων και ενημερώνεται καθ'όλη τη διάρκεια περιήγησης του κάθε χρήστη στο σύστημα, καθώς κάθε νέο άρθρο που επιλέγεται προς ανάγνωση, μετατρέπει άμεσα τον εκάστοτε χρήστη σε λιγότερο ή περισσότερο όμοιο με κάποιον άλλο χρήστη.

- **Εξαγωγή Ονοματισμένων Οντοτήτων:**

Τυπικά, στα ειδησεογραφικά άρθρα οι ονοματισμένες οντότητες περιλαμβάνουν

λέξεις/φράσεις όπως “πότε, πού, τι συνέβη, ποιος εμπλέκεται” κ.λπ. Οι αναγνώστες ενδέχεται να έχουν προτίμηση σε κάποιες συγκεκριμένες ονοματισμένες οντότητες ενός άρθρου. Συνεπώς, οι οντότητες αυτές είναι σημαντικές για ένα σύστημα που θέλει να προσφέρει εξατομικευμένες συστάσεις σε χρήστες.

Για την εξαγωγή των ονοματισμένων οντοτήτων χρησιμοποιήσαμε το ανοιχτού κώδικα εργαλείο επεξεργασίας φυσικής γλώσσας GATE (General Architecture for Text Engineering), το οποίο αναγνωρίζει αυτόματα ονοματισμένες οντότητες σε κείμενα, δεδομένων κάποιων κανόνων. Για τη συγκεκριμένη διαδικασία χρησιμοποιήσαμε τους προεπιλεγμένους κανόνες του εργαλείου.

Κατά την προεπεξεργασία των κειμένων, έχοντας εξάγει το κείμενο κάθε άρθρου της βάσης δεδομένων σε αρχείο κειμένου, φορτώνουμε όλα τα αρχεία στο GATE και προχωρούμε στη δημιουργία ενός corpus κειμένων. Τρέχουμε το dictionary-based Named Entity Recognition εργαλείο ANNIE Gazetteer με είσοδο το corpus κειμένων και επιλέγουμε την εξαγωγή του κειμένου κάθε άρθρου σε μορφή .xml [28]. Με κατάλληλη επεξεργασία των .xml αρχείων κρατάμε την πληροφορία των ετικετών που μας ενδιαφέρουν (ονόματα, τοποθεσίες και οργανισμοί), κάνοντας χρήση της βιβλιοθήκης BeautifulSoup [3] της Python.

Η πληροφορία που εξάγεται για κάθε άρθρο, δηλαδή το όνομα της οντότητας μαζί με τον αντίστοιχο τύπο της (Organization, Person, Location), αποθηκεύεται στη βάση δεδομένων του συστήματος. Μετά την αναγνώριση των οντοτήτων, κάθε άρθρο συσχετίζεται με μία λίστα από ονοματισμένες οντότητες μαζί με τον αντίστοιχο τύπο. Έτσι, είναι εύκολο να εξάγουμε τις ονοματισμένες οντότητες που αφορούν τον εκάστοτε χρήστη, ανατρέχοντας κάθε φορά στο ιστορικό αναγνωσμένων άρθρων, εξάγοντας τις αποθηκευμένες οντότητες για τα άρθρα αυτά και ανανεώνοντας την εν λόγω λίστα με τις οντότητες των άρθρων αυτών.

2.3.4 Υποσύστημα Προσωποποιημένων Συστάσεων

Στο τελευταίο στάδιο του προτεινόμενου συστήματος παρέχονται εξατομικευμένες συστάσεις σε μεμονωμένους χρήστες. Η εξατομικευμένη σύσταση ειδησεογραφικών άρθρων στηρίζεται στη διερεύνηση της σχέσης μεταξύ πρόσφατα δημοσιευμένων άρθρων και του προφίλ του χρήστη. Στην παρούσα διπλωματική εργασία προτάθηκε η χρήση μιας υβριδικής μεθόδου συστάσεων. Ωστόσο, η διαφοροποίηση από προηγούμενες προσεγγίσεις έγκειται στην πρόταση για διεπίπεδη ιεραρχία συστάσεων, όπου το πρώτο επίπεδο δείχνει μια σύντομη σύνοψη για κάθε κατηγορία θεμάτων που μπορεί να προτιμά ο χρήστης και το δεύτερο επίπεδο δίνει μια συγκεκριμένη λίστα με άρθρα ειδήσεων παρόμοια με το αναγνωστικό ενδιαφέρον του χρήστη.

Αντιστοίχιση Αναγνωστικών Προτιμήσεων για την Αναπαράσταση 1ου Επιπέδου

Αφού δημιουργήσουμε την ιεραρχία των ειδησεογραφικών άρθρων (συσταδοποίηση σε clusters και groups εντός αυτών), καθώς και το προφίλ του χρήστη, το πρώτο επίπεδο αναπαράστασης μπορεί να επιτευχθεί με τη διαδοχική αντιστοίχιση του προφίλ του χρήστη στην ιεραρχία ειδήσεων και την επιλογή των κατάλληλων clusters. Σημειώνουμε ότι κάθε cluster αντιστοιχίζεται σε μία κατηγορία θεμάτων.

Τυπικά, η κατανομή θεμάτων αναπαρίσταται ως ένα διάνυσμα θεμάτων $T = \{\langle t1, w1 \rangle, \langle t2, w2 \rangle, \dots\}$. Για να εξασφαλίσουμε ότι όλα τα διανύσματα θεμάτων έχουν την ίδια διάσταση, δημιουργούμε ένα λεξιλόγιο θεμάτων V βασισμένο στα ήδη υπάρχοντα θέματα, όπου $|V|$ είναι ο συνολικός αριθμός από αντιπροσωπευτικές λέξεις. Κάθε διάσταση αντιστοιχεί σε μία ξεχωριστή λέξη. Αν η λέξη υπάρχει μέσα στο κείμενο, η τιμή του διανύσματος δεν είναι μηδενική. Συγκρίνουμε το βαθμό ομοιότητας της κατανομής θεμάτων κάθε ομάδας, T_C , με αυτή του προφίλ του χρήστη, T_U , μέσω της ομοιότητας συνημιτόνου (cosine similarity) [23].

Ομοιότητα Συνημιτόνου: Οι προς σύγκριση κατανομές θεμάτων αναπαρίστανται ως διανύσματα σε ένα πολυδιάστατο χώρο. Η μετρική ομοιότητας συνημιτόνου υπολογίζει την ομοιότητα από το συνημίτονο της γωνίας που σχηματίζουν τα δύο διανύσματα. Η ελάχιστη τιμή της μετρικής είναι -1 , υποδηλώνοντας την απόκλιση των διανυσμάτων και η μέγιστη 1 , υποδηλώνοντας την απόλυτη ταύτιση. Όταν τα διανύσματα είναι κάθετα μεταξύ τους και σχηματίζουν γωνία 90° , το συνημίτονο είναι 0 , υποδηλώνοντας ότι τα διανύσματα είναι ανεξάρτητα.

Η εξίσωση που εφαρμόζουμε για τον υπολογισμό της μετρικής αυτής είναι η ακόλουθη:

$$Sim(\mathbf{T}_C, \mathbf{T}_U) = \frac{\mathbf{T}_C \cdot \mathbf{T}_U}{\|\mathbf{T}_C\| \|\mathbf{T}_U\|} \quad (2.1)$$

όπου $|\mathbf{T}_C| = |\mathbf{T}_U| = |V|$, και $\|\mathbf{T}_C\|, \|\mathbf{T}_U\|$ αποτελούν τις 12-νόρμες.

Η κατάταξη των clusters βασίζεται στο σκορ ομοιότητας που υπολογίζεται από την εξίσωση (2.1). Γενικά, οι χρήστες τείνουν να προτιμούν συγκεκριμένες κατηγορίες θεμάτων, χωρίς να ενδιαφέρονται για όλα τα θέματα. Ως εκ τούτου, επιλέγουμε τα clusters με σκορ ομοιότητας μεγαλύτερο από ένα δυναμικό κατώφλι. Αφού επιλέξουμε τα κατάλληλα clusters, εισχωρούμε σε κάθε cluster και επιλέγουμε το group νέων που είναι πιο κοντά σε ομοιότητα με τις προτιμήσεις του χρήστη, κάνοντας χρήση της ίδιας στρατηγικής μέσω της οποίας επιλέξαμε και τα clusters προηγουμένως. Έτσι,

δημιουργείται μια λίστα από group άρθρων (ένα group από κάθε cluster), η οποία και επιλέγεται ως η βάση για το δεύτερο επίπεδο σύστασης.

Αντιστοίχιση Αναγνωστικών Προτιμήσεων για την Αναπαράσταση 2ου Επιπέδου

Έχοντας εξασφαλίσει τα group άρθρων που πιθανότατα ενδιαφέρουν το χρήστη, το ακόλουθο βήμα είναι η επιλογή συγκεκριμένων άρθρων προς αναπαράσταση στο χρήστη. Αρχικά, διαμορφώνουμε ένα προφίλ για κάθε ειδησεογραφικό άρθρο και εν συνεχεία, μοντελοποιούμε τις προσωποποιημένες συστάσεις ως ένα “προϋπολογισμένο πρόβλημα μέγιστης κάλυψης” (budgeted maximum coverage problem) [20] και το επιλύουμε μέσω ενός άπληστου (greedy) προσεγγιστικού αλγορίθμου. [10]

Δημιουργία Προφίλ Ειδησεογραφικών Άρθρων

Ένα ειδησεογραφικό άρθρο αποτελείται από στατικά χαρακτηριστικά (π.χ. κατανομή θεμάτων, ονοματισμένες οντότητες) και δυναμικά χαρακτηριστικά (π.χ. χρήστες που το διάβασαν, δημοφιλία, πόσο “φρέσκο” είναι από τη σκοπιά του πόσο πρόσφατα δημοσιεύθηκε). Σχετικά με τα στατικά χαρακτηριστικά, η κατανομή θεμάτων κάθε άρθρου καθώς και οι ονοματισμένες οντότητές του έχουν εξαχθεί σε προηγούμενο βήμα της εφαρμογής και βρίσκονται ήδη αποθηκευμένα σε αντίστοιχους πίνακες της βάσης δεδομένων του συστήματος. Σχετικά με τη δημοφιλία, υπολογίζουμε το λόγο των χρηστών που το διάβασαν προς το συνολικό αριθμό χρηστών του συστήματος. Σχετικά με το πόσο πρόσφατα δημοσιεύθηκε, υπολογίζουμε τη διαφορά μεταξύ της ημερομηνίας δημοσίευσης και της τρέχουσας ημερομηνίας.

Στο σύστημά μας, τα προφίλ των άρθρων είναι ιδιαίτερα βοηθητικά στο να συγκρίνουμε μεταξύ τους δύο άρθρα και να μπορέσουμε να αξιολογήσουμε σε τι βαθμό ικανοποιεί το καθένα εξ αυτών τις αναγνωστικές προτιμήσεις του χρήστη. Οι δύο παραπάνω συγκρίσεις, αυτή μεταξύ προφίλ άρθρων και αυτή μεταξύ προφίλ άρθρου και προφίλ χρήστη υπολογίζονται μέσω της ίδια φόρμουλας.

Δεδομένων ενός προφίλ άρθρου $F_n = \langle T_n, P_n, E_n \rangle$ και ενός προφίλ χρήστη $F_u = \langle T_u, P_u, E_u \rangle$, η ομοιότητα μεταξύ των F_n και F_u υπολογίζεται ως εξής:

$$Sim(F_n, F_u) = \frac{\alpha Sim(T_n, T_u) + \beta Sim(P_n, P_u) + \gamma Sim(E_n, E_u)}{\sqrt{\alpha^2 + \beta^2 + \gamma^2}},$$

όπου a , b και c είναι παράμετροι μέσω των οποίων ρυθμίζουμε το πόσο εμπιστευόμαστε τα αντίστοιχα μέρη. Στο δικό μας σύστημα συστάσεων επιλέγουμε να αναθέσουμε

ιση σημαντικότητα σε κάθε έναν από τους παραπάνω παράγοντες, δεδομένου ότι αναπαριστούν διαφορετικές αλλά, παράλληλα, συνδεδεμένες πτυχές ενός ειδησεογραφικού άρθρου ή ενός προφίλ χρήστη, αναθέτοντας και στους τρεις παράγοντες την ίδια τιμή $a = b = c = 1$.

