ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΣΥΣΤΗΜΑ ΣΥΣΤΑΣΕΩΝ ΓΙΑ ΒΙΒΛΙΑ

Διπλωματική Εργασία

Κουρουνιώτου Αντωνία

Επιβλέπων: Βασιλακόπουλος Μιχαήλ

Σεπτέμβριος 2022

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΣΥΣΤΗΜΑ ΣΥΣΤΑΣΕΩΝ ΓΙΑ ΒΙΒΛΙΑ

Διπλωματική Εργασία

Κουρουνιώτου Αντωνία

Επιβλέπων: Βασιλακόπουλος Μιχαήλ

Σεπτέμβριος 2022

UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

RECOMMENDER SYSTEM FOR BOOKS

Diploma Thesis

Kourouniotou Antonia

Supervisor: Vassilakopoulos Michael

September 2022

Εγκρίνεται από την Επιτροπή Εξέτασης:

Επιβλέπων **Βασιλακόπουλος Μιχαήλ**
Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Μέλος **Τσομπανοπούλου Παναγιώτα**
Αναπληρώτρια Καθηγήτρια, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Μέλος **Τουσίδου Ελένη**
Ε.ΔΙ.Π., Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Ευχαριστίες

Με την παράδοση της διπλωματικής εργασίας ολοκληρώνεται ο κύκλος σπουδών μου στο τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Πανεπιστημίου Θεσσαλίας. Ύπήρξε ένα ταξίδι, στο τέλος του οποίου νιώθω την ανάγκη να εκφράσω τις ευχαριστίες μου στους ανθρώπους που στάθηκαν δίπλα μου και υπήρξαν συνοδοιπόροι.

Για αρχή, θα ήθελα να ευχαριστήσω τον κύριο Βασιλακόπουλο, Καθηγητή στο Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Πανεπιστήμιο Θεσσαλίας, που υπήρξε ο επιβλέπων της παρούσας εργασίας. Η επικοινωνία μας μπορεί να χαρακτηριστεί μόνο άριστη, τόσο ως προς την άμεση ανταπόκριση, όσο και ως προς την συμβουλευτική του ικανότητα.

Στη συνέχεια, θα ήθελα να ευχαριστήσω την κυρία Τσομπανοπούλου, Αναπληρώτρια Καθηγήτρια στο Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Πανεπιστήμιο Θεσσαλίας και την κυρία Τουσίδου, Ε.ΔΙ.Π. στο Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Πανεπιστήμιο Θεσσαλίας, που δέχτηκαν να αποτελέσουν μέλη της επιτροπής εξέτασης της διπλωματικής μου εργασίας.

Τέλος, δε θα μπορούσα να μην εκφράσω το μεγαλύτερο μου ευχαριστώ στην οικογένεια μου και τους φίλους μου, που υπήρξαν δίπλα μου από την αρχή αυτού του ταξιδιού και που χωρίς αυτούς όλα τα πράγματα θα είχαν λιγότερη σημασία.

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ ΠΕΡΙ ΑΚΑΔΗΜΑΪΚΗΣ ΔΕΟΝΤΟΛΟΓΙΑΣ ΚΑΙ ΠΝΕΥΜΑΤΙΚΩΝ ΔΙΚΑΙΩΜΑΤΩΝ

«Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ρητά ότι η παρούσα διπλωματική εργασία, καθώς και τα ηλεκτρονικά αρχεία και πηγαίοι κώδικες που αναπτύχθηκαν ή τροποποιήθηκαν στα πλαίσια αυτής της εργασίας, αποτελεί αποκλειστικά προϊόν προσωπικής μου εργασίας, δεν προσβάλλει κάθε μορφής δικαιώματα διανοητικής ιδιοκτησίας, προσωπικότητας και προσωπικών δεδομένων τρίτων, δεν περιέχει έργα/εισφορές τρίτων για τα οποία απαιτείται άδεια των δημιουργών/δικαιούχων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον και πληρούν τους κανόνες της επιστημονικής παράθεσης. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο, αρχεία ή/και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή. Δηλώνω επίσης ότι τα αποτελέσματα της εργασίας δεν έχουν χρησιμοποιηθεί για την απόκτηση άλλου πτυχίου. Αναλαμβάνω πλήρως, ατομικά και προσωπικά, όλες τις νομικές και διοικητικές συνέπειες που δύναται να προκύψουν στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής».

Η Δηλούσα

Κουρουνιώτου Αντωνία

Διπλωματική Εργασία

ΣΥΣΤΗΜΑ ΣΥΣΤΑΣΕΩΝ ΓΙΑ ΒΙΒΛΙΑ

Κουρουνιώτου Αντωνία

Περίληψη

Η πρόοδος και η εξέλιξη που χαρακτηρίζει την εποχή που διανύουμε είναι άμεσα συνυφασμένη με την σώστη αξιοποίηση των πληροφοριών που είναι διαθέσιμες. Το γεγονός αυτό σε συνδυασμό με τον συνεχώς αυξανόμενο όγκο των πληροφοριών, εντείνουν την ανάγκη δημιουργίας Συστημάτων Σύστασης, προκειμένου να επιτευχθεί η διακριτοποίηση των πληροφοριών. Ιδιαίτερη και αρκετά πιο απαιτητική κρίνεται η ανάπτυξη συστημάτων σύστασης, που ως παράμετρο διακριτοποίησης χρησιμοποιούν κείμενα γραμμένα στη φυσική γλώσσα, τα οποία δεν διαθέτουν κάποια υπάρχουσα ταξινόμηση.

Η εργασία αυτή ερευνά τη δημιουργία ενός μη προσωποποιημένου συστήματος σύστασης για βιβλία, αξιοποιώντας για την διακριτοποίηση το κείμενο περιγραφής των βιβλίων. Πιο λεπτομερώς, η ανάπτυξη αυτού του συστήματος υλοποιείται με χρήση του μοντέλου Λανθάνουσας Κατανομής Dirichlet (LDA), προκειμένου να ανακαλυφθούν θεματικές ενότητες που βρίσκονται κρυμμένες μέσα στα κείμενα περιγραφής. Αφού, πραγματοποιηθεί ο κατάλληλος συντονισμός των παραμέτρων του μοντέλου και η αξιολόγηση του με ποικίλες μεθόδους, εκπαιδεύεται το μοντέλο πάνω στα κείμενα της βάσης δεδομένων με τα βιβλία, παράγοντας τις τελικές θεματικές ενότητες. Βάσει αυτών των θεματικών ενοτήτων, κάθε βιβλίο αποκτά μία κατανομή ως προς αυτές, που αξιοποιείται σε συδνυασμό με την απόσταση Jensen-Shannon προκειμένου να προκύψει η συσχέτιση μεταξύ των βιβλίων και τελικώς η σύσταση.

Τέλος, όλη αυτή η διαδικασία οργανώνεται σε μία εύχρηστη εφαρμογή για τον χρήστη, παρέχοντάς του τη δυνατότητα να κάνει ποικίλων ειδών αναζητήσεις προκειμένου να του συσταθούν βιβλία βάσει των επιθυμιών του.

Λέξεις-κλειδιά:

Σύστημα Σύστασης, Λανθάνουσα Κατανομή Dirichlet, Απόσταση Jensen-Shannon, Συνάφεια θέματος, Μοντέλο Word2vec

Diploma Thesis

RECOMMENDER SYSTEM FOR BOOKS

Kourouniotou Antonia

Abstract

The progress and development that characterizes the time we are going through is directly intertwined with the correct utilization of the information that is available. This fact, combined with the ever-increasing volume of information, intensify the need to create Recommendation Systems, in order to achieve the differentiation of information. The development of recommendation systems, which as a differentiation parameter use texts written in natural language, which do not have any existing classification, is particularly and considerably more demanding.

This work investigates the creation of a non-personalized book recommendation system, utilizing the book description text for differentiation. In more detail, the development of this system is implemented using the Latent Dirichlet Allocation (LDA) model, in order to discover thematic units hidden within the description texts. After the appropriate coordination of the model parameters and its evaluation with various methods, the model is trained on the texts of the database with the books, producing the final thematic units. Based on these thematic units, each book obtains a distribution in terms of them, which is used in combination with the Jensen-Shannon distance in order to derive the correlation between the books and the final recommendation.

Finally, this whole process is organized in an easy-to-use application for the user, giving him the possibility to make various kinds of searches in order to recommend books based on his wishes.

Keywords:

Recommender System, Latend Dirichlet Allocation, Jensen-Shannon distance, Topic Coherence, Word2vec Model

Πίνακας περιεχομένων

Κατάλογος σχημάτων

Κατάλογος πινάκων

Συντομογραφίες

βλπ.	βλέπε
κ.λπ.	και λοιπά
ΣΣ	Σύστημα Σύστασης

BoW	Bag of Words
CBF	Content Based Filtering
CBOW	Continuous Bag of Words
CF	Collaborative Filtering
GUI	Graphical User Interface
HTML	HyperText Markup Language
JS	Jensen Shannon
KL	Kullback Leibler
LDA	Latent Dirichlet Allocation
NLP	Natural Language Process
nltk	Natural Language Toolkit
PMI	Pointwise Mutual Information
RS	Recommendation System

Κεφάλαιο 1

Εισαγωγή

Στην εποχή που τα τεχνολογικά μέσα και το Διαδίκτυο έχουν κατακλύσει κάθε πτυχή της ανθρώπινης ζωής, είναι φυσικό και επόμενο να μπορει να δωθεί σε αυτην την εποχή ο χαρακτηρισμός, "Εποχή της πληροφορίας". Αυτή η συσσώρευση πληροφοριών εκ πρώτης όψεως ακούγεται εντυπωσιακή, αλλά αν αναλογιστεί κανείς τον όγκο αυτων των πληροφοριών, θα αντιληφθεί ότι είναι απλά χαοτική, καθώς χωρίς κάποιον τρόπο σύστασης καθίσταται ακατόρθωτο ο χρήστης να προσκομίσει τα δεδομένα βάσει των επιθυμιών του.

Μπορεί η αρχική προσέγγιση να είναι γενική, αφού ως σημείο αναφοράς χρησιμοποιείται το Διαδίκτυο, αλλά το ίδιο πρόβλημα καλείται να αντιμετωπίσει κάθε μορφή ηλεκτρονικής πλατφόρμας, η οποία τίθεται να διαχειριστεί ένα μεγάλο όγκο δεδομένων. Πιο συγκεκριμένα, ο στόχος των ηλεκτρονικών πλατφορμών, πλέον έχει μετατεθεί, καθώς σε συνδυασμό με τα πολυάριθμα και ποιοτικά δεδομένα, που θα πρέπει να παρέχουν, εξίσου σημαντικό ειναι να αποτελούν μια πλατφόρμα που να μπορει να ανταποκρίνεται σωστά στις αναζητήσεις των χρηστών της, αλλα και να γίνεται σωστή διακριτοποίηση των πληροφοριών/προϊόντων που παρέχει. Σε αυτό το σημείο βρίσκονται τα Συστήματα Σύστασης (Recommender Systems). Το Youtube, η Amazon, το Netflix είναι μόνο μερικά παραδείγματα πλατφορμών που έχουν κερδίσει έδαφος τις τελευταίες δεκαετίες, καθώς παρά τον τεράστιο όγκο δεδομένων, με τα ισχυρά συστήματα σύστασης που κατέχουν, έχουν καταφέρει να μετατρέψουν αυτό το χάος σε μία εύκολα διαχειρίσιμη πληροφορία για το χρήστη.

Σημαντικά πιο δύσκολη κρίνεται η κατάσταση στην οποία τα Συστήματα Σύστασης τίθενται να διαχειριστούν δεδομένα κειμένου, που δεν έχουν κάποια ξεκάθαρη διακριτοποί-

ηση, αλλά αυτή πρέπει να επέλθει μέσα από το κείμενο.

1.1 Αντικείμενο της διπλωματικής

Αντικείμενο της παρούσας διπλωματικής εργασίας αποτελεί η δημιουργία ενός μη προσωποποιημένου συστήματος σύστασης βιβλίων βασισμένο μόνο στις θεματικές ενότητες, που είναι λανθάνουσες στο κείμενο περιγραφής, σε συνδυασμό με τον τίτλο του βιβλίου. Συγκεκριμένα, με χρήση του μοντέλου Λανθάνουσας Κατανομης Dirichlet (Latent Dirichlet Allocation), προσπαθεί να ανακαλύψει τις θεματικές ενότητες που είναι κρυμμένες, και να αποδώσει σε κάθε κείμενο μία κατανομή ως προς αυτές. Με βάση τις κατανομές αυτές και με τη χρήση της απόστασης Jensen-Shannon γίνεται η εύρεση παρόμοιων κειμένων και τελικώς η σύσταση.

Επίσης, αναλύονται ποικίλοι μέθοδοι αξιολόγησης του μοντέλου LDA, όπως η Συνάφεια Θέματος, αξιολόγηση με χρήση Word2vec μοντέλα καθώς και αξιολόγηση μέσω υπολογισμού της στατιστικής απόστασης των θεματικών ενοτήτων, προκειμένου να παραχθεί ένα ικανοποιητικό ΣΣ.

Τέλος, αυτό το ΣΣ, οργανώνεται σε μία εύχρηση εφαρμογή για το χρήστη προκειμένου να πραγματοποιείται η σύσταση βιβλίων, χρησιμοποιώντας τόσο μία βάση δεδομένων για βιβλία, όσο μία βάση δεδομένων από βαθμολογήσεις χρηστών στα βιβλία.

1.2 Συναφείς Εργασίες

Η ιδέα χρήσης μοντέλων που ανακαλύπτουν θεματικές ενότητες που είναι κρυμμένες στα κείμενα είναι ιδιαίτερα διαδεδομένη, λόγω ίσως του γεγονότος οτι υπάρχουν απεριόριστες πηγές πληροφορίας σε μορφές κειμένου. Επομένως, έχει αποτελέσει αντικείμενο απασχόλησης για πολλούς ερευνητές.

Συγκεκριμένα, οι Konstantinos Christidis και Gregoris Mentzas [13] υλοποίησαν με χρήση του μοντέλου LDA έναν ιστότοπο δημοπρασιών που συστήνει στους αγοραστές προϊόντα παρόμοια σε περιεχόμενο με την αναζήτη τους, αλλά και στους πωλητές παρόμοια προϊόντα με αυτά που θέλουν να πουλήσουν προκειμένου να τους βοηθήσουν και στην τιμολόγηση του προϊόντος αλλά και στην εύρεση κατάλληλων λέξεων για την περιγραφή του.

Στη συνέχεια, οι Rel Guzman Apaza, Elizabeth Vera Cervantes, Laura Cruz Quispe και

José Ochoa Luna σε έρευνα που διεξήγαγαν στο National University of St. Agustin, Arequipa - Perù [14] δημιούργησαν ένα πρόγραμμα σύστασης διαθέσιμων μαθημάτων που βρίσκονται σε ηλεκτρονικούς ιστοτόπους, όπως Coursera, Udacity, Edx κ.λπ. Αυτή η σύσταση επιτεύχθηκε με χρήση του μοντέλου LDA, που ανακάλυπτε θεματικές ενότητες από το περιεχόμενο που δίνεται σε ένα αναλυτικό πρόγραμμα μαθημάτων κολεγίου.

Επίσης, οι Zhiqiang He, Zhongyi Wu, Bochong Zhou, Lei Xu και Weifeng Zhang [15] πραγματοποίησαν μία έρευνα ως προς τη συμμετοχή του μοντέλου LDA σε συστήματα σύστασης για τουριστικές διαδρομές. Πιο λεπτομερώς, ο μεγάλος όγκος ταξιδιωτικών πληροφοριών σε μορφή κειμένου, που είναι διαθέσιμος στο διαδίκτυο και η τάση των ανθρώπων να δημιουργούν ταξίδια προσαρμοσμένα στις δικές τους ανάγκες οδήγησε στην υλοποίηση ενός τέτοιου ΣΣ, που επέφερε τελικά και θετικά αποτελέσματα ως προς την αποτελεσματικότητα του. Βέβαια, η σύσταση που πραγματοποίησαν επεκτάθηκε και σε σύσταση με συνεργατικό φιλτράρισμα, δηλαδή με συσχέτιση των χρηστών και των προτιμήσεων τους.

Τέλος μία αρκετά όμοια έρευνα με αυτήν της παρούσας διπλωματικής εργασίας υπήρξε η έρευνα που διεξήγαγαν οι Mr. Dhiraj Vaibhav Bagul και Dr. Sunita Barve [16]. Πιο λεπτομερώς εκπαίδευσαν ένα μοντέλο LDA πάνω σε επιστημονικά άρθρα, για τα οποία γνώριζαν τους εκδοτικούς οίκους τους. Με βάση αυτό το μοντέλο και τη μετρική Jensen-Shannon, δημιούργησαν ένα σύστημα σύστασης, στο οποίο ο χρήστης με την εισαγωγή της περίληψης του, του προτείνονται οι καλύτεροι εκδοτικοί οίκοι, προκειμένου να δημοσιεύσει το κείμενο του. Η χρήση του μοντέλου LDA συγκρίθηκε με την εκδοχή σύστασης με ομοιότητα συνημιτόνου και αποδήχτηκε πως το LDA μοντέλο ξεπέρασε την επίδοση της ομοιότητας συνημιτόνου για σύσταση με βάση το περιεχόμενο.

1.3 Οργάνωση του τόμου

Ακολουθεί μία σύντομη παρουσίαση των κεφαλαίων και του περιεχομένου που αυτά καλύπτουν. Συγκεκριμένα, στο *Κεφάλαιο 2*, αναλύονται βασικές θεωρητικές και μαθηματικές έννοιες, απαραίτητες για την κατανόηση αυτής της διπλωματικής εργασίας. Στη συνέχεια, στο *Κεφάλαιο 3*, αναφέρονται πληροφορίες σχετικά με το περιβάλλον και τα εργαλεία που χρησιμοποιήθηκαν προκειμένου να αναπτυχθεί το προγραμματιστικό κομμάτι αυτής της διπλωματικής. Ακολουθεί το *Κεφάλαιο 4*, στο οποίο περιγράφεται η διαδικασία συλλογής, οργάνωσης και επεξεργασίας των δεδομένων. Έπειτα, στο *Κεφάλαιο 5* αναπτύσσονται μέθοδοι

για την αξιολόγηση του μοντέλου LDA. Συνέχεια αποτελεί το *Κεφάλαιο 6*, στο οποίο πραγματοποιείται ο σχεδιασμός και η ανάπτυξη του μοντέλου LDA σε προγραμματιστικό επίπεδο, καθώς επίσης και η επιλογή του καλύτερου μοντέλου, βάσει των μετρικών αξιολόγησης που αναπτύχθηκαν στο Κεφάλαιο 5. Στο *Κεφάλαιο 7* γίνεται η παρουσίαση της εφαρμογής που δημιουργήθηκε και πιο συγκεκριμένα παρουσιάζονται τόσο οι λειτουργίες της εφαρμογής, όσο και η διαδικασία που υλοποιείται παρασκηνιακά για κάθε λειτουργία, προκειμένου να πραγματοποιηθεί η σύσταση. Επίσης παρουσιάζονται και κάποια αποτελέσματα. Τέλος, η παρούσα διπλωματική εργασία κλείνει με το *Κεφάλαιο 8* στο οποίο διατυπώνονται κάποιες παρατηρήσεις, καθώς επίσης και ιδέες για βελτίωση και μελλοντική επέκταση της εργασίας.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

Στο παρόν κεφάλαιο ορίζονται κάποιες βασικές έννοιες που θα χρησιμοποιηθούν στη συνέχεια, προκειμένου να γίνει μία βαθύτερη κατανόηση του αντικειμένου που πραγματεύεται η παρούσα διπλωματική εργασία σε θεωρητικό και σε μαθηματικό επίπεδο.

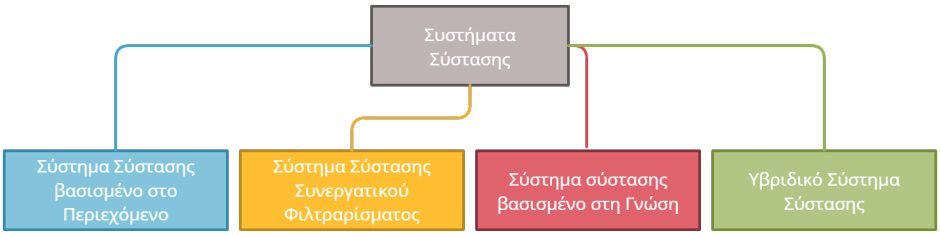
2.1 Θεωρία συστημάτων σύστασης

Τα Συστήματα Σύστασης αποτελούν υποκατηγορία των συστημάτων φιλτραρίσματος πληροφορίας [17]. Συγκεκριμένα, αλληλεπιδρούν με έναν τεράστιο όγκο πληροφοριών / δεδομένων και στο τέλος αυτής της διαδικασίας συστήνουν ή αλλιώς προτείνουν ένα ή περισσότερα αντικείμενα στους χρήστες, πάντα συνυφασμένα με τις επιθυμίες του καθενός. Αυτή η αναγκαιότητα που επικρατεί για δημιουργία και χρήση ικανών ΣΣ, που να εκμεταλλεύονται όλα τα διαθέσιμα δεδομένα, έχει οδηγήσει στην κατηγοριοποίηση τους σε:

- Σύστημα Σύστασης βασισμένο στο Περιεχόμενο (Content Based RS)
- Σύστημα Σύστασης Συνεργατικού Φιλταρίσματος (Collaborative Filtering RS)
- Σύστημα Σύστασης βασισμένο στη Γνώση (Knowledge Based RS)
- Υβριδικό Σύστημα Σύστασης (Hybrid RS)

2.1.1 Σύστημα σύστασης βασισμένο στο περιεχόμενο

Τα ΣΣ βασισμένα στο περιεχόμενο, αποτελούν συστήματα που η τελική τους σύσταση στηρίζεται τόσο στις παρελθοντικές τάσεις και προτιμήσεις των χρηστών, όσο και στα χα-



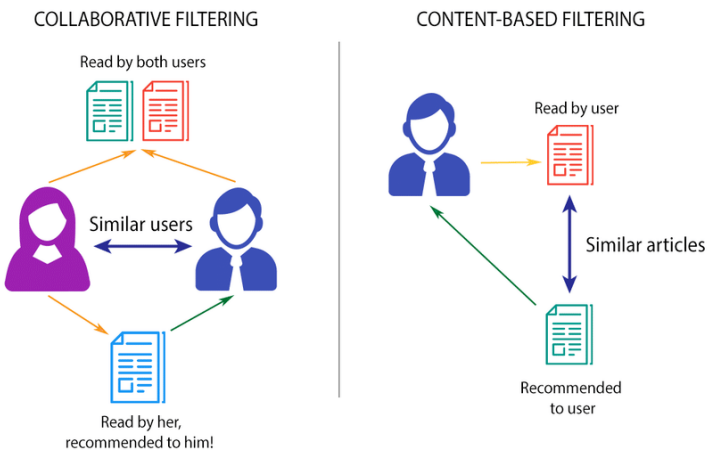
Σχήμα 2.1: Κατηγορίες Συστημάτων Σύστασης

ρακτηριστικά των αντικειμένων που προτίμησαν. Συγκεκριμένα, αν το προφίλ ενός χρήστη δηλώνει την αρέσκεια του ως προς τα βιβλία μαγειρικής, τότε η σύσταση με βάση το περιεχόμενο θα του επιστρέφει αποτελέσματα σχετικά με αυτήν την κατηγορία [17, 18].

Σε αυτό το σημείο όμως είναι σημαντικό να επισημανθούν δύο σημαντικά προβλήματα. Το πρώτο είναι η ύπαρξη δεδομένων που να αντιπροσωπεύουν σε ικανοποιητικό βαθμό αυτό που δηλώνει το αντικείμενο, προκειμένου να μπορούν να διακριτοποιηθούν. Το δεύτερο ζήτημα ονομάζεται Ψυχρή Εκκίνηση (Cold Start) και συναντάται στις περιπτώσεις που πρέπει να πραγματοποιηθεί σύσταση σε έναν καινούριο χρήστη. Όπως γίνεται κατανοητό, το προφίλ των προτιμήσεων ενός νέου χρήστη θα είναι κενό, με αποτέλεσμα να μην μπορούν να προκύψουν κάποια συμπεράσματα ως προς την αρέσκεια του, προκειμένου να του προταθούν αντικείμενα παρόμοιων χαρακτηριστικών [17, 18].

2.1.2 Σύστημα σύστασης συνεργατικού φιλταρίσματος

Στην επόμενη, εξίσου σημαντική κατηγορία, ανήκουν τα ΣΣ Συνεργατικού Φιλτραρίσματος. Αυτά τα συστήματα προκειμένου να λειτουργήσουν απαιτούν την ύπαρξη ιστορικού χρηστών σχετικά με τις προτιμήσεις τους ως προς τα διαθέσιμα αντικείμενα. Πιο λεπτομερώς, η σύσταση τους βασίζεται στην παραδοχή ότι αν δύο χρήστες συμφωνούσαν στο παρελθόν θα συμφωνούν και στο μέλλον και επομένως με βάση την ομοιότητα, ως προς τις προτιμήσεις, ενός χρήστη με άλλους προσπαθεί να προβλέψει την βαθμολογία που θα έβαζε σε ένα αντικείμενο προκειμένου να το προτείνει ή οχι. Το πρόβλημα της Ψυχρής Εκκίνησης εμφανίζεται και σε αυτό το ΣΣ, παράλληλα και με άλλα προβήματα που σχετίζονται με το πλήθος των χρηστών και το πλήθος των αξιολογήσεων τους [18, 19].