Η $Sim(T_C, T_U)$ υπολογίζεται μέσω της εξίσωσης (2.1), ενώ οι $Sim(P_n, P_u)$ και $Sim(E_n, E_u)$ υπολογίζονται μέσω της ομοιότητας Jaccard. Η ομοιότητα Jaccard [24] για δύο σύνολα είναι το μέγεθος της τομής προς το μέγεθος της ένωσής τους.

Εισαγωγή στις Submodular Συναρτήσεις

Μια κατηγορία συναρτήσεων που χρησιμοποιούνται ευρέως και παρουσιάζουν ιδιαίτερο ενδιαφέρον είναι οι submodular συναρτήσεις. Οι συναρτήσεις αυτές προσπαθούν να μοντελοποιήσουν το γεγονός ότι όσο συνεχίζουμε να δίνουμε αγαθά σε κάποιον χρήστη, επέρχεται ένας κορεσμός στην ωφέλειά του. Η αύξηση της ωφέλειας από την απόκτηση νέων αγαθών είναι μικρή όταν ο χρήστης έχει ήδη στην κατοχή του κι άλλα αγαθά. Έχουμε, λοιπόν, μια μονότονη (φθίνουσα) συμπεριφορά στην αύξηση της ωφέλειας.

Ορισμός. Μια συνάρτηση ωφέλειας f για έναν χρήστη i είναι submodular αν για οποιαδήποτε δύο υποσύνολα αντικειμένων S, T , με $S \subseteq T$, και για κάθε αντικείμενο $j \notin T$, ισχύει ότι

$$f(T \cup \{s\}) - f(T) \leq f(S \cup \{s\}) - f(S)$$

Ο παραπάνω ορισμός μάς λέει ότι η αύξηση της ωφέλειας που προκαλεί η προσθήκη του αγαθού s στον χρήστη i είναι μεγαλύτερη όταν προσθέτουμε το s σε ένα σύνολο αγαθών S , από ότι όταν το προσθέτουμε σε ένα μεγαλύτερο σύνολο του S (το T). Πάρα πολλά μικροοικονομικά μοντέλα στηρίζονται στην υπόθεση ότι οι εμπλεκόμενοι χρήστες έχουν submodular συμπεριφορά. Κλασικό παράδειγμα αποτελούν και τα κοινωνικά δίκτυα, όπου η προσθήκη ενός νέου φίλου θα αυξήσει περισσότερο την κοινωνική επιρροή σε μία λιγότερο κοινωνική ομάδα από ό,τι σε μία πιο κοινωνική.

Μία πιο μαθηματική διατύπωση για το πρόβλημα της μεγιστοποίησης της επιρροής είναι η εξής: δεδομένου ενός συνόλου στοιχείων E , όπου κάθε στοιχείο είναι συσχετισμένο με μία τιμή επιρροής και ένα κόστος, και δεδομένου, επίσης, ενός budget B άρθρων, ο σκοπός είναι να βρεθεί ένα υποσύνολο του E όπου θα μεγιστοποιεί την επιρροή (δηλαδή την συνάρτηση $f(S)$) χωρίς το συνολικό κόστος να ξεπερνά το budget B . Ως budget B μπορεί να θεωρηθεί ο μέγιστος αριθμός από προτεινόμενα άρθρα μέσα σε

κάθε group άρθρων. Το πρόβλημα της μεγιστοποίησης της επιρροής είναι δυστυχώς NP-hard. Ωστόσο, το επιλύουμε μέσω ενός άπληστου (greedy) προσεγγιστικού αλγορίθμου, ο οποίος επιλέγει διαδοχικά το στοιχείο το οποίο αυξάνει τη μέγιστη δυνατή επιρροή εντός του ορίου κόστους. Δηλαδή προσθέτουμε κάθε φορά στο S το άρθρο που σε αυτό το σημείο μεγιστοποιεί το οριακό κέρδος (marginal gain). Ο αλγόριθμος αυτός αποδεικνύεται ότι αποτελεί $(1 - \frac{1}{e})$ -προσέγγιση του βέλτιστου, δηλαδή προσέγγιση του βέλτιστου κατά 63%.

Algorithm 1 Άπληστος προσεγγιστικός αλγόριθμος

Start with an empty set S

for B iterations **do**

 Add article ς to S so that it maximizes $I(\varsigma) = f(S \cup \{\varsigma\}) - f(S)$

end for

Submodularity Μοντέλο Συστάσεων

Με βάση τη συγκεκριμένη στρατηγική, ορίζουμε μία συνάρτηση ποιότητας f για να αξιολογήσουμε το επιλεγμένο σετ άρθρων S σε σχέση με ολόκληρο το group νέων N ως εξής:

$$f(S) = \frac{1}{|N \setminus S| \cdot |S|} \sum_{n_1 \in N \setminus S} \sum_{n_2 \in S} \text{sim}(n_1, n_2) \\ + \frac{1}{\binom{|S|}{2}} \sum_{\substack{n_1, n_2 \in S \\ n_1 \neq n_2}} -\text{sim}(n_1, n_2) + \frac{1}{|S|} \sum_{n_1 \in S} \text{sim}(u, n_1),$$

όπου n_1 και n_2 υποδηλώνουν άρθρα, το u αναπαριστά τον δεδομένο χρήστη και η $\text{sim}(\cdot, \cdot)$ αναπαριστά την ομοιότητα μεταξύ των δύο προφίλ, είτε αναφερόμαστε σε προφίλ χρήστη είτε σε προφίλ άρθρου.

Η παραπάνω εξίσωση αποτελείται από τρία μέρη:

Το πρώτο στοχεύει στην εκτίμηση της ποιότητας του πόσο αντιπροσωπευτικό είναι το επιλεγμένο σετ νέων S , το δεύτερο παρέχει μία εικόνα για το πόσο ποικίλα είναι τα θέματα που κρύβονται στα επιλεγμένα άρθρα και τέλος, το τρίτο μέρος μάς δίνει στοιχεία για το πόσο ικανοποιούνται οι προτιμήσεις του χρήστη από το επιλεγμένο σετ S . Η συνάρτηση $f(S)$ εξισορροπεί τη συνεισφορά κάθε ενός μέρους. Και τα τρία αυτά μέρη αποτελούν submodular συναρτήσεις, άρα είναι μη αρνητικές και μονότονες, επομένως και η συνάρτηση $f(S)$ θα είναι submodular.

Υποθέτοντας ότι το άρθρο ς είναι το υποψήφιο άρθρο, η αύξηση της ωφέλειας αναπαρίσταται ως εξής:

$$I(s) = f(S \cup \{s\}) - f(S).$$

Στόχος είναι να βρεθεί μία λίστα από άρθρα η οποία μεγιστοποιεί το οριακό κέρδος εντός του δοσμένου budget. Έτσι, μοντελοποιούμε τις προσωποποιημένες συστάσεις άρθρων ως ένα “προϋπολογισμένο πρόβλημα μέγιστης κάλυψης”.

Αρχικά, φροντίζουμε να αφαιρέσουμε από κάθε επιλεγμένο group άρθρων τα άρθρα τα οποία βρίσκονται ήδη στο ιστορικό ανάγνωσης του χρήστη, καθώς άρθρα που έχουν ήδη αναγνωσθεί δεν πρέπει να βρεθούν στην τελική προτεινόμενη λίστα. Έτσι, καταλήγουμε με μία λίστα από λίστες, κάθε μία εκ των οποίων περιλαμβάνει τα εναπομείναντα υποψήφια προς πρόταση άρθρα από κάθε group.

Στη συνέχεια, υπολογίζουμε το budget για κάθε λίστα, δηλαδή το μέγιστο αριθμό προτεινόμενων άρθρων από κάθε κατηγορία που επιλέχθηκε ως ταιριαστή με τις προτιμήσεις του χρήστη. Το budget υπολογίζεται με βάση την ομοιότητα κάθε χρήστη με το εν λόγω cluster/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.

Στο σημείο αυτό προβήκαμε σε μία παραδοχή: Η εφαρμογή του άπληστου αλγορίθμου (Algorithm 1) σε κάθε group άρθρων (προκειμένου να επιλύσουμε το πρόβλημα μέγιστης κάλυψης, επιλέγοντας διαδοχικά το άρθρο που προσφέρει τη μέγιστη αύξηση ωφέλειας από το επιλεγμένο σετ S , μέσα στο πλαίσιο που ορίζεται από το budget), αποδείχθηκε υπερβολικά χρονοβόρα για ένα τυπικό υπολογιστικό σύστημα και ξέφευγε από την ουσία ενός συστήματος συστάσεων. Ο χρήστης θα έπρεπε να περιμένει υπερβολικά πολλή ώρα προκειμένου να δεχθεί ως σύσταση τη βέλτιστη επιλογή άρθρων, όπως θα μας την υποδείκνυε η συνάρτηση αξιολόγησης. Έτσι, καταφύγαμε στη λύση του να υπολογίσουμε για κάθε group όλους τους δυνατούς συνδυασμούς άρθρων, μεγέθους ίσου με το αντίστοιχο budget, και να επιλέξουμε τυχαία έναν από αυτούς μέσω της συνάρτησης rand. Αν, για παράδειγμα, το budget για ένα group είναι ίσο με την τιμή τρία, τότε υπολογίζουμε όλες τις πιθανές τριάδες άρθρων.

Για να ενσωματώσουμε τα προτεινόμενα άρθρα από κάθε group στην τελική προτεινόμενη λίστα, επιλέγουμε, λοιπόν, μία τυχαία n -άδα από κάθε group και όχι απαραίτητα αυτή με την υψηλότερη κατάταξη.

Η submodularity-based στρατηγική επιλογής άρθρων οδηγεί σε μία αρκετά ποικίλη λίστα άρθρων από κάθε κατηγορία.

Προσαρμογή Κατάταξης Ειδησεογραφικών Άρθρων

Υιοθετώντας τον άπληστο αλγόριθμο που αναλύσαμε νωρίτερα, μπορούμε να εξάγουμε

μία λίστα από ειδησεογραφικά άρθρα για κάθε θεματική κατηγορία. Λαμβάνοντας υπόψιν τα αποκλειστικά χαρακτηριστικά των άρθρων, όπως η δημοφιλία και το πόσο πρόσφατα δημοσιευμένα είναι, η κατάταξη των επιλεγμένων άρθρων χρειάζεται να προσαρμοστεί ώστε να κάνουμε το προτεινόμενο αποτέλεσμα πιο λογικό.

Στο σύστημά μας η δημοφιλία ενός άρθρου και το πόσο πρόσφατα δημοσιεύθηκε αποτελούν μέρος του Προφίλ Άρθρου, τη δημιουργία του οποίου αναλύσαμε ενδελεχώς σε προηγούμενη παράγραφο. Κάνοντας προσαρμογές στη λίστα επιλεγμένων άρθρων, συνδυάζουμε τις κανονικοποιημένες τιμές αυτών των δύο τύπων ιδιοτήτων. Τυπικά, δεδομένου ενός άρθρου n , η δημοφιλία n_P και το πόσες μέρες είναι δημοσιευμένο n_I μπορούν να συνδυαστούν ως εξής:

$$n_\phi = \frac{n_P - n_{P_{min}}}{n_{P_{max}} - n_{P_{min}}} - \frac{n_I - n_{I_{min}}}{n_{I_{max}} - n_{I_{min}}}.$$

Παρατηρούμε ότι όσο πιο πρόσφατα δημοσιεύθηκε ένα άρθρο (και άρα, όσο πιο μικρή η τιμή του εν λόγω χαρακτηριστικού), τόσο υψηλότερη θέση παίρνει το άρθρο στην τελική κατάταξη. Δεδομένης μια λίστας άρθρων προς σύσταση, επιλέγουμε διαδοχικά δύο γειτονικά άρθρα n_i και n_j από την κορυφή της λίστας προς τα κάτω και τα συγκρίνουμε ως προς το δυναμικό τους σκορ, n_f . Εάν η εν λόγω διαφορά είναι μεγαλύτερη από μηδέν, ανταλλάσσουμε τη θέση των δύο αυτών άρθρων. Αλλιώς, τα προσπερνούμε και συνεχίζουμε με τη σύγκριση του επόμενου ζεύγους άρθρων.