Σχήμα 2.2: Σύστημα σύστασης συνεργατικού φιλτραρίσματος (αριστερά) και σύστημα σύστασης βασισμένο στο περιεχόμενο (δεξιά) [1]

2.1.3 Σύστημα σύστασης βασισμένο στη γνώση

Άλλη μία κατηγορία που δεν την ακούμε συχνά, καθώς συνήθως υλοποιείται στις περιπτώσεις που δεν μπορούν να πραγματοποιηθούν οι δύο παραπάνω κατηγορίες ΣΣ, είναι τα ΣΣ βασισμένα στη Γνώση. Πιο συγκεκριμένα, αυτά τα ΣΣ δε βασίζονται στο ιστορικό προτιμήσεων του χρήστη, αλλά ζητάει από τον ίδιο τον χρήστη να εισάγει τις προτιμήσεις του, είτε με κάποιο φιλτράρισμα είτε με κάποια εισαγωγή παραδείγματος, είτε με συμπλήρωση πεδίων, προκειμένου να του συσταθούν προτάσεις συνυφασμένες με αυτά που εισήγαγε. Ακριβώς επειδή δεν απαιτείται το προφίλ του χρήστη για να του συσταθούν, είναι και ΣΣ που δεν αντιμετωπίζουν το πρόβλημα της Ψυχρής Εκκίνησης, απο την άλλη πλευρά όμως προκειμένου η σύσταση να κριθεί ποιοτική απαιτείται σαφή αποτύπωση των επιθυμητών χαρακτηριστικών από τον χρήστη. Για να γίνει πιο κατανοητό, τα ΣΣ βασισμένα στη Γνώση τα συναντώνται συχνά σε ηλεκτρονικές πλατφόρμες για αγορά αυτοκινήτων, πλατφόρμες για ενοικίαση/αγορά σπιτιου κ.λπ. [20, 21].

2.1.4 Υβριδικό σύστημα σύστασης

Μία τελευταία κατηγορία ΣΣ, αποτελούν τα Υβριδικά ΣΣ. Όπως διαπιστώθηκε και από τις παραπάνω κατηγορίες, κάθεμια λειτουργεί για διαφορετικούς σκοπούς και απαιτεί διαφορετικά δεδομένα. Ως Υβριδικά ΣΣ ορίζεται η συνεργατική δράση του ΣΣ Συνεργατικού Φιλτραρίσματος, του ΣΣ βασισμένο στο Περιεχόμενο και άλλων προσεγγίσεων σύστασης. Οι μελέτες δείχνουν πως αυτά τα Υβριδικά Συστήματα επιτυγχάνουν καλύτερες συστάσεις

από την μεμονομένη λειτουργία των παραπάνω ΣΣ [17].

2.2 Μη εποπτευόμενη μηχανική μάθηση

Η Μηχανική Μάθηση (Machine Learning) είναι ένας κλάδος της επιστήμης των υπολογιστών και της τεχνητής νοημοσύνης, που αξιοποιώντας τα δεδομένα, που είναι διαθέσιμα, προσπαθεί μέσω μοντέλων και αλγορίθμων να προσεγγίσει τον τρόπο με τον οποίο μαθαίνουν οι άνθρωποι [22]. Υπάγεται σε κατηγορίες, μία εξ αυτών να αποτελεί η Μη Εποπτευόμενη Μηχανική Μάθηση (Unsupervised Machine Learning). Συγκεκριμένα, ο όρος "Μη Εποπτευόμενη Μηχανική Μάθηση" είναι συνυφασμένος με μοντέλα τα οποία εκπαιδεύονται πάνω σε δεδομένα που δεν έχουν κάποιο είδος ετικέτας, δηλαδή δεν απαιτεί δεδομένα εξόδου κατά τη φάση εκπαίδευσης, αρκείται μόνο στα δεδομένα εισόδου. Επομένως, τα Unsupervised μοντέλα δεν εκπαιδεύονται προκειμένου να ανακαλύψουν τη σχέση μεταξύ εισόδου και εξόδου, αλλά προσπαθούν βασισμένα στα δεδομένα εισόδου να ανακαλύψουν κοινα χαρακτηριστικά. Για τους παραπάνω λόγους, είναι κατάλληλα για τον εντοπισμό αόρατων τάσεων και σχέσεων μεταξύ των ίδιων των δεδομένων, όπως επίσης και για την ομαδοποίηση των δεδομένων σε συγκεκριμένο αριθμό ομάδων βάσει μετρικων αξιολόγησης [23].

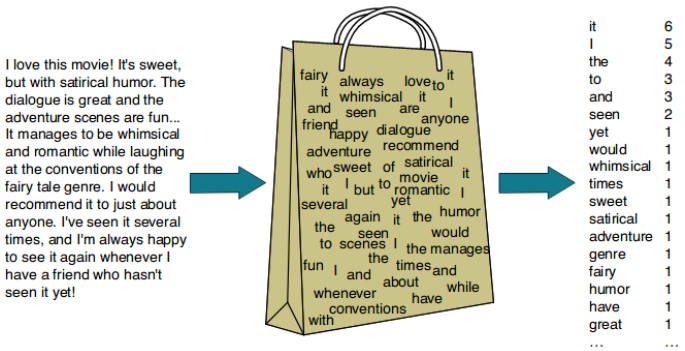
2.3 Μοντέλα θέματος

Τα Μοντέλα Θέματος είναι μία Μη Εποπτευόμενη τεχνική της Μηχανικής Μάθησης, που ως βασικό στόχο έχουν την ανακάλυψη κρυμμένων θεματικών ενοτήτων στα δεδομένα. Συχνά αναφέρονται και ως Πιθανοτικά Μοντέλα Θέματος (Probabilistic Topic Models), επειδή ακριβώς για την ανακάλυψη των λανθάνοντων σημασιολογικών ενοτήτων στα δεδομένα, χρησιμοποιούν αλγορίθμους της στατιστικής. Τα δεδομένα αυτά πέρα από δεδομένα κειμένου, μπορεί να είναι εικόνες, γενετικές πληροφορίες, η παρούσα διπλωματική εργασία όμως επικεντρώνεται στα κείμενα. Τα Μοντέλα Θέματος στηρίζονται στην παραδοχή, ότι ένα κείμενο ανταποκρίνεται σε μια θεματική ενότητα και στην ίδια θεματική ανήκουν κατά πλειοψηφία και οι λέξεις από τις οποίες αποτελείται αυτό. Για παράδειγμα, σε ένα κείμενο σχετικό με τον Αθλητισμό, πολύ πιθανόν είναι να εντοπιστούν λέξεις όπως "άθλημα", "αγώνας" και "προπονητής", που όπως είναι κατανοητό αντιπροσωπεύουν αυτή τη θεματολογία, χωρίς βέβαια να σημαίνει οτι συναντόνται μόνο σε αυτήν [24].

Επομένως, τα Μοντέλα Θέματος θέτουν όλη αυτή τη διαδικασία δημιουργίας συνόλων όμοιων λέξεων σε ένα μαθηματικό πλαίσιο προκειμένου να μπορεί να επιτευχθεί σε ένα σύνολο δεδομένων κειμένου, με σκοπό την ανακάλυψη της κατανομής θέματος σε καθένα από αυτά [25]. Ενα μοντέλο που είναι άμεσα συνυφασμένο με όσα περιγράφηκαν είναι το μοντέλο της Λανθάνουσας Κατανομής Dirichlet (Latent Dirichlet Allocation), που είναι και το βασικό μοντέλο της παρούσας διπλωματικής εργασίας.

2.3.1 Υπόθεση Bag of Words

Δύο βασικά προβλήματα με τα οποία έρχονται αντιμέτωποι οι αλγόριθμοι Μηχανικής Μάθησης όταν συναντούν δεδομένα κειμένου, και επομένως και τα μοντέλα θέματος, είναι η ακαταστασία που υπάρχει στα κείμενα, καθώς προτιμούν μία καθορισμένη είσοδο σταθερού μήκους, όπως επίσης και οι λέξεις έναντι των αριθμών που πρέπει να δέχονται ως εισοδο. Σε αυτά τα προβλήματα έρχονται να δώσουν λύση διάφορες τεχνικές, μία εκ των οποίων είναι η Υπόθεση Bag of Words (BoW) [26].



Σχήμα 2.3: Απεικόνιση υπόθεσης Bag Of Words [2]

Από την ονομασία του γίνεται κατανοητό ότι αντιμετωπίζει κάθε κείμενο σαν μία τσάντα από λέξεις χωρίς να δίνει σημασία στη σειρά ή τη δομή των λέξεων, παρα μόνο στο αν εμφανίζονται ή όχι μέσα στο κείμενο (βλπ. Σχήμα 2.3) [26]. Περιλαμβάνει δύο πράγματα:

- Λεξιλόγιο: κάθε λέξη έχει τον δικό της αναγνωριστικό κωδικό
- Μετρική καταμέτρησης της συχνότητας εμφάνισης των λέξεων

2.3.2 Εισαγωγή στις πιθανοτικές έννοιες του μοντέλου Λανθάνουσας Κατανομής Dirichlet

Η υποενότητα αυτή αποτελεί μία εισαγωγή σε πιθανοτικές έννοιες, απαραίτητες, προκειμένου να γίνει πλήρως αντιληπτό το επόμενο υποκεφάλαιο επεξήγησης του μοντέλου Λανθάνουσας Κατανομής Dirichlet (Latent Dirichlet Allocation).

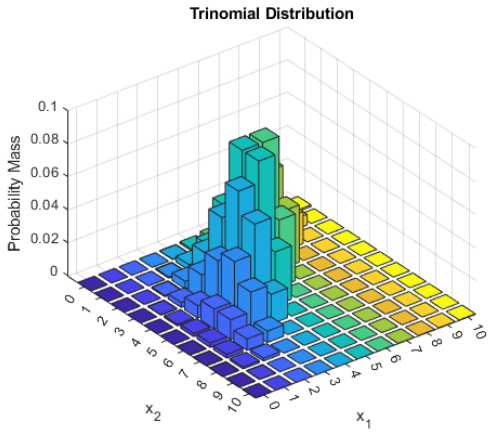
Θεώρημα Bayes

Το θεώρημα του Bayes [27] είναι ευρέως γνωστό στον τομέα των πιθανοτήτων και της στατιστικής. Συγκεκριμένα, εκφράζει την πιθανότητα να συμβεί ένα γεγονός, δεδομένου ότι ένα άλλο γεγονός ισχύει. Ο τύπος που αποτυπώνει στατιστικά αυτό που μόλις διατυπώθηκε δίνεται από τη σχέση (2.1)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, P(B) \neq 0 \tag{2.1}$$

όπου $P(A|B)$ είναι η πιθανότητα να συμβεί το γεγονός A, δεδομένου ότι το B είναι αληθές και ονομάζεται μεταγενέστερη πιθανότητα(posterior probability), $P(B|A)$ η πιθανότητα να συμβεί το γεγονός B, δεδομένου ότι το A είναι αληθές και $P(A), P(B)$ εκφράζουν τη πιθανότητα να συμβεί το A και B γεγονός αντίστοιχα, χωρίς κάποια δέσμευση.

Πολυωνυμική κατανομή



Σχήμα 2.4: Απεικόνιση Πολυωνυμικής Κατανομής (Multinomial Distribution) για k=3 [3]

Στη θεωρία των πιθανοτήτων η Πολυωνυμική Κατανομή (Multinomial Distribution) [28, 29] αποτελεί γενίκευση της Διωνυμικής Κατανομής (Binomial Distribution). Χρησιμοποιεί-

ται για την πρόβλεψη του αποτελέσματος μιας σειράς επαναληπτικών δοκιμών, με κάθε δοκιμή να είναι εντελώς τυχαία και ανεξάρτητη και να έχει μία δική της πιθανότητα να συμβεί.

Πιο συγκεκριμένα, έστω k σταθερός, πεπερασμένος αριθμός, που υποδηλώνει τα αμερόληπτα πιθανά ενδεχόμενα, με πιθανότητες για κάθε ενδεχόμενο $p_1, p_2, \ldots, p_k$ ($p_i \geq 0$ για $i = 1, \ldots, k$ και $\sum_{i=1}^{k} p_i = 1$), n ο αριθμός των αμερόληπτων δοκιμών και X_i τυχαία μεταβλητή που υποδηλώνει τον αριθμό των φορών, που παρατηρείται ο αριθμός έκβασης του ενδεχομένου i στις n δοκιμές. Τότε το διάνυσμα τυχαίων μεταβλητών $X = X_1, \ldots, X_k$ ακολοθεί μία Πολυωνυμική Κατανομή.