Μέσω αυτής της μικρής προσαρμογής η παραγόμενη κατάταξη δίνει έμφαση στα πιο δημοφιλή και “φρέσκα” ειδησεογραφικά άρθρα, καθώς, επίσης, επικεντρώνεται σε άρθρα που ικανοποιούν σε μεγαλύτερο βαθμό τις αναγνωστικές προτιμήσεις του χρήστη.

Κεφάλαιο 3

Τεχνολογίες Υλοποίησης

Στο κεφάλαιο αυτό θα περιγράψουμε τα εργαλεία και τις τεχνολογίες που χρησιμοποιήθηκαν για την υλοποίησή μας.

3.1 Η γλώσσα προγραμματισμού Python

Η γλώσσα προγραμματισμού Python [1], [18] είναι μια γλώσσα διερμηνευόμενη και object-oriented. Η σύλληψη της Python έγινε στα τέλη της δεκαετίας του 1980 και η εφαρμογή της έγινε το Δεκέμβριο του 1989 από τον Ολλανδό Guido Van Rossum, ο οποίος είναι και ο κύριος συγγραφέας της. Το όνομα της γλώσσας προέρχεται από την ομάδα Άγγλων κωμικών Monty Python. Ο κώδικάς της διανέμεται με την άδεια Python Software Foundation. Ανάμεσα στα κύρια χαρακτηριστικά της είναι τα εξής: γλώσσα ανοιχτού κώδικα (open source), αναγνωσιμότητα, εύκολη εκμάθηση, επεκτασιμότητα, εύκολη συντήρηση, δυνατότητα απλοποίησης στην υλοποίηση δύσκολων συναρτήσεων, γρήγορη ανάπτυξη εφαρμογών.

Η ίδια η γλώσσα είναι επεκτάσιμη καθώς ένα βασικό σύνολο της γλώσσας αποτελεί τον πυρήνα της, ενώ όλα τα υπόλοιπα είναι βιβλιοθήκες (modules) που επεκτείνουν την λειτουργικότητά της. Οι κύριοι τύποι δεδομένων που χρησιμοποιεί είναι οι λίστες, τα λεξικά και οι πλειάδες. Έχει μεγάλο εύρος εφαρμογών, όπως για παράδειγμα στον επιστημονικό υπολογισμό, στην τεχνητή νοημοσύνη, στην επεξεργασία φυσικής γλώσσας κλπ.

Στην παρούσα διπλωματική εργασία χρησιμοποιήσαμε την έκδοση 3.4.0 της γλώσσας, η οποία σε συνδυασμό με την πλατφόρμα NLTK χρησιμεύουν σε πολλές εφαρμογές της επεξεργασίας φυσικής γλώσσας.

3.2 MySQL

Η MySQL [16] είναι ένα ανοιχτού λογισμικού σύστημα διαχείρισης σχεσιακών βάσεων δεδομένων. Η ονομασία MySQL περιέχει δύο στοιχεία: το My είναι το όνομα της κόρης του συνιδρυτή του συστήματος, Monty Widenius και το SQL αναφέρεται στη γλώσσα SQL (Structured Query Language), μια γλώσσα υπολογιστών που αναπτύχθηκε ξεχωριστά από τις υλοποιήσεις συστημάτων διαχείρισης βάσεων δεδομένων (όπως της MySQL, της PostgreSQL, της Oracle κλπ). Το πρόγραμμα τρέχει έναν εξυπηρετητή (server) παρέχοντας πρόσβαση πολλών χρηστών σε ένα σύνολο βάσεων δεδομένων. Το λογισμικό αυτό επιτρέπει στους χρήστες να δημιουργούν και να χρησιμοποιούν μια βάση δεδομένων, παρέχοντάς τους δυνατότητες όπως ο ορισμός, η κατασκευή, η χρήση/προσπέλαση και η διαγραφή αυτής.

Θα μπορούσαμε να υλοποιήσουμε μια βάση δεδομένων και χωρίς τη χρήση συστήματος διαχείρισης (πχ. με χρήση αρχείων). Τα σημαντικότερα πλεονεκτήματα του συστήματος διαχείρισης είναι η ευκολία στη σχεδίαση και την υλοποίηση, η γρήγορη ανάπτυξη εφαρμογών, η ακεραιότητα των δεδομένων, ο έλεγχος πρόσβασης χρηστών, η ταυτόχρονη χρήση από πολλούς χρήστες, ο έλεγχος ορθότητας/πλεονασμών και οι έτοιμες συναρτήσεις/αλγόριθμοι.

Η MySQL χρησιμοποιείται ευρέως σε διάφορες εφαρμογές όπως οι TYPO3, Joomla, Wordpress, phpBB, MyBB, Drupal. Χρησιμοποιείται επίσης και σε κάποιες από τις πιο διαδεδομένες διαδικτυακές υπηρεσίες, όπως το Flickr, το YouTube, η Wikipedia, το Google, το Facebook και το Twitter.

Στην παρούσα διπλωματική εργασία χρησιμοποιήσαμε την έκδοση 5.6.33.

3.3 HTML

Η HTML (αρχικοποίηση του αγγλικού HyperText Markup Language [8], ελλ. Γλώσσα Σήμανσης Υπερκειμένου) είναι η κύρια γλώσσα σήμανσης για τις ιστοσελίδες, και τα στοιχεία της είναι τα βασικά δομικά στοιχεία των ιστοσελίδων.

Η HTML γράφεται υπό μορφή στοιχείων HTML τα οποία αποτελούνται από ετικέτες (tags), οι οποίες περικλείονται μέσα σε σύμβολα «μεγαλύτερο από» και «μικρότερο από», μέσα στο περιεχόμενο της ιστοσελίδας. Οι ετικέτες HTML συνήθως λειτουργούν ανά ζεύγη, με την πρώτη να ονομάζεται ετικέτα έναρξης και τη δεύτερη ετικέτα λήξης (ή σε άλλες περιπτώσεις ετικέτα ανοίγματος και ετικέτα κλεισίματος αντίστοιχα). Ανάμεσα στις ετικέτες οι σχεδιαστές ιστοσελίδων μπορούν να τοποθετήσουν κείμενο, πίνακες, εικόνες κλπ.

Ο σκοπός ενός web browser είναι να διαβάσει τα έγγραφα HTML και τα συνθέτει σε σελίδες που μπορεί κανείς να διαβάσει ή να ακούσει. Ο browser δεν εμφανίζει τις

ετικέτες HTML, αλλά τις χρησιμοποιεί για να ερμηνεύσει το περιεχόμενο της σελίδας. Τα στοιχεία της HTML χρησιμοποιούνται για να χτίσουν όλους τους ιστότοπους. Η HTML επιτρέπει την ενσωμάτωση εικόνων και άλλων αντικειμένων μέσα στη σελίδα και μπορεί να χρησιμοποιηθεί για να εμφανίσει διαδραστικές φόρμες. Παρέχει τις μεθόδους δημιουργίας δομημένων εγγράφων (δηλαδή εγγράφων που αποτελούνται από το περιεχόμενο που μεταφέρουν και από τον κώδικα μορφοποίησης του περιεχομένου) καθορίζοντας δομικά σημαντικά στοιχεία για το κείμενο, όπως κεφαλίδες, παραγράφους, λίστες, συνδέσμους, παραθέσεις και άλλα. Μπορούν, επίσης, να ενσωματώνονται σενάρια εντολών σε γλώσσες όπως η JavaScript, τα οποία επηρεάζουν τη συμπεριφορά των ιστοσελίδων HTML.

Οι Web browsers μπορούν επίσης να αναφέρονται σε στυλ μορφοποίησης CSS για να ορίζουν την εμφάνιση και τη διάταξη του κειμένου και του υπόλοιπου υλικού. Ο οργανισμός W3C, ο οποίος δημιουργεί και συντηρεί τα πρότυπα για την HTML και τα CSS, ενθαρρύνει τη χρήση των CSS αντί διαφόρων στοιχείων της HTML για σκοπούς παρουσίασης του περιεχομένου.

3.4 CSS

Η CSS (Cascading Style Sheets [22] - Διαδοχικά Φύλλα Στυλ ή αλληλουχία φύλλων στυλ) είναι μια γλώσσα υπολογιστή που ανήκει στην κατηγορία των γλωσσών φύλλων στυλ που χρησιμοποιείται για τον έλεγχο της εμφάνισης ενός εγγράφου που έχει γραφτεί με μια γλώσσα σήμανσης. Χρησιμοποιείται δηλαδή για τον έλεγχο της εμφάνισης ενός εγγράφου που γράφτηκε στις γλώσσες HTML και XHTML, δηλαδή για τον έλεγχο της εμφάνισης μιας ιστοσελίδας και γενικότερα ενός ιστοτόπου. Η CSS είναι μια γλώσσα υπολογιστή προορισμένη να αναπτύσσει στυλιστικά μια ιστοσελίδα, δηλαδή να διαμορφώνει περισσότερα χαρακτηριστικά, χρώματα, στοίχιση και δίνει περισσότερες δυνατότητες σε σχέση με την html. Η χρήση της CSS κρίνεται ως απαραίτητη για μια όμορφη και καλοσχεδιασμένη ιστοσελίδα .

3.5 Apache HTTP Server

Ο Apache HTTP Server [21], γνωστός και απλά σαν Apache, είναι ένας εξυπηρετητής του παγκόσμιου ιστού (web). Όποτε ένας χρήστης επισκέπτεται έναν ιστότοπο, το πρόγραμμα πλοήγησης (browser) επικοινωνεί με έναν διακομιστή (server) μέσω του πρωτοκόλλου HTTP, ο οποίος παράγει τις ιστοσελίδες και τις αποστέλλει στο πρόγραμμα πλοήγησης. Κυκλοφόρησε υπό την άδεια λογισμικού Apache και είναι λογισμικό ανοιχτού κώδικα.

Ο Apache χρησιμοποιείται και σε τοπικά δίκτυα σαν διακομιστής συνεργαζόμενος με

συστήματα διαχείρισης Βάσης Δεδομένων, π.χ. MySQL, όπως και χρησιμοποιήθηκε στην παρούσα διπλωματική (έκδοση 2.4.7).

3.6 Flask Web Framework

Τα Web Frameworks είναι σχεδιασμένα για να υποστηρίζουν την ανάπτυξη διαδικτυακών εφαρμογών. Κάθε εφαρμογή που απαιτεί τη χρήση βάσεων δεδομένων, φορμών, συνόδων (sessions), cookies ή κάποιας απομακρυσμένης υπηρεσίας (όπως είναι το Twitter, το Facebook κλπ) θα ωφεληθεί αν υλοποιηθεί με κάποιο Framework.

Το Flask [14] είναι ένα micro web framework γραμμένο σε Python.

3.7 Εργαλεία Επεξεργασίας Φυσικής Γλώσσας

3.7.1 Tree Tagger

Ο Tree Tagger [27] είναι ένα εργαλείο που δημιουργήθηκε από τον Helmut Schmid στο Ινστιτούτο για τον γλωσσικό υπολογισμό στο Πανεπιστήμιο της Στουτγκάρδης. Αποτελείται από ένα εργαλείο για σχολιασμό κειμένου με βάση το σε τι μέρος του λόγου αντιστοιχίζεται η κάθε λέξη, καθώς και πληροφορία για λήμματα. Ο Tree Tagger δουλεύει πολύ καλά για την Αγγλική γλώσσα, αλλά υποστηρίζει επίσης και άλλες γλώσσες όπως Ισπανικά, Γαλλικά, Ιταλικά, Ολλανδικά, Βουλγαρικά, Ρωσικά, Ελληνικά, Πορτογαλικά κ.λπ. Ο Tree Tagger περιλαμβάνει κατηγορίες στις οποίες ανήκουν οι λέξεις. Οι κατηγορίες αυτές αντιπροσωπεύουν στην ουσία τα μέρη του λόγου μιας πρότασης. Ένα παράδειγμα μιας τέτοιας κατηγορίας είναι το NN, το οποίο υποδηλώνει ουσιαστικό. Παρακάτω ακολουθεί ένας αναλυτικός πίνακας με τις κατηγορίες του Tree Tagger και τι σημαίνει η καθεμία.

3.7.2 NLTK

Το NLTK (Natural Language Toolkit) [4] είναι ένα πακέτο βιβλιοθηκών και προγραμμάτων της Python για εφαρμογές της Επεξεργασίας Φυσικής Γλώσσας και αναπτύχθηκε απ'τους Steven Bird, Edward Loper και Ewan Klein. Περιλαμβάνει πολλά γνωστά σώματα κειμένων, γραφικές αναπαραστάσεις και δειγματικά δεδομένα. Συνοδεύεται από ένα βιβλίο το οποίο εξηγεί τις έννοιες που σχετίζονται με τα εργαλεία που παρέχει. Βασικός στόχος του NLTK είναι να υποστηρίξει την έρευνα και την εκμάθηση της Επεξεργασίας Φυσικής Γλώσσας καθώς και άλλων σχετικών πεδίων, όπως η Γλωσσολογία, η Τεχνητή Νοημοσύνη, η Ανάκτηση Πληροφορίας και η Μηχανική Μάθηση. Όταν λέμε φυσική γλώσσα εννοούμε μια γλώσσα η οποία χρησιμοποιείται για την καθημερινή επικοινωνία των ανθρώπων. Σε αντίθεση

TreeTagger Tag Set (58 tags)

POSTag	Description	Example	POSTag	Description	Example
CC	coordinating conjunction	<i>and, but, or, &</i>	VB	verb <i>be</i> , base form	<i>be</i>
CD	cardinal number	<i>1, three</i>	VBD	verb <i>be</i> , past	<i>was /were</i>
DT	determiner	<i>the</i>	VBG	verb <i>be</i> , gerund/participle	<i>being</i>
EX	existential there	<i>there is</i>	VBN	verb <i>be</i> , past participle	<i>been</i>
FW	foreign word	<i>d'œuvre</i>	VBZ	verb <i>be</i> , pres, 3rd p. sing	<i>is</i>
IN	preposition/subord. conj.	<i>in, of, like, after, whether</i>	VBP	verb <i>be</i> , pres non-3rd p.	<i>am /are</i>
IN/that	complementizer	<i>that</i>	VD	verb <i>do</i> , base form	<i>do</i>
JJ	adjective	<i>green</i>	VDD	verb <i>do</i> , past	<i>did</i>
JJR	adjective, comparative	<i>greener</i>	VDG	verb <i>do</i> gerund/participle	<i>doing</i>
JJS	adjective, superlative	<i>greenest</i>	VDN	verb <i>do</i> , past participle	<i>done</i>
LS	list marker	<i>(1),</i>	VDZ	verb <i>do</i> , pres, 3rd per. sing	<i>does</i>
MD	modal	<i>could, will</i>	VDP	verb <i>do</i> , pres, non-3rd per.	<i>do</i>
NN	noun, singular or mass	<i>table</i>	VH	verb <i>have</i> , base form	<i>have</i>
NNS	noun plural	<i>tables</i>	VHD	verb <i>have</i> , past	<i>had</i>
NP	proper noun, singular	<i>John</i>	VHG	verb <i>have</i> , gerund/participle	<i>having</i>
NPS	proper noun, plural	<i>Vikings</i>	VHN	verb <i>have</i> , past participle	<i>had</i>
PDT	predeterminer	<i>both the boys</i>	VHZ	verb <i>have</i> , pres 3rd per. sing	<i>has</i>
POS	possessive ending	<i>friend's</i>	VHP	verb <i>have</i> , pres non-3rd per.	<i>have</i>
PP	personal pronoun	<i>I, he, it</i>	VV	verb, base form	<i>take</i>
PP\$	possessive pronoun	<i>my, his</i>	VVD	verb, past tense	<i>took</i>
RB	adverb	<i>however, usually, here, not</i>	VVG	verb, gerund/participle	<i>taking</i>
RBR	adverb, comparative	<i>better</i>	VVN	verb, past participle	<i>taken</i>
RBS	adverb, superlative	<i>best</i>	VVP	verb, present, non-3rd p.	<i>take</i>
RP	particle	<i>give up</i>	VVZ	verb, present 3d p. sing.	<i>takes</i>
SENT	end punctuation	<i>?, !, .</i>	WDT	wh-determiner	<i>which</i>
SYM	symbol	<i>@, +, *, ^, /, =</i>	WP	wh-pronoun	<i>who, what</i>
TO	to	<i>to go, to him</i>	WP\$	possessive wh-pronoun	<i>whose</i>
UH	interjection	<i>uhhuhhuhh</i>	WRB	wh-abverb	<i>where, when</i>
			:	general joiner	<i>,, , --</i>
			\$	currency symbol	<i>\$, £</i>

Σχήμα 3.1: Κατηγορίες Tree Tagger.

με τις τεχνητές γλώσσες, όπως οι γλώσσες προγραμματισμού και η μαθηματική σημειογραφία, οι φυσικές γλώσσες έχουν εξελιχθεί καθώς περνούν από γενιά σε γενιά και είναι αρκετά δύσκολο να μπορέσουν να οριστούν με ρητούς κανόνες. Έχει χρησιμοποιηθεί με επιτυχία ως εργαλείο διδασκαλίας, μελέτης και ως πλατφόρμα για την ανάπτυξη πρωτότυπων ερευνητικών συστημάτων.

Το NLTK δημιουργήθηκε με βάση τέσσερις πρωταρχικούς σκοπούς:

- **Απλότητα (Simplicity):** Το να παρέχεται ένα διαισθητικό πλαίσιο εργασίας παράλληλα με ουσιώδεις οικοδομικούς λίθους, δίνοντας στους χρήστες μια πρακτική γνώση της φυσικής επεξεργασίας γλώσσας χωρίς να βαλτώνουν στη βαρέτη εργασία της επεξεργασίας σχολιασμένων γλωσσικών δεδομένων.
- **Συνέπεια (Consistency):** Το να παρέχεται ένα αμετάβλητο πλαίσιο εργασίας με συνεπείς διεπαφές και δομές δεδομένων και εύκολα προβλέψιμα ονόματα μεθόδων.

- **Επεκτασιμότητα (Extensibility):** Το να παρέχεται μια δομή μέσα στην οποία νέες υπομονάδες λογισμικού μπορούν εύκολα να φιλοξενηθούν, περιλαμβάνοντας διαφορετικές υλοποιήσεις και περιλαμβάνοντας συναγωνιστικές προσεγγίσεις στην ίδια αποστολή.
- **Συναρμολογισιμότητα (Modularity):** Το να παρέχονται συστατικά που μπορούν να χρησιμοποιηθούν ανεξάρτητα χωρίς την ανάγκη κατανόησης του υπόλοιπου πακέτου.

Κατά την υλοποίηση του συστήματός μας χρησιμοποιήσαμε την έκδοση 3.2.1 του NLTK για τις παρακάτω διαδικασίες:

- **Χωρισμός του κειμένου σε προτάσεις (Sentence Tokenization/Segmentation).**

Αν θέλουμε να χωρίσουμε σε προτάσεις ένα κομμάτι κειμένου, τότε μπορούμε να ακολουθήσουμε την παρακάτω διαδικασία:

```
>>> para = "Hello World. It's good to see you. Thanks for buying this book."
```

Now we want to split `para` into sentences. First we need to import the sentence tokenization function, and then we can call it with the paragraph as an argument.

```
>>> from nltk.tokenize import sent_tokenize
>>> sent_tokenize(para)
['Hello World.', "It's good to see you.", 'Thanks for buying this book.']
```

Σχήμα 3.2: Χωρισμός προτάσεων.

Έτσι, τώρα έχουμε μια λίστα με τις προτάσεις και μπορούμε να τις χρησιμοποιήσουμε για περαιτέρω επεξεργασία.

- **Χωρισμός των προτάσεων σε λέξεις (Word Tokenization).**

Αν θέλουμε να διαχωρίσουμε μια πρόταση σε μεμονωμένες λέξεις, θα ακολουθήσουμε τη διαδικασία που φαίνεται στην εικόνα:

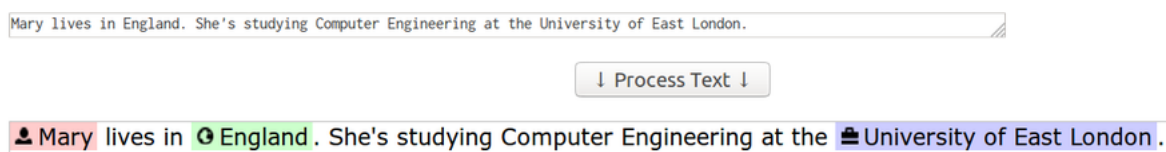
```
>>> from nltk.tokenize import word_tokenize
>>> word_tokenize('Hello World.')
['Hello', 'World', '.']
```

Σχήμα 3.3: Χωρισμός προτάσεων σε λέξεις.

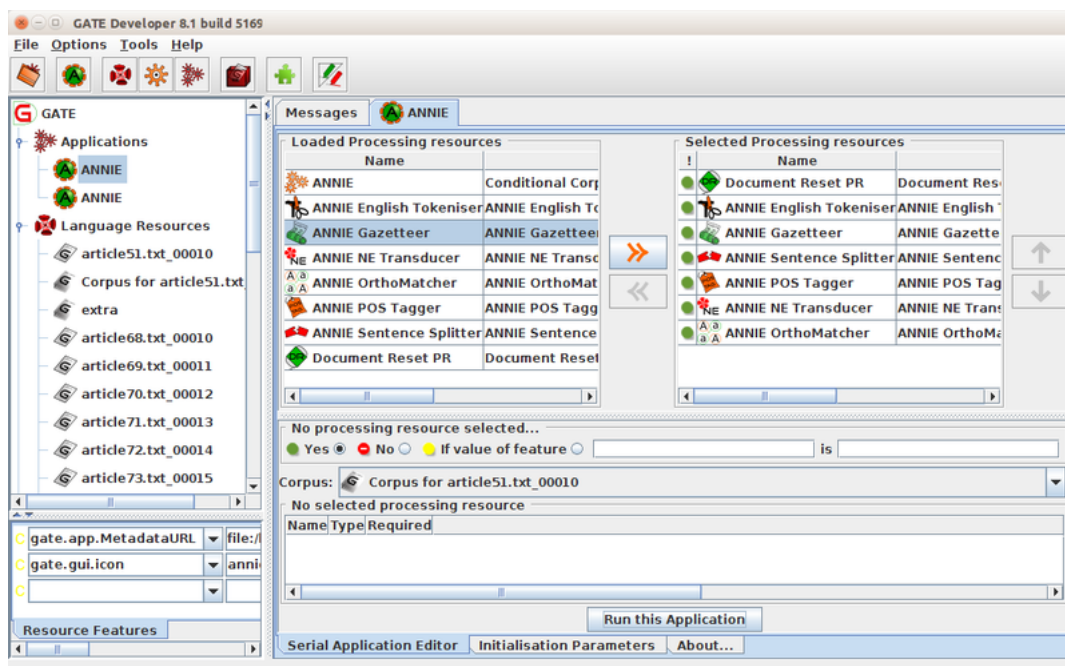
- **Απομάκρυνση ανεπιθύμητων λέξεων (Removing Stopwords).**