Η Συνάρτηση Μάζας Πιθανότητας (PMF) είναι μία συνάρτηση που δίνει την πιθανότητα ότι μια διακριτή τυχαία μεταβλητή είναι ακριβώς ίση με κάποια τιμή ($P(X = x)$). Η PMF της Πολυωνυμικής Κατανομής δίνεται από τον παρακάτω τύπο

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \frac{n!}{x_1! x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}, \tag{2.2}$$

όπου $\sum_{i=1}^{k} x_i = n$

Dirichlet κατανομή

Η Dirichlet Κατανομή (Dirichlet Distribution) [30] είναι μία γενίκευση της Beta Κατανομής (Beta Distribution). Η Συνάρτηση Πυκνότητας Πιθανότητας (PDF) είναι μία συνάρτηση, που χρησιμοποιείται για τον καθορισμό της πιθανότητας η συνεχής τυχαία μεταβλητή να εμπίπτει σε ένα συγκεκριμένο εύρος τιμών:

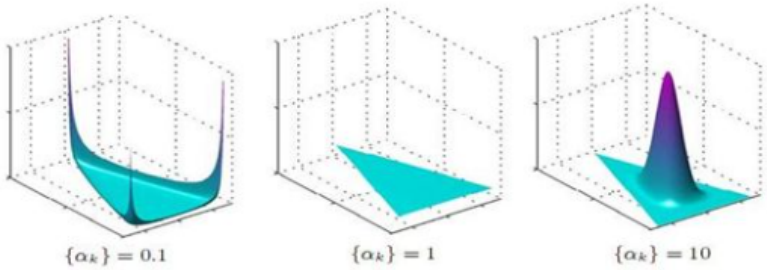
$$P(a \leq X \leq b) = \int_a^b f_x \, dx. \tag{2.3}$$

Η PDF της Dirichlet Κατανομής δίνεται από τον τύπο:

$$f(x_1, \ldots, x_k | \alpha_1 \ldots \alpha_k) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K} \alpha_i)} \prod_{i=1}^{K} x_i^{\alpha_i - 1} = \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}, \tag{2.4}$$

όπου $\sum_{i=1}^{k} x_i = 1$ με $x_i \in [0, 1] \forall i$ και $\alpha = (\alpha_1, \ldots, \alpha_k)$. Σε περιπτώσεις που το $\alpha_1 = \alpha_2 = \ldots = \alpha_k$ τότε η κατανομή ονομάζεται συμμετρική.

Σε αυτό το σημείο είναι σημαντική μία αναφορά στο Σχήμα 2.5. Συγκεκριμένα, παρατηρείται ότι για $\alpha_i < 1$ απωθεί το x_i προς τα άκρα, για $\alpha_i > 1$ έλκει το x_i προς κάποια κεντρική τιμή, ενώ για $\alpha_i = 1$ δημιουργείται μία Κανονική Κατανομή.



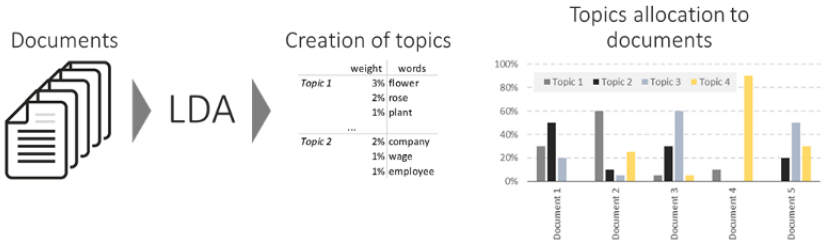
Σχήμα 2.5: Dirichlet Κατανομή (Dirichlet Distribution) για K=3 και απεικόνιση μεταβολής της κατανομής για διαφορετικές τιμές του α [4]

2.3.3 Λανθάνουσα Κατανομή Dirichlet

Ο βασικός αλγόριθμος της παρούσας διπλωματικής εργασίας είναι ο αλγόριθμος της Λανθάνουσας Κατανομής Dirichlet (Latent Dirichlet Allocation-LDA). Είναι ένα μοντέλο που όπως έχει αναφερθεί στο τέλος της εισαγωγής της ενότητας 2.3, ανήκει στις τεχνικές Πιθανοτικής Μοντελοποίησης Θέματος, επομένως ισχύουν τα χαρακτηριστικά, όπως το οτι είναι Μη Εποπτευόμενο μοντέλο, και οι υποθέσεις, που είναι άμεσα συνυφασμένες με αυτά.

Αρχικά, η Λανθάνουσα Κατανομή Dirichlet εισήχθη για πρώτη φορά από τους Dr. David Blei, Andrew Ng και Michael Jordan, οι οποίοι περιέγραψαν αυτό το μοντέλο ως «Ένα γενεσιουργό πιθανολογικό μοντέλο για συλλογές διακριτών δεδομένων, όπως τα δεδομένα κειμένου »[31]. Το LDA είναι ένα Μπεϋζιανό μοντέλο πιθανοτήτων τριών επιπέδων, λέξεων, θεμάτων και κειμένων, που το όνομά του προδιαθέτει για τη λειτουργία του και συγκεκριμένα η λέξη “Latent” αναφέρεται στις θεματικές ενότητες που είναι λανθάνουσες, δηλαδή κρυφές δομές μέσα στο κείμενο, ενώ η λέξη “Dirichlet”, δηλώνει την κατανομή Dirichlet με βάση την οποία καθορίζονται οι κατανομές των θεματικών ενοτήτων στα κείμενα και των λέξεων στις θεματικές ενότητες [32]. Βασικός στόχος του αποτελεί η ανακάλυψη των θεματικών ενοτήτων που είναι κρυμμένες στα κείμενα και στη συνέχεια η ανάθεση μίας κατανομής ως προς τις θεματικές ενότητες για κάθε κείμενο, βασιζόμενο στην κατανομή των λέξεων σε αυτό (βλπ. Σχήμα 2.6)

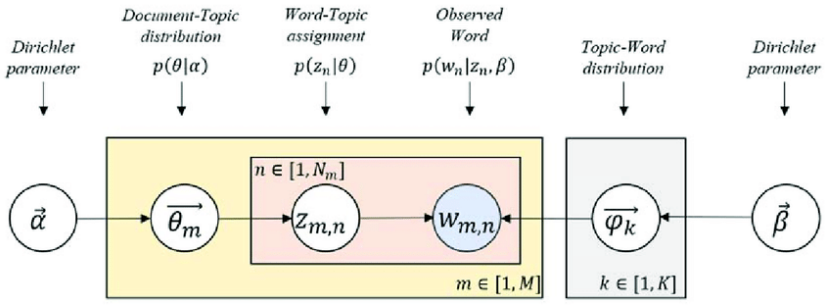
Η παραδοχή στην οποία στηρίζεται η Λανθάνουσα Κατανομή Dirichlet είναι ότι η συγγραφή ενός κειμένου αρχίζει με τον προκαθορισμό των θεματικών ενοτήτων που θα προσεγγισθούν. Στη συνέχεια, με βάση αυτήν την κατανομή γίνεται η επιλογή των λέξεων από τις δεξαμενές λέξεων των αντίστοιχων θεματικών ενοτήτων. Επίσης, κάθε κείμενο απαντά σε μία Dirichlet Κατανομή από θεματικές ενότητες και κάθε θεματική ενότητα σε μία Dirichlet Κατανομή από τις λέξεις. Όλη αυτήν την παραδοχή για το πώς προέκυψαν τα κείμενα



Σχήμα 2.6: Αναπαράσταση απόδοσης κατανομής θέματος σε κείμενο [5]

της συλλογής, ο LDA την έχει υπάγει σε ένα στατιστικο μοντέλο, τα βήματα του οποίου αναγράφονται στον Αλγόριθμο 1 [16, 31, 33, 34, 35, 36].

Συγκεκριμένα, ο Αλγόριθμος 1 ουσιαστικά αυτό που περιγράφει είναι ότι για ένα κείμενο Ν λέξεων γίνεται επιλογή μίας κατανομής θεμάτων θ σύμφωνα με την Dirichlet κατανομή και για κάθε θεματική ενότητα επιλογή μίας κατανομής λέξεων ϕ, βασισμένη και αυτήν στην Dirichlet Κατανομή. Στη συνέχεια, για κάθε πιθανή λέξη w που θα ενταχθεί στο κείμενο, πρώτα διαλέγεται η θεματική ενότητα στην οποία θα ανήκει η λέξη z, βάσει του αποτελέσματος μίας Multinomial Κατανομής στην ήδη υπάρχουσα κατανομή θεματικών ενοτήτων του κειμένου(θ). Έπειτα, βάσει αυτής της θεματικής ενότητας (z) και μέσω μίας Multinomial Κατανομής στην κατανομή θεματικής ενότητας-λέξεις (ϕ) προκύπτει η λέξη που θα ενταχθεί τελικά στο κείμενο w (βλπ. Σχήμα 2.7).



Σχήμα 2.7: Γραφική Αναπαράσταση Διαδικασίας Δημιουργίας Κειμένου LDA [6]

Τέλος, ο τύπος (2.5) υπολογίζει την πιθανότητα να παραχθεί ένα συγκεκριμένο κείμενο, βάσει της παραπάνω διαδικασίας.

$$P(\mathbf{W}, \mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\phi}; \alpha, \beta) = \prod_{i=1}^{K} P(\phi_i; \beta) \prod_{j=1}^{M} P(\theta_j; \alpha) \prod_{t=1}^{N} P(Z_{j,t}|\theta_j) P(W_{j,t}|\phi_{z_{j,t}}) \tag{2.5}$$

Αυτή η μεταγενέστερη πιθανότητα, δεδομένου βέβαια ενός κειμένου, αποτελεί ένα δύσκολο κομμάτι υπολογισμού και υπάρχουν αρκετοί αλγόριθμοι που προσπαθούν να προσεγ-

Algorithm 1 Αλγόριθμος Διαδικασίας Δημιουργίας Κειμένου LDA

Παράμετροι:

M: Ο αριθμός των κειμένων

N: Ο αριθμός των λέξεων σε ένα δωθέν κείμενο (το κείμενο i περιέχει N_i λέξεις)

K: Ο αριθμός των θεματικών ενοτήτων

$\boldsymbol{\alpha}$ **:** Είναι η παράμετρος Dirichlet Κατανομής, για την κατανομή κείμενο-θεματική ενότητα

$\boldsymbol{\beta}$ **:** Είναι η παράμετρος Dirichlet Κατανομής, για την κατανομή θεματική ενότητα-λέξεις

$\boldsymbol{\theta_i}$ **:** Είναι η κατανομή θεμάτων για το κείμενο i

$\boldsymbol{\varphi_k}$ **:** Είναι η κατανομή λέξεων για τη θεματική ενότητα i

$\mathbf{z_{ij}}$: Είναι η θεματική ενότητα για τη j λέξη στο i κείμενο

$\mathbf{w_{ij}}$: Είναι η συγκεκριμένη λέξη

Βήματα Αλγορίθμου:

1: Επιλέξτε $N_i \sim$ Poisson(ξ) όπου $i \in [1, \ldots, M]$

2: Επιλέξτε $\theta_i \sim$ Dir(α) όπου $i \in [1, \ldots, M]$ και $\mathrm{Dir}(\alpha)$ είναι Dirichlet Κατανομή με συμμετρικό α και τυπικά $\alpha < 1$

3: Επιλέξτε $\varphi_k \sim$ Dir(β) όπου $k \in [1, \ldots, K]$ και $\mathrm{Dir}(\beta)$ είναι Dirichlet Κατανομή με συμμετρικό β και τυπικά $\beta < 1$

4: **for** Για κάθε λέξη $w_{i,j}$ με $i \in [1, \ldots, M]$ και $j \in [1, \ldots, N_i]$ **do**

5: Επιλέξτε ένα θέμα $z_{i,j} \sim \mathrm{Multinomial}(\theta_i)$.

6: Επιλέξτε μία λέξη $w_{i,j} \sim \mathrm{Multinomial}(\varphi_{z_{i,j}})$.