3.7.3 GATE for Named Entity Recognition

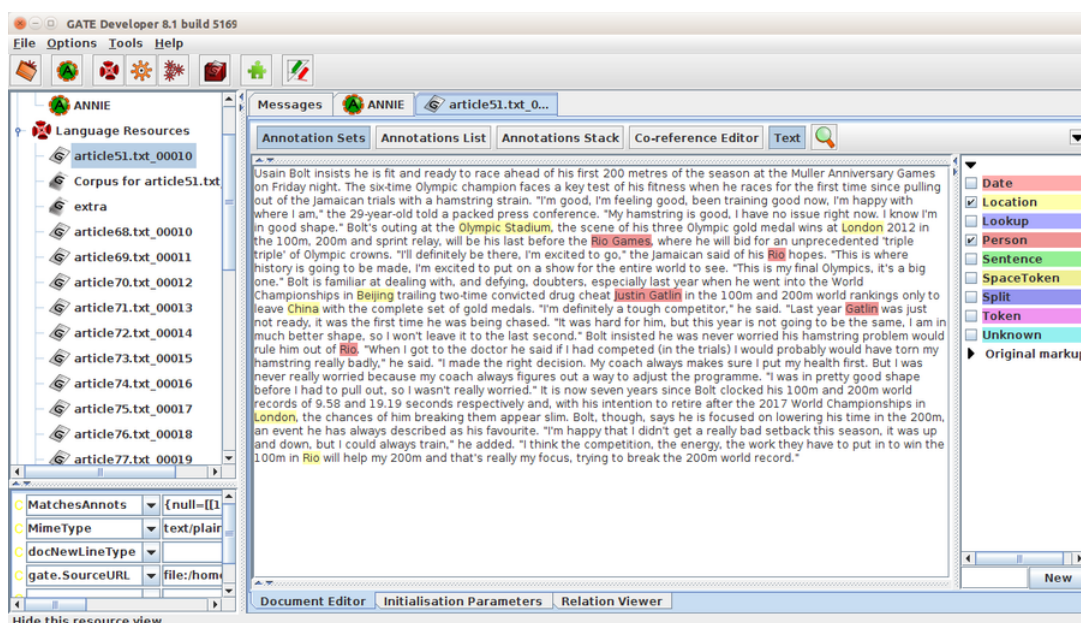
Το GATE (General Architecture for Text Engineering) [6] αποτελεί πλατφόρμα ανοιχτού κώδικα για την επεξεργασία και ανάλυση κειμένων φυσικής γλώσσας, παρέχοντας εργαλεία ανάλυσης κειμένου. Το GATE διανέμεται με το σύστημα διεξαγωγής πληροφοριών γνωστό ως ANNIE, το οποίο έχει αποτελέσει τη βάση πολλών εμπορικών και ερευνητικών συστημάτων. Το ANNIE είναι σε θέση να αναγνωρίσει έναν αριθμό διαφορετικών τύπων οντοτήτων, όπως ονόματα, τοποθεσίες και οργανισμούς, απαντώντας έτσι σε ερωτήσεις όπως “Τι συνέβη, ποιος εμπλέκεται, πότε συνέβη” κ.λπ, φράσεις οι οποίες ενδέχεται να κεντρίζουν το ενδιαφέρον αρκετών χρηστών ως προς την επιλογή άρθρων προς ανάγνωση.



Σχήμα 3.4: Αναγνώριση ονοματισμένων οντοτήτων.



Σχήμα 3.5: Εφαρμογή του εργαλείου ANNIE Gazetteer στο corpus κειμένων.



Σχήμα 3.6: Επιλογή των ετικετών που μας ενδιαφέρουν μέσω της δεξιάς στήλης.

Κεφάλαιο 4

PELOMA: A Personalized News Recommendation System

Στην ενότητα αυτή θα παρουσιάσουμε την διεπαφή χρήστη. Η διεπαφή αποτελεί έναν σημαντικό διαμεσολαβητή μεταξύ συστήματος και χρήστη. Μέσω της διεπαφής ο χρήστης έχει την δυνατότητα να επικοινωνήσει με το σύστημα, δηλαδή να δώσει τα δεδομένα του σε αυτό και να πάρει αποτελέσματα από αυτό.

Για να μπορέσει να λειτουργήσει η εφαρμογή θα πρέπει να εκτελεστούν μια φορά τα απαραίτητα υποσυστήματα, η λειτουργία των οποίων έχει αναλυθεί εκτενώς στο Κεφάλαιο 2:

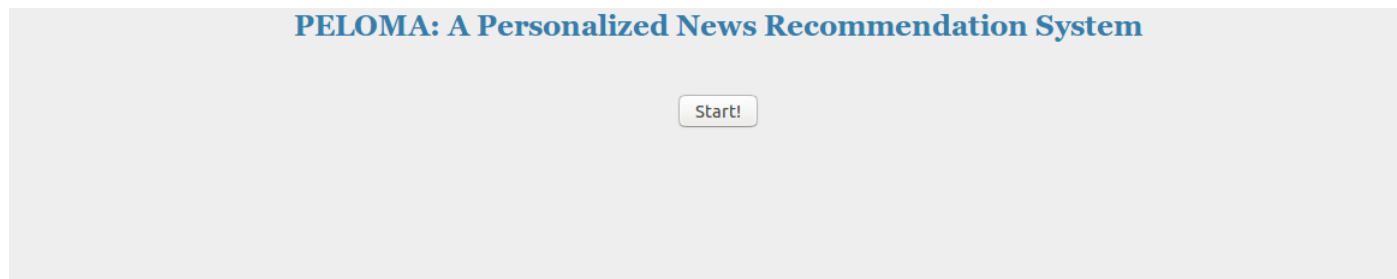
- **Υποσύστημα Δημιουργίας Βάσης Δεδομένων και Καταχώρησης Πληροφορίας**

Το υποσύστημα αυτό θα δημιουργήσει τη βάση δεδομένων και θα καταχωρήσει την απαραίτητη για τα επόμενα υποσυστήματα πληροφορία σχετικά με τα άρθρα.

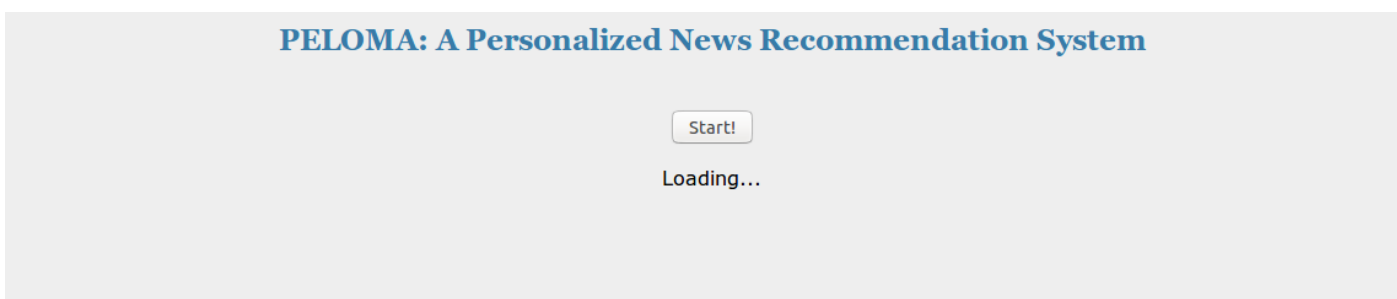
- **Υποσύστημα Ομαδοποίησης Ειδησεογραφικών Άρθρων**

Στο υποσύστημα αυτό προβαίνουμε σε ομαδοποίηση των ειδησεογραφικών άρθρων που βρίσκονται αποθηκευμένα στη βάση δεδομένων σε κατηγορίες (clusters), ομαδοποίηση των ειδησεογραφικών άρθρων εντός των clusters σε ομάδες (groups), καθώς και σε εντοπισμό ειδησεογραφικών θεμάτων μέσω πιθανοτικών μοντέλων θεμάτων για κάθε cluster, group και μεμονωμένο άρθρο της συλλογής. Επιπρόσθετα, κάνουμε εξαγωγή ονοματισμένων οντοτήτων για κάθε άρθρο. Όλη η παραχθείσα νέα πληροφορία αποθηκεύεται στη βάση δεδομένων.

Παρακάτω παρουσιάζουμε το περιβάλλον της διεπαφής:



Σχήμα 4.1: Αρχική εικόνα περιβάλλοντος διεπαφής.



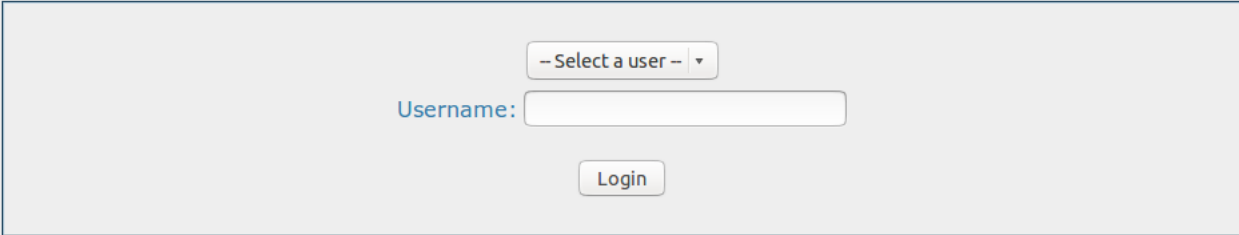
Σχήμα 4.2: Αναμονή για δημιουργία προφίλ αποθηκευμένων χρηστών και άρθρων.

Αρχικά, μέσω του κουμπιού Start! δημιουργούμε κάποια έτοιμα προφίλ, τόσο για τους αποθηκευμένους χρήστες, όσο και για τα άρθρα που βρίσκονται στο ιστορικό τους, πριν επιτρέψουμε σε ένα νέο χρήστη να εισέλθει στο σύστημα. Η αναγκαιότητα δημιουργίας έτοιμων προφίλ χρηστών προκύπτει από την απαίτηση του συστήματος για εμπλουτισμό κάθε προφίλ χρήστη με μία λίστα από άλλους χρήστες του συστήματος που επιδεικνύουν παρόμοια αναγνωστική συμπεριφορά.

Στο στάδιο αυτό υπολογίζονται και καταχωρούνται για κάθε άρθρο του ιστορικού των αποθηκευμένων χρηστών τόσο στατικά χαρακτηριστικά (π.χ. ονοματισμένες οντότητες), όσο και δυναμικά χαρακτηριστικά (π.χ. χρήστες που το διάβασαν, δημοφιλία, πόσο “φρέσκο” είναι από τη σκοπιά του πόσο πρόσφατα δημοσιεύθηκε).

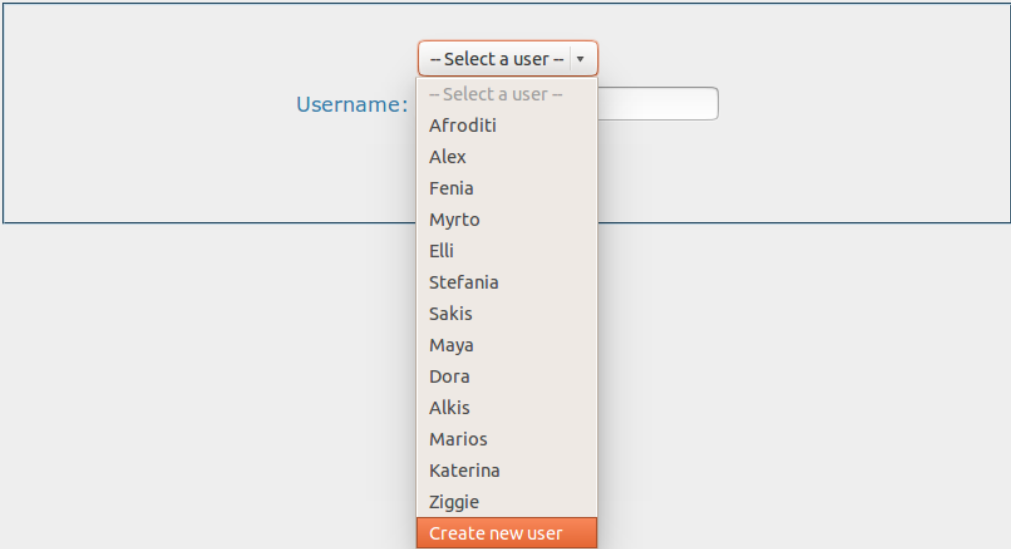
Επιπρόσθετα, υπολογίζονται η κατανομή των θεμάτων των ειδησεογραφικών άρθρων τα οποία ο κάθε αποθηκευμένος χρήστης έχει διαβάσει στο παρελθόν, η ομοιότητά του με άλλους αποθηκευμένους χρήστες, καθώς και οι ονοματισμένες οντότητες των άρθρων του ιστορικού του.

Στο επόμενο βήμα πραγματοποιείται η είσοδος του χρήστη στο σύστημα μέσω της φόρμας της εικόνας 4.3. Μπορούμε είτε να επιλέξουμε το username ενός από τους ήδη αποθηκευμένους χρήστες κάνοντας χρήση του drop-down list που εμφανίζεται και επιλέγοντας ένα από τα ονόματα της λίστας, είτε να δημιουργήσουμε έναν νέο χρήστη εισάγοντας το επιθυμητό username στο κενό πεδίο της φόρμας.



The screenshot shows the login interface for 'PELOMA: A Personalized News Recommendation System'. It features a dropdown menu labeled '- Select a user -' and a text input field labeled 'Username:'. Below the input field is a 'Login' button.

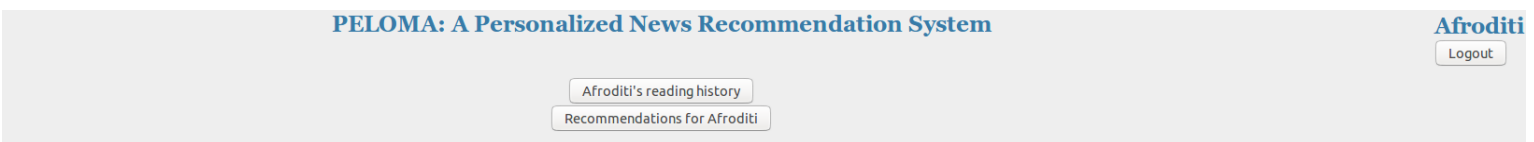
Σχήμα 4.3: Είσοδος χρήστη στο σύστημα.



The screenshot shows the same login interface as Figure 4.3, but with the dropdown menu open. The menu displays a list of usernames: Afroditi, Alex, Fenia, Myrto, Elli, Stefania, Sakis, Maya, Dora, Alkis, Marios, Katerina, and Ziggy. At the bottom of the dropdown is a button labeled 'Create new user'.

Σχήμα 4.4: Είσοδος χρήστη στο σύστημα.

Αμέσως μετά τη σύνδεση στο σύστημα, ο χρήστης οδηγείται στην κεντρική σελίδα της εφαρμογής, μέσω της οποίας αποκτά πρόσβαση στα άρθρα που είναι αποθηκευμένα στη βάση δεδομένων. Στην επάνω δεξιά γωνία του συστήματος αναγράφεται το username του συνδεδεμένου χρήστη και παρέχεται κουμπί για την αποσύνδεσή του από το σύστημα. Κατά την αποσύνδεση, ο χρήστης οδηγείται στην αρχική φόρμα εισόδου στην εφαρμογή.



Σχήμα 4.5: Κουμπί αποσύνδεσης από το σύστημα.

Σε περίπτωση όπου ο χρήστης επιχειρήσει τη λήψη συστάσεων χωρίς να έχει διαβάσει κάποιο από τα άρθρα του συστήματος, εμφανίζεται σχετικό μήνυμα, ενημερώνοντάς τον πως πρέπει να διαβάσει τουλάχιστον ένα άρθρο από κάποια κατηγορία προκειμένου να δοθεί στο σύστημα η απαραίτητη είσοδος σχετικά με τις προτιμήσεις του.



Σχήμα 4.6: Προσπάθεια λήψης συστάσεων πριν την ανάγνωση άρθρων.

Τα άρθρα παρουσιάζονται μέσω μιας λίστας στην οποία αναγράφονται το id της κατηγορίας, το id και ο τίτλος του κάθε άρθρου, ο οποίος τίτλος αποτελεί και link προς το πλήρες κείμενο.

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
2	21	Restricting immigration would make America smaller, not greater
	22	Brexit: British workers 'facing explosion of zero-hours contracts and fewer rights'
	23	Angela Merkel's party falls behind Germany's centre-left opposition SPD for first time in a decade
	24	Putin approves change to law decriminalising domestic violence
	25	All praise to John Bercow for refusing to bow to racist, sexist Donald Trump
	26	White House's 'under-reported' terror list includes many well-known attacks
	27	Angela Merkel 'explains' to Donald Trump the obligations of Geneva refugee convention after his immigration ban
	28	Government to tweak planning laws to solve housing crisis
	29	Activists blockade London meeting of 'secret Neo Nazi society'
	30	Cyprus fears russian meddling in its settlement talks
	31	Nicolas Sarkozy to face trial over 2012 campaign financing
	32	The phrase Putin never uses about terrorism (and Trump Does)
	33	What can Mexico do about Trump?
	34	Ministers refused to stop bomb sales to Saudi Arabia despite being told to do so by own export control chief
	35	UK will fund scheme to get refugees to move to Asia and Latin America
	36	Government must play a role again in job creation
	37	Government looking to amend childcare package
	38	U.S. tech titans lead legal brief against Trump travel ban
	39	During the Brexit years, Theresa May hasn't got time for domestic politics. She badly needs a deputy
	40	Ministers refused to stop bomb sales to Saudi Arabia despite being told to do so by own export control chief

Σχήμα 4.7: Προβολή λίστας άρθρων βάσης δεδομένων.

Παρακάτω βλέπουμε τη μορφή με την οποία παρουσιάζεται στο χρήστη το πλήρες κείμενο ενός άρθρου. Κατά το τέλος της ανάγνωσης ο χρήστης μπορεί να επιστρέψει στην αρχική λίστα άρθρων κάνοντας χρήση του κουμπιού *Back to Articles* που βρίσκεται στο επάνω μέρος της σελίδας.

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 message. 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.

Σχήμα 4.8: Προβολή πλήρους κειμένου ενός άρθρου.

Κάθε άρθρο που αναγνώσθηκε αποθηκεύεται στη βάση δεδομένων στο ιστορικό ανάγνωσης του εν λόγω χρήστη. Ο χρήστης μπορεί ανά πάσα στιγμή να ενημερωθεί για τα άρθρα που έχει διαβάσει, μέσω του κουμπιού *Reading history* που βρίσκεται στο επάνω μέρος της αρχικής σελίδας.

Αφού ο χρήστης περιηγηθεί μέσω της εφαρμογής διαβάζοντας άρθρα που τον ενδιαφέρουν, μπορεί να δεχθεί τις προσωποποιημένες συστάσεις άρθρων του συστήματος, πατώντας το κουμπί *Recommendations for you* που βρίσκεται στο επάνω μέρος της σελίδας. Τότε, το σύστημα εμφανίζει μία λίστα με άρθρα που είναι πιθανό να τον ενδιαφέρουν, χωρισμένα σε κατηγορίες. Κάθε λίστα συστάσεων μπορεί να περιέχει άρθρα από το πολύ τρεις διαφορετικές κατηγορίες, δεδομένης της προτίμησης των χρηστών προς συγκεκριμένες κατηγορίες άρθρων. Για κάθε προτεινόμενο άρθρο αναγράφονται το id της κατηγορίας στην οποία ανήκει, το id και ο τίτλος του άρθρου, ο οποίος τίτλος αποτελεί και link προς το πλήρες κείμενο.

Το σύστημα εμφανίζει διαφορετικό αριθμό προτεινόμενων άρθρων ανά κατηγορία, ανάλογα με το ενδιαφέρον του χρήστη ως προς τη συγκεκριμένη θεματική ενότητα.

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

Diversity of recommended news list	<input type="radio"/> Execrable <input type="radio"/> Below Average <input type="radio"/> Average <input type="radio"/> Above Average <input checked="" type="radio"/> Exceptional
Ordering of recommended news articles	<input type="radio"/> Execrable <input type="radio"/> Below Average <input type="radio"/> Average <input checked="" type="radio"/> Above Average <input type="radio"/> Exceptional

[Send your feedback!](#)

Σχήμα 4.9: Προβολή προσωποποιημένων συστάσεων.

Τέλος, ο χρήστης έχει την επιλογή να αξιολογήσει την ποιότητα των προσωποποιημένων συστάσεων του συστήματος. Μέσω των checkboxes που εμφανίζονται δίπλα στα άρθρα, μπορεί να επιλέξει τις προτάσεις του συστήματος που σχετίζονται πραγματικά με τα ενδιαφέροντά του. Ακριβώς κάτω από τη φόρμα συστάσεων μπορεί να αξιολογήσει το σύστημα τόσο ως προς την ποικιλία της λίστας συστάσεων, όσο και ως προς την κατάταξη των άρθρων της λίστας.

Το feedback του χρήστη δίνεται στο σύστημα πατώντας το κουμπί Send your feedback!.

Ο αριθμός των άρθρων που ο χρήστης βρήκε ενδιαφέροντα αναγράφεται σε μορφή ποσοστού ως προς το συνολικό αριθμό προτεινόμενων άρθρων.

PELOMA: A Personalized News Recommendation System

[Back to Articles](#)

Afroditi's rating	
Article preference:	8/9
Recommended list's diversity:	Exceptional
Recommended articles' ordering:	Above Average

We appreciate your feedback!

Σχήμα 4.10: Αξιολόγηση συστήματος από το χρήστη.

Ο χρήστης μπορεί είτε να αποσυνδεθεί από το σύστημα, είτε να συνεχίσει την ανάγνωση περισσότερων άρθρων, δεχόμενος εκ νέου συστάσεις.