7: **end for**

γίσουν τη λύση όσο καλύτερα γίνεται. Παρακάτω θα γίνει μία αναφορά σε δύο από αυτούς τους αλγορίθμους:

- Gibbs Sampling
- Variational Bayesian Inference

Gibbs Sampling

Η δειγματοληψία Gibbs είναι μία τεχνική της αλυσίδας Markov Monte Carlo που μπορεί να χρησιμοποιηθεί για την εκτίμηση παραμέτρων. Συγκεκριμένα, η παράμετρος που χρήζει περισσότερης σημασίας να εκτιμηθεί στο μοντέλο LDA είναι η Z, δηλαδή η απόδοση θεματικών ενοτήτων στις λέξεις των κειμένων (όπου $z_{i,j}$ σημαίνει η θεματική ενότητα j στη λέξη i), καθώς μέσα από αυτήν μπορούν να υπολογιστούν οι θ και ϕ [36, 37]. Αυτή η οπτική

Algorithm 2 Collapsed Gibbs sampler

Παράμετροι:

M: Ο αριθμός των κειμένων

N: Ο αριθμός των λέξεων σε ένα δωθέν κείμενο (το κείμενο i περιέχει N_i λέξεις)

K: Ο αριθμός των θεματικών ενοτήτων

$\boldsymbol{\alpha}$ **:** Είναι η παράμετρος Dirichlet Κατανομής, για την κατανομή κείμενο-θεματικές ενότητες

$\boldsymbol{\beta}$ **:** Είναι η παράμετρος Dirichlet Κατανομής, για την κατανομή θεματική ενότητα-λέξεις

$\mathbf{z_{ij}}$: Είναι η θεματική ενότητα για τη j λέξη στο i κείμενο

$\mathbf{w_{ij}}$: Είναι η j λέξη στο i κείμενο

$\mathbf{n_{k,v}}$: Είναι οι φορές που η λέξη που αντιστοιχίζεται στη λέξη w_{ij}, η w_v έχει αποδωθεί στη θεματική ενότητα k

$\mathbf{n_{d,k}}$: Μετράει τις φορές που έχει αποδωθεί η θεματική ενότητα k στο d κείμενο

Βήματα Αλγορίθμου:

1: Αρχικοποίηση: Τυχαία επιλογή θεματικής ενότητας για κάθε λέξη σε κάθε κείμενο βασισμένη σε μία Multinomial Κατανομή

2: **for** Για κάθε λέξη $w_{i,j}$ με $i \in [1, \ldots, M]$ και $j \in [1, \ldots, N_i]$ **do**

3: Υπολογισμός των μεταβλητών $n_{k,v}, n_{d,k}$

4: **for** Για κάθε $k \in [1, \ldots, K]$ **do**

5: Υπολογισμός της πιθανότητας:

$$P(z_{i,j} = k | w_{i,j} = w_v, d, \alpha, \beta) \propto \frac{n_{kv,-ij} + \beta}{\Sigma_{v'=1}^{V}(n_{kv,-ij} + \beta)} \cdot \frac{n_{dk-ij} + \alpha}{N_d + \Sigma_{k'=1}^{K}\alpha}, \tag{2.6}$$

όπου -ij σημαίνει χωρίς να υπολογίζεται στους υπολογισμούς η εμφάνιση της λέξης w_{ij}

6: **end for**

7: Ανάθεση της θεματικής ενότητας k με τη μεγαλύτερη πιθανότητα, στη λέξη $w_{i,j}$

8: **end for**

ονομάστηκε “Collapsed Gibbs sampler” και περιγράφεται στον Αλγόριθμο 2, η εξωτερική επαναληπτική διαδικασία του οποίου εκτελείται έως ότου ικανοποιηθεί η εκάστοτε συνθήκη σύγκλισης.

Η βασική ιδέα του Αλγορίθμου 2 είναι για κάθε λέξη σε κάθε κείμενο να υπολογίζει την πιθανότητα ανάθεσης μιας θεματικής ενότητας ως το γινόμενο δύο επιμέρους αναλογιών

1. Η πρώτη αναλογία εκφράζει την πιθανότητα η λέξη w_{ij} να αντιστοιχηθεί στο θέμα k και

2. η δεύτερη αναλογια εκφράζει την πιθανότητα ανάθεσης του θέματος k στο κείμενο d.

Αυτές οι δύο αναλογίες, αν δεν είναι εκφρασμένες ως προς κάποια λέξη w_{ij}, αντιπροσωπεύουν και τις μεταβλητές θ και ϕ.

Variational Bayesian Inference

Η γενική ιδέα της μεθόδου Variational Bayesian Inference [33, 34] είναι να προσεγγιστεί η μεταγενέστερη κατανομή P μέσω μιας άλλης κατανομής Q_v που περιέχει V ελεύθερες παραμέτρους. Ο στόχος είναι να βρεθεί ο κατάλληλος συνδυασμός ελεύθερων παραμέτρων που μειώνουν την απόκλιση Kullback-Leibler (KL) μεταξύ της Q_v και της P, όπου P θα είναι μία σταθερή τιμή.

Συγκεκριμένα, η μεταγενέστερη πιθανοτητα για ένα γνωστό σύνολο κειμένων έχει μετατραπεί σε $P(Z, \theta, \phi|W, \alpha, \beta)$, και οι ελεύθερες παράμετροι της Q είναι οι γ, ϕ_v, λ, όπου λ είναι μία παράμετρος Dirichlet Κατανομής που καθορίζει το ϕ (θεματική ενότητα-λέξεις κατανομη), η γ είναι κι αυτή μία παράμετρος Dirichlet Κατανομής που καθορίζει το θ (θεματική ενότητα-κείμενο κατανομη) και τέλος το ϕ_v είναι μία παράμετρος Multinomial Κατανομής που καθορίζει το Z (τη θεματική ενότητα για τη συγκεκριμένη λέξη, στο συγκεκριμένο κείμενο)

Με αυτές τις αλλαγές, το παραπάνω πρόβλημα υπάγεται σε ένα πρόβλημα βελτιστοποίησης που μπορεί να περιγραφεί μαθηματικά με τον εξής τρόπο:

$$\gamma^*, \phi_v^*, \lambda^* = argmin_{\gamma,\phi_v,\lambda} KL(q(Z, \theta, \phi|\gamma, \phi_v, \lambda) \parallel p(Z, \theta, \phi|w, \alpha, \beta)), \tag{2.7}$$

όπου KL είναι η απόκλιση Kullback-Leibler (KL), ο τύπος της οποίας βρίσκεται στη σχέση (2.9).

Επομένως, για διαφορετικό συνδυασμό των μεταβλητών $\gamma^*, \phi_v^*, \lambda$, προκύπτει και άλλη τιμή στην πιθανότητα q, και στόχος είναι η εύρεση του συνδυασμού που επιτυγχάνει μικρότερη απόσταση από την πραγματική μεταγενέστερη πιθανότητα p.

2.4 Απόκλιση Jensen-Shannon

Η Απόκλιση Jensen Shannon (Jensen–Shannon divergence) [38] συναντάται στην στατιστική και στις πιθανότητες. Συγκεκριμένα, αποτελεί μία μετρική, που υπολογίζει την ομοιότητα μεταξύ πιθανοτικών κατανομών, η τετραγωνική ρίζα της οποίας είναι γνωστή ως Από-

σταση Jensen-Shannon(Jensen-Shannon Distance). Ο υπολογισμός της ομοιότητας μεταξύ δύο κατανομών P, Q δίνεται από τον τύπο:

$$JSD(P \parallel Q) = \frac{1}{2}D(P \parallel M) + \frac{1}{2}D(Q \parallel M), \tag{2.8}$$

όπου $M = \frac{1}{2}(P + Q)$ και $D(P \parallel M), D(Q \parallel M)$ ορίζονται από την απόκλιση Kullback–Leibler (Kullback–Leibler divergence ή Relative Entropy), που δίνεται από τον τύπο:

- Για κατανομές P και Q διακριτής τυχαίας μεταβλητής:

$$D_{KL}(P \parallel Q) = \sum_{x \in X} P(x) log(\frac{P(x)}{Q(x)}) = -\sum_{x \in X} P(x) log(\frac{Q(x)}{P(x)}) \tag{2.9}$$

- Για κατανομές P και Q συνεχής τυχαίας μεταβλητής

$$D_{KL}(P \parallel Q) = \int_{-\infty}^{\infty} p(x) log(\frac{p(x)}{q(x)})\, dx = -\int_{-\infty}^{\infty} p(x) log(\frac{q(x)}{p(x)})\, dx, \tag{2.10}$$

όπου p, q είναι οι πυκνότητες πιθανότητες των P, Q αντίστοιχα.

Όπως προκύπτει και από τη σχέση (2.8) παρατηρούμε ότι η Απόκλιση Jensen-Shannon είναι μια συμμετρική εκδοχή της Απόκλισης Kullback–Leibler. Τέλος, για το πεδίο τιμών της Απόκλισης Jensen -Shannon όταν η βάση στον λογάριθμο είναι δύο, ισχύει ότι:

$$0 \leq JSD(P \parallel Q) \leq 1 \tag{2.11}$$

2.5 Ομοιότητα συνημιτόνου

Η Ομοιότητα Συνημιτόνου [39] είναι μία μετρική ομοιότητας διανυσμάτων. Ο τύπος με τον οποίο υπολογίζεται είναι:

$$cosine\_similatiry = cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} \tag{2.12}$$

Πιο συγκεκριμένα, πρόκειται για μία μετρική που υπολογίζει την ομοιότητα μεταξύ δύο διανυσμάτων από τη γωνία που σχηματίζουν μεταξύ τους. Ορίζεται ως το εσωτερικό γινόμενο των διανυσμάτων διαιρεμένο με το γινόμενο των μήκων τους. Το εύρος των τιμών του, είναι όπως και το εύρος της συνάρητησης συνημιτόνου, δηλαδή από το [-1,1]. Δηλαδή, δύο διανύσματα που έχουν την ίδια κατεύθυνση, η ομοιότητα συνημιτόνου ειναι 1, δύο κάθετα διανύσματα έχουν ομοιότητα συνημιτόνου 0, ενώ δύο εντελώς αντίθετα διανύσματα έχουν ομοιότητα -1. Παρόλα αυτά πολλές φορές ο τρόπος υλοποίησής της μπορεί να περιορίζει το εύρος στα θετικά, δηλαδή στο [0,1]

Κεφάλαιο 3

Περιβάλλον και Εργαλεία

Στο κεφάλαιο αυτό γίνεται μία μικρή αναφορά στα χαρακτηριστικά του συστήματος, που υποστήριξε την υλοποίηση του προγραμματιστικού τμήματος της εργασίας. Επίσης πραγματοποιείται και μία μικρή παρουσίαση των εργαλειών που χρησιμοποιήθηκαν.

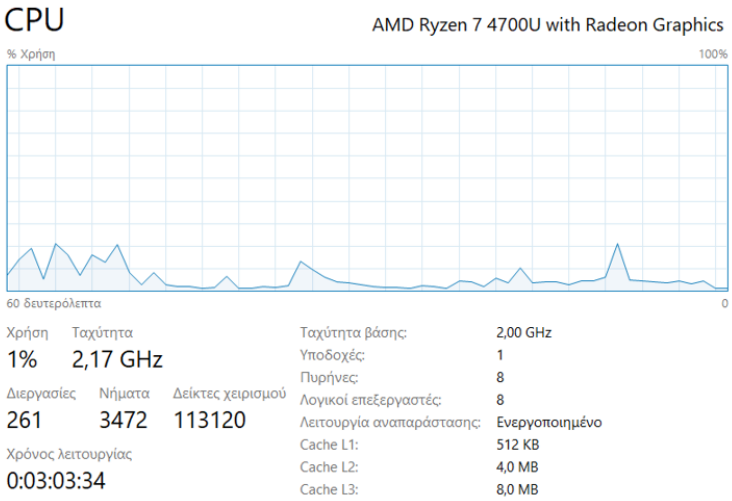
3.1 Πληροφορίες συστήματος

Το σύστημα πάνω στο οποίο εκτελέστηκε το προγραμματιστικό κομμάτι της διπλωματικής εργασίας είχε τα χαρακτηριστικά που αναγράφονται στό Σχήμα 3.1 και 3.2.

Σχήμα 3.1: Προδιαγραφές συσκευής

3.2 Python

Το πρόγραμματιστικό κομμάτι της παρούσας διπλωματικής εργασίας αναπτύχθηκε με τη γλώσσα προγραμματισμού Python και συγκεκριμένα στην έκδοση “3.8.5” [40]. Η Python



Σχήμα 3.2: Πληροφορίες για τη CPU του συστήματος.

είναι μία υψηλού επιπέδου γλώσσα, η χρήση της οποίας τα τελευταία χρόνια αυξάνεται συνεχώς. Αποτελεί μία εκ των βασικών γλωσσών προγραμματισμού στην επιστήμη των δεδομένων και αυτό πιθανότατα να οφείλεται στην ευκολία της χρήσης της, καθώς επίσης και στις απεριόριστες επιλογές που προσφέρονται στον προγραμματιστή εξαιτίας των πολυάριθμων βιβλιοθηκών που διαθέτει.

3.3 Anaconda

Το Anaconda [41] πρόκειται για μια διανομή ανοιχτού κώδικα, των γλωσσών Python και R. Αποτελεί ένα πολύ χρήσιμο εργαλείο για την Επιστήμη των Δεδομένων και τη Μηχανική Μάθηση, καθώς απλοποιεί σημαντικά την διαδικασία διαχείρισης και εγκατάστασης πακέτων.

3.4 Jupyter Notebook

Το Jupyter Notebook, που βρίσκεται προεγκατεστημένο στο περιβάλλον Anaconda είναι μια εφαρμογή ανοιχτού κώδικα, που χρησιμοποιείται για την ανάπτυξη αρχείων κώδικα. Προκειμένου να υλοποιήθει το προγραμματιστικό μέρος της εργασίας, χρησιμοποιήθηκε αυτό το εργαλείο καθώς είναι ευρέως γνωστό για την προτίμηση του στον τομέα της επιστήμης των δεδομένων.

3.5 Tkinter

Τέλος, το γραφικό περιβάλλον του χρήστη υλοποιήθηκε χρησιμοποιώντας τη GUI βιβλιοθήκη *tkinter* (Tk Interface) της Python.

Κεφάλαιο 4

Συλλογή και Προεπεξεργασία Δεδομένων

Σε αυτό το κεφάλαιο θα γίνει μία γνωριμία με τα δεδομένα που χρησιμοποιήθηκαν, προκειμένου να υλοποιηθεί το προγραμματιστικό κομμάτι της εργασίας. Συγκεκριμένα, τα δεδομένα τόσο των βιβλίων, όσο και των χρηστών είναι βασισμένα στην ιστοσελίδα goodreads.