Κεφάλαιο 5

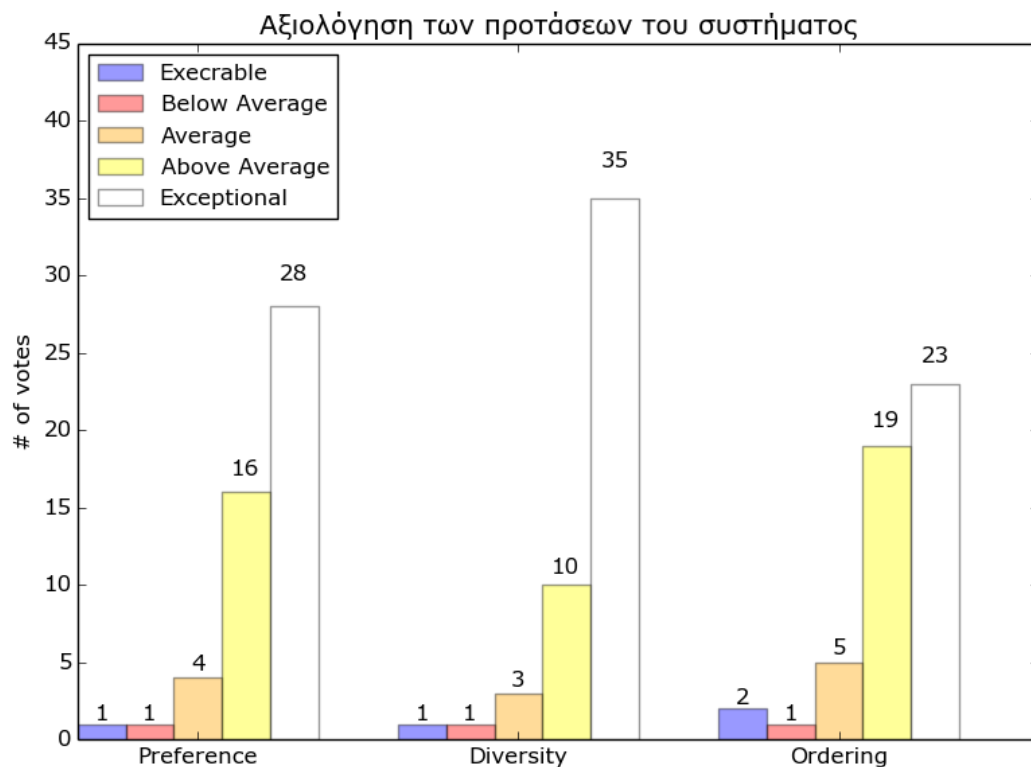
Αξιολόγηση Συστήματος

Στο κεφάλαιο αυτό γίνεται αξιολόγηση του μηχανισμού συστάσεων του συστήματος που υλοποιήσαμε για τις προσωποποιημένες προτάσεις άρθρων ειδήσεων σε μεμονωμένους χρήστες.

Εφόσον ο στόχος του συστήματος είναι η παραγωγή χρήσιμων ως προς τους χρήστες συστάσεων, πραγματοποιήσαμε ένα πείραμα για να ελέγξουμε την ακρίβεια της λειτουργίας αυτής. Το πείραμα βασίστηκε στην αξιολόγηση του συστήματος από 50 χρήστες. Οι εθελοντές χρήστες εισήλθαν στο σύστημα, επέλεξαν προς ανάγνωση άρθρα της αρεσκείας τους και το σύστημα τα καταχώρησε στο αναγνωστικό τους ιστορικό. Στη συνέχεια, το σύστημα δημιούργησε σύνολα προτάσεων για καθέναν από αυτούς. Η ικανοποίηση κάθε χρήστη υπολογίστηκε ως προς τα εξής τρία κριτήρια:

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

Τα αποτελέσματα του πειράματος παρουσιάζονται στο παρακάτω ιστόγραμμα:



Σχήμα 5.1: Αξιολόγηση των προτάσεων του συστήματος.

Το συμπέρασμα που εξάγουμε από την πειραματική διαδικασία είναι ότι η πλειοψηφία των χρηστών επωφελείται από τις συστάσεις που παράγει το σύστημα, καθώς το μεγαλύτερο ποσοστό των χρηστών δηλώνει ικανοποιημένο ως προς τις απαιτήσεις για άρθρα σχετικά με τα ενδιαφέροντά του, ποικιλία στη λίστα συστάσεων και σωστή κατάταξη των άρθρων στην τελική λίστα. Συγκεκριμένα, ιδιαίτερα ικανοποιητικά είναι τα αποτελέσματα σχετικά με την ποικιλία της λίστας συστάσεων, όπου το σύστημα φαίνεται να εναρμονίζεται πλήρως με τα ενδιαφέροντα του χρήστη και να μην αφήνει καμία απ' τις ενδιαφέρουσες κατηγορίες άρθρων εκτός της τελικής σύστασης.

Άξιο καταγραφής αποτελεί και το πώς διαμορφώνονται οι συστάσεις προς έναν χρήστη κατά το χρόνο παραμονής του στο σύστημα, όπου δέχεται συστάσεις και συνεχίζει την ανάγνωση άρθρων. Κάθε λίστα συστάσεων περιέχει άρθρα από το πολύ τρεις διαφορετικές κατηγορίες, λαμβάνοντας υπόψη την τάση των χρηστών να προτιμούν συγκεκριμένες κατηγορίες άρθρων.

Στον Πίνακα 5.1 αποτυπώνονται οι πληροφορίες που αφορούν έναν τυχαίο χρήστη τη στιγμή που επιλέγει για πρώτη φορά να δεχθεί τις συστάσεις άρθρων του συστήματος.

Στον πίνακα καταγράφονται το ιστορικό ανάγνωσης του χρήστη, δηλαδή ο αριθμός των άρθρων που έχουν αναγνωσθεί από κάθε κατηγορία έως τη δεδομένη στιγμή, η τιμή ομοιότητας του χρήστη με κάθε κατηγορία άρθρων, καθώς και ο αριθμός προτεινόμενων άρθρων από τις επιλεγμένες τρεις κατηγορίες με τη μεγαλύτερη ομοιότητα.

<i>1st Recommendation</i>	# 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

Πίνακας 5.1: Ιστορικό ανάγνωσης και 1η φάση συστάσεων

Παρατηρούμε πως το σύστημα εντόπισε ορθώς την προτίμηση του χρήστη για τις τρεις αυτές κατηγορίες, δεδομένου ότι είναι και αυτές από τις οποίες έχει αναγνώσει το μεγαλύτερο αριθμό άρθρων. Λαμβάνοντας υπόψη το μικρό αριθμό άρθρων που έχουμε αποθηκευμένα στη βάση δεδομένων από κάθε κατηγορία, ο αλγόριθμος του συστήματος έχει σχεδιαστεί να προτείνει τέσσερα άρθρα από την πιο ταιριαστή κατηγορία, τρία άρθρα από τη δεύτερη πιο ταιριαστή και δύο άρθρα από την τρίτη σε σειρά πιο ταιριαστή κατηγορία. Αυτή η παραδοχή προϋποθέτει την ύπαρξη του αντίστοιχου αριθμού άρθρων μέσα στην επιλεγμένη ομάδα (group) κάθε επιλεγμένης κατηγορίας (cluster).

```
Selected groups (one for each of the selected clusters): [2, 0, 2]
Articles inside selected groups:
--- Cluster 2 Group 2: [24, 25, 26, 27, 30, 32]
--- Cluster 3 Group 0: [42, 43, 48, 49, 52, 53, 54, 55, 56, 57, 60]
--- Cluster 6 Group 2: [98, 100, 102, 105, 106, 109, 111, 114, 123, 124, 125]
```

Σχήμα 5.2: Επιλεγμένες ομάδες από κάθε κατηγορία.

Στην Εικόνα 5.2 αποτυπώνονται οι ομάδες άρθρων που επιλέχθηκαν και οι οποίες μας οδήγησαν στην 1η φάση συστάσεων άρθρων. Μερικά από τα άρθρα εντός

των επιλεγμένων ομάδων έχουν ήδη αναγνωσθεί από το χρήστη και είναι αυτά που οδήγησαν το σύστημα να συμπεράνει την ομοιότητα. Έτσι, γίνεται εύκολα αντιληπτό ότι εξαιτίας του περιορισμένου αριθμού άρθρων στη βάση δεδομένων, είναι πιθανό η ομοιότητα ενός χρήστη με μία κατηγορία να αυξάνεται καθώς αυτός συνεχίζει να επιλέγει άρθρα από την εν λόγω κατηγορία, όμως να μην υπάρχει διαθέσιμος ο επιθυμητός αριθμός άρθρων προς σύσταση.

Στον Πίνακα 5.2 αποτυπώνονται το ιστορικό ανάγνωσης του χρήστη, η τιμή ομοιότητας του χρήστη με κάθε κατηγορία άρθρων και η δεύτερη φάση συστάσεων. Παρατηρούμε πως ο χρήστης επέλεξε να αναγνώσει τόσο άρθρα από τις προτεινόμενες κατηγορίες, όσο και ένα άρθρο από την Κατηγορία #5, εξ ου και η αύξηση της ομοιότητάς του με την κατηγορία αυτή. Το σύστημα εξακολουθεί να συμπεριφέρεται ορθά, δίνοντας μεγαλύτερη βαρύτητα στην Κατηγορία #6.

<i>2nd Recommendation</i>	# 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

Πίνακας 5.2: Ιστορικό ανάγνωσης και 2η φάση συστάσεων

Στον Πίνακα 5.3 αποτυπώνονται το ιστορικό ανάγνωσης του χρήστη, η τιμή ομοιότητας του χρήστη με κάθε κατηγορία άρθρων και η τρίτη φάση συστάσεων. Παρατηρούμε πως το σύστημα λειτουργεί ομαλά, επιστρέφοντας ξανά συστάσεις από την Κατηγορία #2, απ' όπου ο χρήστης έχει διαβάσει και το μεγαλύτερο αριθμό άρθρων.

<i>3rd Recommendation</i>	# 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

Πίνακας 5.3: Ιστορικό ανάγνωσης και 3η φάση συστάσεων

Το σύστημα καλύπτει επαρκώς τις αναγνωστικές ανάγκες του χρήστη, καθώς κατά τις τρεις διαφορετικές φάσεις έγιναν συστάσεις από τέσσερις διαφορετικές κατηγορίες άρθρων, από τις οποίες είχε διαβαστεί ένας σημαντικός αριθμός άρθρων.

Τέλος, ως μέρος της πειραματικής ανάλυσης επιλέγουμε να παραθέσουμε και ένα 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. Πρώτη επιλογή προς ανάγνωση αποτελεί το άρθρο #12.

Σκοπός μας είναι να διαπιστώσουμε αν το σύστημα θα λειτουργήσει ορθώς, προτείνοντας στο χρήστη άρθρα μόνο από τη συγκεκριμένη κατηγορία, καθώς και να παρατηρήσουμε τον τρόπο με τον οποίο επιλέγονται τα άρθρα από το πιο όμοιο γκρουπ έως το λιγότερο όμοιο, καθώς ο χρήστης συνεχίζει την ανάγνωση των προτεινόμενων άρθρων μέχρι την εξάντληση όλων των διαθέσιμων άρθρων της κατηγορίας αυτής.

```

--- 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 5.3: Groups άρθρων της Κατηγορίας #1.

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

Σχήμα 5.4: Κατηγορία #1 - 1η σύσταση - Άρθρα από group 0 .

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

Σχήμα 5.5: Κατηγορία #1 - 2η σύσταση - Άρθρα από group 1 .

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

Σχήμα 5.6: Κατηγορία #1 - 3η σύσταση - Άρθρα από group 1 .

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

Σχήμα 5.7: Κατηγορία #1 - 4η σύσταση - Άρθρα από group 2 .

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

Σχήμα 5.8: Κατηγορία #1 - 5η σύσταση - Άρθρα από group 2 .

Κλείνοντας, η πειραματική αξιολόγηση έδειξε πως το σύστημα είναι σε θέση να καλύψει επαρκώς τις αναγνωστικές ανάγκες του χρήστη με βάση τα δεδομένα που του έχουν δοθεί ως είσοδος. Επιπλέον, μπορούμε να παρατηρήσουμε ότι σε ένα τέτοιο σύστημα είναι δύσκολο να πραγματοποιηθεί μια αντικειμενική αξιολόγηση, καθώς οι προτιμήσεις του χρήστη διαφέρουν από αυτές των υπολοίπων χρηστών ή ακόμα και από τις δικές του, ανάλογα με τις περιστάσεις.

Κεφάλαιο 6

Συμπεράσματα και Μελλοντικές Επεκτάσεις

6.1 Συμπεράσματα

Καθώς όλο και περισσότερες πληροφορίες γίνονται διαθέσιμες στο διαδίκτυο, οι χρήστες όλο και περισσότερο ψάχνουν απεγνωσμένα κάποια εργαλεία που θα τους βοηθήσουν να φιλτράρουν αυτή τη ροή των πληροφοριών και να βρουν άρθρα ειδήσεων που να τους ενδιαφέρουν. Ακριβώς σε ένα τέτοιο σενάριο, το σύστημα που αναπτύχθηκε στο πλαίσιο αυτής της διπλωματικής εργασίας προσπαθεί να περιορίσει το πρόβλημα που δημιουργείται από την διαρκή ροή ειδήσεων από διαφορετικές πηγές ενημέρωσης. Αυτό που ουσιαστικά θέλουμε να δημιουργήσουμε είναι να παρέχουμε μια εξατομικευμένη υπηρεσία συστάσεων για άρθρα ειδήσεων, στην οποία ο χρήστης θα επιλέγει τις κατηγορίες σχετικά με τις οποίες θέλει να ενημερώνεται και με βάση το προφίλ του και το ιστορικό χρήσης, το σύστημα θα παρέχει εξατομικευμένες συστάσεις που ταιριάζουν περισσότερο με τα ενδιαφέροντά του.

Στο πλαίσιο της εργασίας παρουσιάσαμε ένα σύστημα δημιουργίας εξατομικευμένων συστάσεων σε εφαρμογή διαδικτυακού περιεχομένου, η οποία λαμβάνει υπόψη το προφίλ και το ιστορικό των χρηστών του συστήματος για να προτείνει άρθρα ειδήσεων. Τα άρθρα αναπαρίστανται με τη βοήθεια θεματικών μοντέλων, δηλαδή μοντέλων για την ανακάλυψη θεμάτων που υπάρχουν σε μία συλλογή κειμένων. Τα προφίλ των χρηστών αναπαρίστανται μέσω μιας τριπλέτας αποτελούμενης από τα εξής χαρακτηριστικά: την κατανομή των θεμάτων των αναγνωσμένων άρθρων, τη λίστα χρηστών οι οποίοι έχουν παρόμοια πρότυπα πρόσβασης με τον εν λόγω χρήστη και τέλος, τη λίστα από ονοματισμένες οντότητες, δηλαδή λέξεις οι οποίες απαντούν σε φράσεις όπως ‘Τι συνέβη, ποιος εμπλέκεται, πότε συνέβη’ κλπ. Στόχος μας ήταν να βρούμε σύντομες περιγραφές των άρθρων της συλλογής και να εξερευνήσουμε

τους συσχετισμούς μεταξύ των clusters (ή των groups) άρθρων και του προφίλ του δοθέντος χρήστη, συγκρίνοντας την ομοιότητα των θεμάτων που “κρύβονται” μέσα στα άρθρα τους. Οι σχέσεις μεταξύ αυτών των εννοιών εμπλουτίζουν τις παραπάνω αναπαραστάσεις και ενσωματώνονται στις διαδικασίες δημιουργίας συστάσεων.

Πιο συγκεκριμένα, συλλέξαμε άρθρα ειδήσεων τόσο από τον παγκόσμιο ιστό, όσο και από τη συλλογή Reuters του NLTK και τα αποθηκεύσαμε σε μία βάση δεδομένων. Το λειτουργικό μέρος του μηχανισμού εξάγει χρήσιμο κείμενο από αυτά, πραγματοποιεί μεθόδους εξαγωγής λέξεων κλειδιών από κάθε άρθρο που υπάρχει στο σύστημα, εφαρμόζει τα θεματικά μοντέλα σε κάθε κατηγορία, γκρουπ και άρθρο, επιτρέποντας με αυτό τον τρόπο την περαιτέρω επεξεργασία τους βάσει σημασιολογικών συσχετίσεων μεταξύ των άρθρων. Το ορατό στους χρήστες είναι ο δικτυακός τόπος που εμφανίζει τα άρθρα του συστήματος σημασιολογικά κατηγοριοποιημένα. Ο μηχανισμός προτάσεων του συστήματος, με βάση το προφίλ του χρήστη, παράγει ένα σύνολο εξατομικευμένων προτάσεων ειδήσεων που ταιριάζουν περισσότερο με τα σημασιολογικά ενδιαφέροντα του χρήστη, κάνοντας δυνατή με αυτό τον τρόπο την παρουσίαση ειδήσεων που σχετίζονται σημασιολογικά με αυτές που ήδη έχει αναγνώσει.

Κατά την πειραματική αξιολόγηση του συστήματος, η ικανοποίηση κάθε χρήστη υπολογίστηκε ως προς τα εξής τρία κριτήρια: συσχετισμός των προτεινόμενων άρθρων με τα πραγματικά του ενδιαφέροντα (Preference), ποικιλία της λίστας συστάσεων (Diversity) και κατάταξη των άρθρων της λίστας συστάσεων (Ordering). Επιπρόσθετα, παρουσιάστηκαν σενάρια χρήσης του συστήματος κατά τα οποία ο χρήστης είτε έχει αναγνώσει άρθρα από διάφορες κατηγορίες, είτε μόνο από μία κατηγορία και επιλέγει να δεχτεί τις συστάσεις του συστήματος. Διαπιστώθηκε ότι 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, όπου το σύστημα φαίνεται να εναρμονίζεται πλήρως με τα ενδιαφέροντα του χρήστη και να μην αφήνει καμία απ’ τις ενδιαφέρουσες κατηγορίες άρθρων εκτός της τελικής σύστασης.

Από τα πειραματικά ευρήματα συμπεραίνουμε, επίσης, ότι το σύστημά μας έχει καλή απόδοση σε διαφορετικούς τύπους κειμένων και ιδιαίτερα, όταν τα κείμενα είναι μικρά σε μέγεθος. Αυτό ερμηνεύεται μέσω του διανύσματος αναπαράστασης ενός κειμένου που προκύπτει από την εφαρμογή του αλγορίθμου LDA, στο οποίο επιλέγουμε τον αριθμό αντιπροσωπευτικών λέξεων απ’ τις οποίες θα αποτελείται. Έτσι, ένα τέτοιο διάνυσμα έχει μεγαλύτερες πιθανότητες να αποτυπώσει ορθότερα το νόημα ενός άρθρου, όταν το άρθρο αποτελείται από σχετικά μικρό αριθμό λέξεων.

Αξίζει να σημειώσουμε ότι τα θεματικά μοντέλα είναι ένα χρήσιμο εργαλείο εξερεύνησης. Τα θέματα παρέχουν μία περίληψη ενός σώματος κειμένων που είναι αδύνατο να γίνει με το χέρι. Η θεματική ανάλυση μπορεί να ανακαλύψει συνδέσεις ανάμεσα και μέσα στα κείμενα που δεν είναι φανερές με γυμνό μάτι και να βρει συσχετίσεις όρων που δε θεωρούνται δεδομένες.

Πρόκληση για το σύστημά μας αποτελεί η αύξηση του πλήθους των άρθρων ανά κατηγορία στη βάση δεδομένων του συστήματος, καθώς και η αξιολόγηση των ομάδων που θα δημιουργηθούν από την εφαρμογή του αλγορίθμου k-means εντός κάθε κατηγορίας.

Τέλος, άξιο αναφοράς αποτελεί το πόσο βοηθητική υπήρξε η χρήση της γλώσσας προγραμματισμού Python σε μία τέτοια εφαρμογή και συγκεκριμένα η πλατφόρμα NLTK (Natural Language Toolkit), μια πλατφόρμα με έτοιμα εργαλεία επεξεργασίας φυσικής γλώσσας.

6.2 Μελλοντικές Επεκτάσεις

Το σύστημα που αναπτύχθηκε στο πλαίσιο αυτής της διπλωματικής εργασίας θα μπορούσε να βελτιωθεί και να επεκταθεί περαιτέρω, τουλάχιστον ως προς τις κάτωθι κατευθύνσεις:

Συστάσεις σε ομάδες χρηστών

Η περιγραφή των προτιμήσεων με τη βοήθεια διανυσμάτων επιτρέπει το συνδυασμό πολλαπλών προφίλ για τη δημιουργία ενός κοινού προφίλ για μια ομάδα χρηστών.

Ανατροφοδότηση του συστήματος με την αξιολόγηση του χρήστη

Το σύστημα μπορεί να δέχεται ως είσοδο την αξιολόγηση του χρήστη, αποθηκεύοντας στη βάση δεδομένων τα προτεινόμενα άρθρα που ο χρήστης βρήκε πραγματικά ενδιαφέροντα σε σχέση με ολόκληρη τη λίστα συστάσεων. Έτσι, μπορούν να πραγματοποιηθούν κάποιες πιο ισχυρές συνάψεις 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 n -άδων σύστασης που προσφέρουν τη μέγιστη αύξηση ωφέλειας για το χρήστη (Εφαρμογή submodular μοντέλου συστάσεων με χρήση άπληστου προσεγγιστικού αλγορίθμου).

Βιβλιογραφία

- [1] Python 3. <http://www.diveintopython3.net/>, 2017.
- [2] Kmeans algorithm. <https://nlp.stanford.edu/ir-book/html/htmledition/k-means-1.html>, 2017.
- [3] BeautifulSoup. <https://www.crummy.com/software/beautifulsoup/>, 2017.
- [4] NLTK Book. <http://www.nltk.org/book/>, 2017.
- [5] Andrew Y. Ng David M. Blei και Michael I. Jordan. Latent dirichlet allocation. 2003.
- [6] GATE. <https://gate.ac.uk/>, 2017.
- [7] Lda gensim. <https://radimrehurek.com/gensim/models/ldamodel.htmls>, 2017.
- [8] HTML. <https://en.wikipedia.org/wiki/html>, 2017.
- [9] G. Mentzas K. Christidis, D. Apostolou. Exploring customer preferences with probabilistic topics models. 2010.
- [10] Jeremy Kun. When greedy algorithms are good enough: The submodularity condition. 2017.
- [11] L. Console L. Ardissono και I. Torre. An adaptive system for the personalized access to news. 2001.
- [12] LDA. https://en.wikipedia.org/wiki/latent_dirichlet_allocation, 2017.
- [13] Tao Li Daniel Knox Balaji Padmanabhan Lei Li, Dingding Wang. Scene: a scalable two-stage personalized news recommendation system. 2011.
- [14] Flask (A Python Microframework). <http://flask.pocoo.org/>, 2017.
- [15] B. Mobasher. Data mining for web personalization. 2007.
- [16] MySQL. <https://en.wikipedia.org/wiki/mysql>, 2017.

- [17] A.Kobsa P. Brusilovsky και W. Nejdl. Hybrid web recommender systems. 2007.
- [18] Python. <https://www.python.org/>, 2017.
- [19] Stuart Russell και Peter Norvig Third Edition. Artificial intelligence: A modern approach. Published by Prentice Hall, 2009.
- [20] Joseph Naor Samir Khuller, Anna Moss. The budgeted maximum coverage problem. 2003.
- [21] Apache HTTP Server. <https://httpd.apache.org/>, 2017.
- [22] CSS Cascading Style Sheets. <https://www.w3.org/style/css/overview.en.html>, 2017.
- [23] Cosine Similarity. https://en.wikipedia.org/wiki/cosine_similarity, 2017.
- [24] Jaccard Similarity. https://en.wikipedia.org/wiki/jaccard_index, 2017.
- [25] Wongkot Sriurai, Phayung Meesad και Choochart Haruechaiyasak. Recommending related articles in wikipedia via a topic-based model. 2008.
- [26] Tf-idf. <https://nlp.stanford.edu/ir-book/html/htmledition/tf-idf-weighting-1.html>, 2017.
- [27] TreeTagger. <http://www.cis.uni-muenchen.de/~schmid/tools/treetagger/>, 2017.
- [28] GATE Tutorial. <https://devo-evo.lab.asu.edu/methods/?q=system/files/week4-gatetutorial.pdf>, 2017.
- [29] Tfidf Vectorizer. http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.tfidfvectorizer.html, 2017.