Το goodreads [42] είναι μία ιστοσελίδα κοινωνικής δικτύωσης που ανήκει στην εταιρία της Amazon, και πρόκειται για μία ηλεκτρονική πλατφόρμα, που επιτρέπει στον οποιοδήποτε να αναζητήσει ελεύθερα στην εκτενή βάση δεδομένων από βιβλία, που παρέχει. Ένα μεγάλο κομμάτι αυτής της βάσης έχει δημιουργηθεί από τους ίδιους τους χρήστες, καθώς εάν γίνει δεκτή η αίτηση τους να γίνουν βιβλιοθηκάριοι της Goodreads, τους δίνεται η δυνατότητα να προσθέτουν βιβλία. Επιπλέον, δίνεται η δυνατότητα σε όποιον το επιθυμεί να εγγραφεί και να δημιουργήσει το δικό του προφιλ από καταλόγους ή συλλογές βιβλίων, να βαθμολογήσει, καθώς και να αξιολογήσει ένα βιβλίο. Η κοινότητα του goodreads συνεχώς αυξάνεται και βρίσκεται στην ευχάριστη θέση να απαριθμεί παραπάνω από 125 εκατομμύρια χρήστες και μία βάση από βιβλία που ξεπερνά τα 3,5 δισεκατομμύρια βιβλία.

4.1 Δεδομένα

Σε αυτό το κεφάλαιο θα πραγματοποιηθεί μία σύντομη περιγραφή των αρχικών δεδομένων και πώς αυτά οργανώθηκαν προκειμένου να αποτελέσουν τελικά μία βάση με χρήστες και τις βαθμολογήσεις τους σε βιβλία και μία βάση με βιβλία.

4.1.1 Μετατροπή των αρχικών δεδομένων σε διαχειρίσιμες βάσεις

Οι αρχικές βάσεις δεδομένών ήταν οι εξής:

1. Μια βάση δεδομένων, που περιείχε βιβλία με τα χαρακτηριστικά και το κείμενο περιγραφής τους.

2. Μια βάση δεδομένων, που περιείχε βιβλία με τα χαρακτηριστικά τους, αλλά χωρίς το κείμενο περιγραφής τους. Στα χαρακτηριστικά συμπεριλαμβάνεται και το αναγνωριστικό κάθε βιβλίου, το ISBN, δίνοντας τη δυνατότητα να συλλεχθεί η περίληψη μέσω εφαρμογής της διαδικασίας Web Scraping στην ιστοσελίδα goodreads.

3. Μια βάση με χρήστες, που περιείχε τη βαθμολογία τους για βιβλία

4. Μια βάση με χρήστες, που περιείχε τη βαθμολογία τους για βιβλία, αλλα ταυτόχρονα και πληροφορίες για το αντίστοιχο βιβλίο, μαζί με το αναγνωριστικό ISBN και την περίληψη του.

Η διαδικασία οργάνωσης των τεσσάρων βάσεων σε δύο ξεχωριστές, μία για τα βιβλία και τις πληροφορίες τους και μία με τις βαθμολογίες των χρηστών, υπήρξε αρκετά χρονοβόρα, καθώς τα δεδομένα ήταν πολυάριθμα και διαφορετικά οργανωμένα σε κάθε βάση. Επειδή, η εξήγησή της δεν αποτελεί το αντικείμενο ενδιαφέροντος αυτής της εργασίας, θα γίνει μόνο μια μικρή αναφορά. Επιτεύχθηκε με χρήση κυρίως της βιβλιοθήκης *pandas* και τη χρήση συναρτήσεων, όπως η *merge*, που χρησιμοποιείται γενικότερα για τη συνένωση βάσεων δεδομένων βασιζόμενοι πάνω στις κοινές τιμές μεταξύ στηλών των δύο βάσεων, και η *concat* που συνενώνει βάσεις δεδομένων την μία κάτω από την άλλη ή τη μία δίπλα στην άλλη.

Για τη συνένωση των δύο βάσεων με χρήστες έπρεπε να αποφευχθεί η ταύτιση του ID των χρηστών και αυτό επιτεύχθηκε με χρήση του *LabelEncoder* της βιβλιοθήκης *sklearn*, το οποίο χρησιμοποιείται για το χειρισμό κατηγορικών μεταβλητών. Συγκεκριμένα εφαρμόστηκε αυτή η συνάρτηση στην πρώτη βάση με τους χρήστες και για κάθε διαφορετικό ID χρήστη γινόταν μία ανάθεση ενός συγκεκριμένου αριθμού από το 0 εως το (όσα είναι τα διαφορετικά ID's - 1). Στη δεύτερη βάση εφαρμόστηκε και εδώ *LabelEncoder*, μόνο που μετά από αυτή τη διαδικασία σε κάθε ID προστέθηκε η τιμή max(αριθμός ID στην πρώτη βάση) και έτσι τα ID's των χρηστών δεν μπερδεύτηκαν.

Τελικώς, μαζί με την αφαίρεση των χαρακτηριστικών που δεν χρησιμοποιούνται στην παρούσα εργασία, προέκυψαν οι παρακάτω βάσεις:

- Βάση με βιβλία με τα παρακάτω χαρακτηριστικά:
  - **Name:** Όνομα βιβλίου

  - **ISBN:** Αναγνωριστικό βιβλίου
  - **Authors:** Όνομα συγγραφέα βιβλίου
  - **Description:** Κείμενο περιγραφής βιβλίου/περίληψη βιβλίου

- Βάση με χρήστες με τα παρακάτω χαρακτηριστικά:
  - **ID:** Αναγνωριστικό χρήστη
  - **ISBN:** Αναγνωριστικό βιβλίου που βαθμολόγησε
  - **Name:** Όνομα βιβλίου που βαθμολόγησε
  - **Rating:** Βαθμολογία του χρήστη στο αντίστοιχο βιβλίο

Πρέπει να σημειωθεί πως η βάση με τους χρήστες περιλαμβάνει μόνο βαθμολογίες, που αντιστοιχίζονται σε βιβλία που διαθέτουν περίληψη και βρίσκονται στη βάση με τα βιβλία. Επειδή ακριβώς παρατηρήθηκε μείωση του μεγέθους της βάσης χρηστών με τη συνθήκη αυτή, για να ενισχυθεί η βάση με τους χρήστες, εφαρμόστηκε η διαδικασία Web Scraping. Μέσω αυτής της διαδικασίας αναζητήθηκαν βιβλία (χωρίς περίληψη) τα οποία έχουν αξιολογήσει οι χρήστες και για τα οποία είναι διαθέσιμο το αναγνωριστικό ISBN στην αρχική βάση 2, προκειμένου να συλλεχθεί η περίληψη τους. Η διαδικασία του Web Scraping περιγράφεται στο επόμενο κεφάλαιο.

4.2 Διαδικασία Web Scraping

Το "Web Scraping" [43], είναι μία διαδικασία που επιτρέπει τη συλλογή μεγάλων ποσοτήτων δεδομένων που βρίσκονται διάχυτες στις ιστοσελίδες και την αποθήκευσή τους σε οργανωμένες δομές. Υπάρχουν πολλές τεχνικές επίτευξης αυτής της διαδικασίας όπως διαδικτυακές υπηρεσίες, τα API's ή γράφοντας κώδικα. Στην παρούσα διπλωματική εργασία, η διαδικασία "Web Scraping" υλοποιήθηκε με έναν κώδικα στη γλώσσα προγραμματισμού Python μέσω της βιβλιοθήκης Selenium. Το Selenium είναι ικανό να αυτοματοποιεί διαφορετικά προγράμματα περιήγησης όπως είναι το Chrome και το Firefox, μέσω ενδιάμεσου λογισμικού, που ονομάζεται selenium webdriver.

Στην προκειμένη περίπτωση, με χρήση του Chrome driver, το πρόγραμμα εκτελούσε την εξής λειτουργία:

1. Πήγαινε αυτόματα στη σελίδα **Goodreads Sign In**.

2. Φίλτραρε το αρχείο HTML με τη συνάρτηση προκειμένου να βρει το σημείο που αναγράφει. “Sign in with email” και αφού έβρισκε τα αντικείμενα που αντιστοιχούσαν στην εισαγωγή email και κωδικού, συνδεόταν σε έναν λογαριασμό, που δημιουργήθηκε για την επίτευξη αυτής της διαδικασίας.

3. Φιλτράροντας πάλι το αρχείο HTML της αντίστοιχης σελίδας, έβρισκε το αντικείμενο που αντιστοιχούσε στο σημείο αναζήτησης και εισήγαγε το ISBN του βιβλίου .

4. Στη νέα σελίδα που βρισκόταν, αναζητούσε στον HTML κώδικα το αντικείμενο που εμπεριέχει το κείμενο περιγραφής του βιβλίου, και αφού το σύλλεγε το επέστρεφε.

5. Επέστρεφε στο βήμα 3 έως ότου βρει τις περιλήψεις για όλα τα ISBN, που του τέθηκαν.

Κάποιες συναρτήσεις που χρησιμοποιήθηκαν για την υλοποίηση αυτών των βημάτων ήταν η *find_element()* για εύρεση των αντικειμένων ενδιαφέροντος, η *send_keys()* για την εισαγωγή δεδομένων κειμένου στα αντικείμενα που βρέθηκαν μέσω της *find_element()*, όπως η αναζήτηση. Επίσης, η *click()* για το πάτημα κάποιου κουμπιού και η *move_to_element()* για την έμμεση μετακίνηση του ποντικιού σε ένα συγκεκριμένο στοιχείο της σελίδας.

Ήταν σημαντικό επίσης να αποφευχθούν προβλήματα που θα σταματούσαν την εκτέλεση της διαδικασίας, όπως το πεδίο που περιείχε την περίληψη να ήταν κενό και να μην υπήρχε το αντικείμενο που αναζητούσε στο αρχείο HTML και το να μπορέσει να πατηθεί το αντικείμενο που αντιστοιχούσε στην επιλογή “...more” όταν η περίληψη ήταν μεγάλη, προκειμένου να παρθεί ολόκληρη.

Πάρθηκαν γύρω στα 15.000 καινούρια βιβλία, τα οποία προστέθηκαν στη βάση με τα βιβλία και αντίστοιχα κρατήθηκαν και οι βαθμολογίες των χρηστών που αντιστοιχίζονταν σε αυτά τα βιβλία.

4.3 Διαχείριση ελλιπών τιμών και απαλοιφή διπλοτύπων

Σε αυτό το κεφάλαιο θα παρουσιαστούν δύο στάδια προεπεξεργασίας που εφαρμόστηκαν και στις δύο βάσεις, δηλαδή στη βάση με τα βιβλία και στη βάση με τις βαθμολογίες των χρηστών πριν την προεπεξεργασία των δεδομένων της κάθε μίας βάσης ξεχωριστά.

Ένα μεγάλο πρόβλημα, που συναντάται στην επιστήμη των δεδομένων, είναι οι ελλιπείς

τιμές στα σύνολα δεδομένων, ή αλλιώς η ύπαρξη NaN τιμών. Υπάρχουν ποικίλοι τρόποι επίλυσης αυτού του προβλήματος, όπως:

- Διαγραφή της στήλης/χαρακτηριστικού, αν οι περισσότερες τιμές της στήλης είναι NaN
- Αντικατάσατση της ελλιπής τιμής με κάποια συγκεκριμένη τιμή όπως μέση τιμή της στήλης, πιο συχνά εμφανιζόμενη τιμή στη στήλη κ.λπ.
- Αφαίρεση ολόκληρης της γραμμής που περιέχει NaN τιμή

Η λύση που εφαρμόστηκε στις βάσεις, για την εξάλειψη αυτού του προβλήματος είναι η αφαίρεση των εγγραφών που σε οποιαδήποτε από τις στήλες/χαρακτηριστικά εμφάνιζαν NaN τιμές. Η διαδικασία αυτή επιτεύχθηκε με τη χρήση της εντολής *notna()*, που ανήκει στη βιβλιοθήκη *pandas*.

Στη συνέχεια, το επόμενο ζήτημα είναι η εύρεση των διπλότυπων εγγραφών στη βάση δεδομένων και η διαχείριση τους. Στην περίπτωση αυτή, ο τρόπος επίλυσης του προβλήματος είναι περιορισμένος καθώς μπορεί είτε να κρατήσει την πρώτη εγγραφή, είτε την τελευταία, είτε καμία από τις δύο. Με χρήση της συνάρτησης *drop_duplicates(keep='first')* και την εφαρμογή της σε όλες τις στήλες και στις δύο βάσεις, διατηρούσε μόνο την πρώτη εγγραφή για κάθε διπλότυπη εμφάνιση. Και αυτή η συνάρτηση, είναι συνάρτηση της βιβλιοθήκης *pandas*.

4.4 Προεπεξεργασία δεδομένων βιβλίων

Η προεπεξεργασία της βάσης με τα βιβλία στην παρούσα διπλωματική εργασία, διακρίνεται σε δύο φάσεις. Η μία φάση είναι η βασική προεπεξεργασία που εφαρμόστηκε στα δεδομένα και αποτελείται από τα στάδια που εφαρμόζονται γενικότερα κατά την επεξεργασία δεδομένων κειμένου. Αυτά τα προεπεξεργασμένα δεδομένα δοκιμάστηκαν στο μοντέλο LDA και διαπιστώθηκαν αλλαγές που μπορούν να συμβούν στη φάση της προεπεξεργασίας δεδομένων προκειμένου να προκύψουν καλύτερα αποτελέσματα. Αυτές οι αλλαγές, που επινοήθηκαν για την καλύτερη λειτουργία του μοντέλου, θα περιγραφούν στη δεύτερη φάση προεπεξεργασίας των δεδομένων.

4.4.1 Πρώτη φάση προεπεξεργασίας δεδομένων βιβλίων

Η πρώτη φάση προεπεξεργασίας δεδομένων βιβλίου αποτελείται από τα στάδια που θα εξηγηθούν στις παρακάτω υποενότητες και που θα εφαρμοστούν για αρχή στα δεδομένα που υπάρχουν στη στήλη Description της βάσης. Πρόκειται για βασικές τεχνικές επεξεργασίας δεδομένων κειμένου [44].

Καθαρισμός δεδομένων κειμένου

Ο καθαρισμός δεδομένων κειμένου είναι μία απαραίτητη διαδικασία, προκειμένου να αφαιρεθεί όλος ο θόρυβος που υπάρχει στα δεδομένα. Συγκεκριμένα, η παρούσα διπλωματική, περιλαμβάνει τις εξής κατηγορίες καθαρισμού δεδομένων:

- αφαίρεση χαρακτήρων αλλαγής γραμμής (π.χ.\n), html tags (π.χ. <h1>) , συνδέσμου (link), σημείων στίξης, αριθμών και κενών χαρακτήρων,
- μετατροπή όλων των γραμμάτων σε πεζά γράμματα,
- διόρθωση λέξεων που μπορεί να επαναλαμβάνουν κάποιο χαρακτήρα από λάθος,
- Επέκταση των λέξεων συστολής (π.χ. isn't σε is not)

Όλα τα παραπάνω υλοποιήθηκαν κατά βάση με χρήση της βιβλιοθήκης *re* της Python. Μία συνάρτηση που χρησιμοποιήθηκε ήταν η *sub(pattern, repl, text)*, που ουσιαστικά αναζητά στο text, το pattern που του όρισε ο χρήστης και το αντικαθιστά με το repl, ενώ επίσης χρησιμοποιήθηκε και η *compile()* της ίδιας βιβλιοθήκης, που δέχεται ως είσοδο ένα μοτίβο σε μορφή συμβολοσειράς και το μετατρέπει σε re.Pattern μορφή. Τέλος, χρησιμοποιήθηκε και η συνάρτηση *replace*, που απλά πραγματοποιεί αντικατάσταση μιας συμβολοσειράς με μία άλλη.

Αναγνώριση γλώσσας

Μετά τον καθαρισμό των δεδομένων ακολούθησε η διαδικασία αναγνώρισης γλώσσας των περιλήψεων των βιβλίων. Αυτό συνέβη, επειδή παρατηρήρθηκε οτι υπάρχουν βιβλία και σε άλλες γλώσσες και όχι μόνο στην αγγλική. Η αναγνώριση γλώσσας έγινε με χρήστη της συνάρτησης *detect(text)* της *langdetect* βιβλιοθήκης, η οποία υποστηρίζει την αναγνώριση 55 γλωσσών.

Εφαρμόστηκε στις περιλήψεις των βιβλίων και από τις συνολικές γλώσσες που εντοπίστηκαν, κυρίαρχη ήταν η αγγλική γλώσσα, που ήταν και η μόνη που διατηρήθηκε.

Tokenization

Η προεπεξεργασία των δεδομένων κειμένου συνεχίζεται με τη διαδικασία του Tokenization. Πιο λεπτομερώς, το tokenization είναι η διάσπαση ενός κειμένου σε μικρότερα κομμάτια προτάσεις ή λέξεις, τα οποία ονομάζονται tokens. Σε αυτήν την περίπτωση, το καθαρισμένο πλέον κείμενο, διασπάστηκε σε λέξεις. Για την υλοποίηση αυτού του βήματος χρησιμοποιήθηκε η συνάρτηση *split()*.

Lemmatization

Ακολουθεί το Lemmatization, που αποτελεί μία διαδικασία αναγωγής μίας λέξης στη ρίζα της [45]. Με ένα απλό παράδειγμα οι λέξεις "played" και "playing" να αντιστοιχίζονται στη λέξη "play". Για την αντιστοίχιση της λέξης με τη ρίζα της, γίνεται ένας προσδιορίσμος του μέρους του λόγου της λέξης στην πρόταση.

Στην προκειμένη περίπτωση, η διαδικασία Lemmatization πραγματοποιήθηκε με χρήση του *WordNetLemmatizer*. Η λημματοποίηση κάθε λέξης γινόταν με βάση το αρχικό γράμμα του μέρους του λόγου που την κατεταζε. Δηλαδή αν ξεκινούσε με "J", τότε το έκανε λημματοποίηση ως επίθετο, αν ξεκινούσε με "V" ως ρήμα, αν ξεκινούσε με "N" ως ουσιαστικό και αν ξεκινούσε με "R" ως επίρρημα. Σε κάθε άλλη περίπτωση το λημματοποιούσε ως ουσιαστικό.

Δημιουργία συνδυασμών λέξεων

Σε αυτό το βήμα προεπεξεργασίας, γίνεται η προσπάθεια εντοπισμού φράσεων ή συνδυασμό λέξεων που υπάρχουν μέσα στα κείμενα, όπως ο εντοπισμός της φράσης "New York" προκειμένου να μη λαμβάνονται ως δύο ξεχωριστές λέξεις, αφου η μία προσδιορίζει την άλλη. Στον τομέα της Επεξεργασίας της Φυσικής Γλώσσας μέσω υπολογιστών αυτοί οι συνδυασμοί λέγονται collocations ή n-grams, όπου το n υποδηλώνει τον αριθμό λέξεων, που θα αποτελείται η φράση.

Στην παρούσα εργασία θα πραγματοποιηθεί προσπάθεια εύρεσης φράσεων με n=2 και n=3, που ονομάζονται Bigrams και Trigrams αντίστοιχα. Η διαδικασία αυτή υλοποιήθηκε μέσω της βιβλιοθήκης *nltk* της python και των συναρτήσεων *BigramCollocationFinder()*,

που ουσιαστικά βρίσκει bigrams από το κείμενο, καθώς επίσης και με τη συνάρτηση *BigramAssocMeasures*, που προσφέρει μια συλλογή από μέτρα συσχέτισης bigram, δηλαδή τα βάρη που θα αποδώσει σε κάθε bigram. Υπάρχουν οι αντίστοιχες και για τα trigrams. Για κάθε συνδυασμό bigram και trigram θα χρησιμοποιηθούν τρείς διαφορετικές μετρικές συσχέτισης: μετρική συχνότητας (frequency counting), μετρική PMI (Pointwise Mutual Information) και τη μετρική του στατιστικού ελέγχου υποθέσεων chi-square (hypothesis testing chi-square).

Πριν αναλυθούν αυτές οι μετρικές, είναι σημαντικό να τονιστεί η ύπαρξη ενός προβλήματος. Συγκεκριμένα, μέσω αυτής της διαδικασίας μπορεί να υπάρξει συνδυασμός λέξεων που δεν παρέχει πληροφορία, όπως “there is a”. Για την αντιμετώπιση αυτού του ζητήματος δημιουργήθηκε ένα φίλτρο για τα bigrmas και ένα για τα trigrams, που πραγματοποιεί τα εξής:

- **Για τα bigrams:** Ελέγχει σε τι μέρος του λόγου ανήκουν οι λέξεις από τις οποίες αποτελείται το Bigram και κρατάει μόνο όσα έχουν τους συνδυασμούς (Ουσιαστικό Ουσιαστικό) ή (Επίθετο Ουσιαστικό). Τέλος, ελέγχει να μην περιέχονται λέξεις που ανήκουν στα stopwords, όπως “the”, “at” και να μην έχουν μέγεθος μικρότερο ή ίσο του 2, εκτος αν η δεύτερη λέξη είναι “fi”

- **Για τα trigrams:** Ελέγχει σε τι μέρος του λόγου ανήκουν οι λέξεις από τις οποίες αποτελείται το Trigram και κρατάει μόνο όσα αναγνωρίζονται η πρώτη ή η τρίτη λέξη τους ως Ουσιαστικό ή Επίθετο. Τέλος, ελέγχει να μην περιέχονται στα stopwords η πρώτη και η τελευταία λέξη και για τις ίδιες λέξεις να μην έχουν μέγεθος μικρότερο ή ίσο του 2.

Για την αναγνώριση του μέρους του λόγου χρησιμοποιήθηκε η συνάρτηση *pos_tag* της βιβλιοθήκς *nltk*, ενώ για τα stopwords χρησιμοποιήθηκε μία λίστα που παρέχει πάλι η ίδια βιβλιοθήκη (βλπ. 4.4.1).

Σχετικά με τις μετρικές συσχέτισης [46]:

1. Frequency Counting

Ουσιαστικά γίνεται μία καταμέτρηση των φορών που εμφανίζεται ένα bigram/trigram στα κείμενα και με αυτόν τον τρόπο ορίζεται πόσο σημαντικό είναι.

2. Pointwise Mutual Information (PMI)

Το PMI υπολογίζει την πιθανότητα να εμφανιστεί μία λέξη, δεδομένου ότι κάποια άλλη λέξη έχει εμφανιστεί και το μέτρο συσχέτισης για τα bigrams υπολογιζόταν μέσω του τύπου:

$$PMI(w_1, w_2) = log_2 \frac{P(w_1, w_2)}{P(w_1)P(w_2)} \tag{4.1}$$

ενώ για τα trigrams μέσω:

$$PMI(w_1, w_2, w_3) = log_2 \frac{P(w_1, w_2, w_3)}{P(w_1)P(w_2)P(w_3)} \tag{4.2}$$

3. Hypothesis Testing Chi-square

Τέλος, στο Hypothesis Testing Chi-square για απλότητα θα γίνει αναφορά μόνο στα bigrams και συγκεκριμένα για το παράδειγμα όπου το bigram είναι το "New York". Αρχικά γίνεται υπολογισμός του πίνακα 4.1, όπου Count(w_1, w_2) σημάινει καταμέτρηση των φορών που εμφανίζονται μαζί οι λέξεις w_1, w_2 και X,Y τυχαίες λέξεις εκτός του "New" και του "York".

Πίνακας 4.1: Καταμέτρηση εμφανίσεων για τον υπολογισμό του chi-square για το bigram

	$\mathbf{w_1 = New}$	$\mathbf{w_1 \neq New}$
$\mathbf{w_2 = York}$	Count(New York)	Count(X York)
$\mathbf{w_2 \neq York}$	Count(New X),	Count(X,Y)

Για τον υπολογισμό του chi-square εφαρμόζεται ο παρακάτω τύπος:

$$\chi^2 = \sum_{i,j} \frac{(Observed_{i,j} - Expected_{i,j})^2}{Expected_{i,j}}, \tag{4.3}$$

όπου $Observed_{i,j}$ ο πίνακας 4.1 και i,j είναι η στήλη και η γραμμή αντίστοιχα. Επίσης το $Expected_{i,j}$ είναι η προσδοκώμενη τιμή αν τα δεδομένα είναι ανεξάρτητα και για αυτό υπολογίζεται από το παρακάτω γινόμενο.

$$Expected_{w_1,w_2} = \frac{Count(w_1)}{N} \frac{Count(w_2)}{N} N, \tag{4.4}$$

όπου N είναι ο συνολικός αριθμός των λέξεων σε όλα τα κείμενα. Ουσιαστικά γίνεται μία σύγκριση των παρατηρούμενων συχνοτήτων με τις συχνότητες όταν οι λέξεις εμφανίζονται ανεξάρτητα.

Εφαρμόζοντας σε κάθε περίπτωση και το φίλτρο που αναφέρθηκε ως προς το από τι λέξεις θα αποτελούνται τα bigrams/trigrams και κρατώντας σε κάθε περίπτωση αυτά με την υψηλότερη συσχέτιση, συνενώθηκαν όλα τα αποτελέσματα που προέκυψαν. Αυτό το βήμα προεπεξεργασίας ολοκληρώθηκε με την αντικατάσταση των bigrams/trigrams μέσα στα κείμενα, προκειμένου να συνυπολογίζονται ως μία λεξη.

Αφαίρεση stopwords και ασήμαντων λέξεων

Τελευταίό βήμα και από τα πιο σημαντικά είναι η αφαίρεση των stopwords και των ασήμαντων λέξεων. Ως stopwords ορίζονται οι πιο χρησιμοποιούμενες λέξεις σε ένα κείμενο, μερικά παραδείγματα των οποίων είναι οι “the”, “is”, “at”, “there” και “a”. Αυτές οι λέξεις δεν κατέχουν κάποια σημασιολογική σημασία σε ένα σύνολο λέξεων, επομένως η αφαίρεση τους θα ξεκαθαρίσει το τοπίο, και θα φέρει στην επιφάνεια λέξεις που έχουν υψηλή πληροφορία. Η διαδικασία της αφαίρεσης αυτών των λέξεων πραγματοποιήθηκε με χρήση της βιβλιοθήκης *nltk* που έχει ενσωματωμένη μία λίστα αποτελούμενη από αγγλικά stopwords. Στη λίστα με τα stopwords προστέθηκαν και κάποιες ακόμα λέξεις όπως “book”, “writer”, “author”, που υπήρχε η υποψία της ύπαρξής τους, και επειδή και αυτές είναι λέξεις που δεν έχουν κάποια ιδιαίτερη σημασία αφαιρέθηκαν.

Τέλος, οι ασήμαντες λέξεις είναι οι λέξεις που το μήκος τους είναι μικρότερο και ίσο του 2, επομένως διατηρήθηκαν μόνο λέξεις που είχαν μέγεθος μεγαλύτερο του 2, όπως επίσης αφαιρέθηκαν και οι λέξεις που είναι γραπτοί αριθμοί, όπως “one”, αφού αναγνωρίστηκαν με χρήστη της συνάρτησης *isdigit()*.

4.4.2 Δεύτερη φάση προεπεξεργασίας δεδομένων βιβλίων

Αλλαγές στα δεδομένα μετά από δοκιμές με το μοντέλο LDA

Αφού εφαρμόστηκε η βασική προεπεξεργασία στα δεδομένα, εκτελέσθηκαν κάποιες δοκιμές του μοντέλου πάνω στα συγκεκριμένα δεδομένα. Από αυτές τις δοκιμές, προέκυψαν

κάποια αποτελέσματα, τα οποία επιδέχονταν βελτίωση και συγκεκριμένα αυτό μπορούσε να επιτευχθεί με την αναπροσαρμογή των δεδομένων εισόδου. Συγκεκριμένα, οι επιθυμητές αλλαγές που θα μπορούσαν να εφαρμοστούν είναι οι εξής:

- Το dataset με τα βιβλία να περιοριστεί σε βιβλία που περιλαμβάνουν περιλήψεις με αριθμό λέξεων, που να ανήκει στο διάστημα [60,400]. Οι δοκιμές, που οδήγησαν σε αυτήν την απόφαση, εφαρμόστηκαν σε περιλήψεις που περιείχαν 15 λέξεις, πριν ακόμα από το στάδιο της προεπεξεργασίας. Συνεπώς, τόσο μικρά κείμενα δε συνεισφέρουν χρήσιμη πληροφορία στο μοντέλο.

- Από τις λέξεις κάθε περίληψης να παραμένουν μόνο οι λέξεις που αναγνωριζονται ως ουσιαστικά. Συγκεκριμένα, με μία καταμέτρηση των μερών του λόγου που περιλαμβάνουν συνολικά όλα τα κείμενα, παρατηρήθηκε ότι κυριαρχούσαν τα ουσιαστικά (βλπ. 4.1). Αυτή η διαπίστωση σε συνδυασμό με τις δοκιμές στο μοντέλο, οδήγησαν στο συμπέρασμα ότι τα υπόλοιπα μέρη του λόγου, όπως τα ρήματα και τα επίθετα, αποπροσανατολίζουν το μοντέλο και δεν προσφέρουν ουσιώδη πληροφορία στο κείμενο.

- Να αφαιρεθούν τα ξένα ονόματα και τα ονόματα των συγγραφέων από το κείμενο περιγραφής κάθε βιβλίου, καθώς μπέρδευε το μοντέλο.

- Τέλος, η είσοδος στο μοντέλο να μην είναι μόνο τα προεπεξεργασμένα κείμενα περιγραφής των βιβλίων, αλλά να ενταχθεί μέσα σε αυτά και ο τίτλος των βιβλίων για να προσφέρει παραπάνω πληροφορία.

Στάδια τελικής προεπεξεργασίας

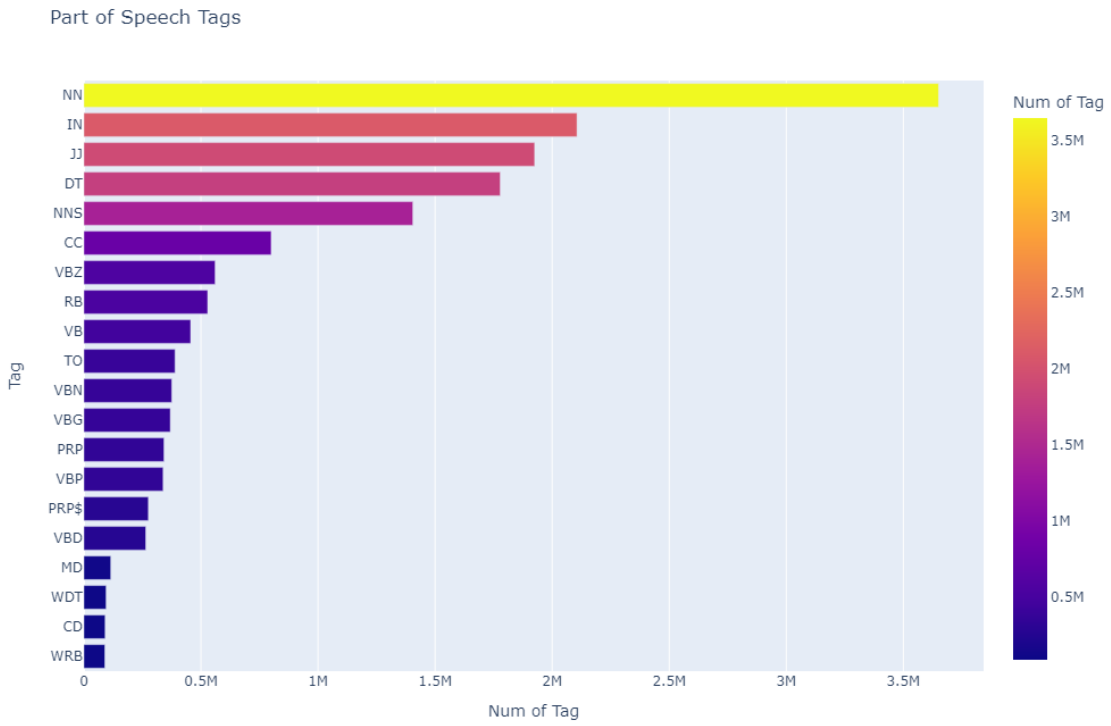
Παρακάτω αναφέρονται αριθμημένα όλα τα βήματα επεξεργασίας των δεδομένων με την τελική σειρά που συνέβησαν αφού πάρθηκαν οι παραπάνω αποφάσεις για τις αλλαγές.

1. Καθαρισμός δεδομένων κειμένου στις στήλες Name και Description (βλπ. 4.4.1).

2. Αναγνώριση γλώσσας στο Description και στο Name και διατήρηση μόνο των βιβλίων που έχουν αγγλική γλώσσα στην περιγραφή και το όνομα (βλπ. 4.4.1).

3. Αφαίρεση ονόματος συγγραφέα ή υποτμήμα του ονόματος, από το κείμενο περιγραφής του αντίστοιχου βιβλίου.

4. Tokenization και στη στήλη Name και στη στήλη Description (βλπ. 4.4.1).



Σχήμα 4.1: Τα μέρη του λόγου που εμφανίζονται στα κείμενα και η καταμέτρηση τους.

5. Καταμέτρηση των λέξεων στη στήλη Description και διατήρηση μόνο των βιβλίων που διαθέτουν περιλήψεις με αριθμό λέξεων από 60 εως 400 λέξεις.

6. Lemmatization των στηλών Name και Description (βλπ. 4.4.1).

7. Συνένωση της στήλης Description και Name σε μία, πάνω στην οποία θα υλοποιηθούν όλα τα υπόλοιπα βήματα.

8. Δημιουργία Bigrams και Trigrams (βλπ. 4.4.1).

9. Αφαίρεση stopwords και ασήμαντων λέξεων (βλπ. 4.4.1), καθώς επίσης και των αγγλικών ονομάτων χρησιμοποιώντας μία βάση που παρέχει 18.000 αγγλικά ονόματα.

10. Διατήρηση μόνο των λέξεων που αναγνωρίζονται ως ουσιαστικά μέσω της συνάρτησης *pos_tag* της βιβλιοθήκης *nltk*, εξαιρώντας από τη διαδικασία τα bigrams και trigrams.

	ISBN	Name	Authors	Description	TokenLemNounsBigramTrigramRemStopWords
0	097669400X	agile web development with rails: a pragmatic ...	dave thomas	rails is a full stack open source web framewor...	['rail', 'stack', 'source', 'web', 'framework'...
1	0517545357	the restaurant at the end of the universe (hit...	douglas adams	just when you thought it was safe to go back t...	['think', 'end', 'time', 'infinity', 'place', ...
2	046504512X	the clock of the long now: time and responsibi...	stewart brand	using the designing and building of the clock ...	['building', 'clock', 'framework', 'time', 'ge...
3	0439023483	the hunger games (the hunger games, #1)	suzanne collins	winning means fame and fortune losing means ce...	['mean', 'fame', 'death', 'hunger', 'game', 'p...
4	0747409382	the authoritative calvin and hobbes: a calvin ...	bill watterson	the authoritative calvin and hobbes is a large...	['hobbes', 'treasury', 'cartoon', 'yukon', 'we...
...	...	...	...	...	...
106427	1841500283	development through technology transfer: creat...	mohammed saad	the first study in technology transfer to use ...	['study', 'technology', 'transfer', 'company',...
106428	052135644X	the purpose of the biblical genealogies: with ...	marshall d johnson	with special reference to the setting of the g...	['reference', 'material', 'occur', 'literature...
106429	0415165024	popper's open society after fifty years	ian c jarvie	popper s open society after fifty years presen...	['society', 'coherent', 'survey', 'reception',...
106430	1855675277	democratization in central and eastern europe	ivan vejvoda	between december and june association or europ...	['association', 'europe', 'agreement', 'conclu...
106431	044450432X	journal of chromatography library, volume 62: ...	zdenek deyl	this book discusses the evolution and uses for...	['evolution', 'electrochromatography', 'dimens...

106432 rows × 5 columns

Σχήμα 4.2: Βάση δεδομένων βιβλίων

4.5 Προεπεξεργασία δεδομένων χρηστών

Η προεπεξεργασία των χρηστών αποτέλεσε μία πιο απλή διαδικασία. Συγκεκριμένα, υλοποιήθηκαν τα παρακάτω:

- Διατήρηση μόνο των βαθμολογήσεων σε βιβλία που υπάρχουν στην τελική βάση των βιβλίων.
- Επειδή η βάση των χρηστών προήλθε από δύο ξεχωριστές βάσεις, στη στήλη Rating παρατηρήθηκε ότι οι βαθμολογίες των χρήστων στο πρώτο dataset ήταν αριθμητικές από το 1 εως το 10, ενώ στο δεύτερο dataset ήταν λεκτικές. Επομένως πραγματοποιήθηκε μία αντιστοίχιση σε βαθμολογίες από το 1 εως το 5 προκειμένου να προκύψει ομοιομορφία.
- Αφαίρεση των χρηστών που είχαν βαθμολογήσει μόνο ένα βιβλίο.
- Διατήρηση μόνο των θετικών βαθμολογιών και στην προκειμένη περίπτωση βαθμολογίες που ανήκουν στο κλειστό διάστημα [3,5].

Τελικά, η τελική βάση δεδομένων, περιέχει 147.773 βαθμολογήσεις από 10.181 χρήστες, πάνω σε 31.048 βιβλία.

	Name	Rating	ID	ISBN
0	agile web development with rails: a pragmatic ...	5.0	49979	097669400X
1	agile web development with rails: a pragmatic ...	4.0	49992	097669400X
2	agile web development with rails: a pragmatic ...	5.0	50001	097669400X
3	agile web development with rails: a pragmatic ...	4.0	50034	097669400X
4	agile web development with rails: a pragmatic ...	4.0	50050	097669400X
...	...	...	...	...
147768	fault lines: stories of divorce	3.0	49882	0425188531
147769	is it too late to run away and join the circus...	5.0	49891	0028620585
147770	mistress of falcon court	3.5	49900	0821726595
147771	inner hunger: a young woman's struggle through...	5.0	49944	0393045900
147772	all elevations unknown: an adventure in the he...	3.0	49976	0767907752

147773 rows × 4 columns

Σχήμα 4.3: Βάση δεδομένων χρηστών

Κεφάλαιο 5

Μέθοδοι Αξιολόγησης Μοντέλων LDA

Πριν από το σχεδιασμό και την ανάπτυξη του μοντέλου LDA θα επεξηγηθούν οι μετρικές που δημιουργήθηκαν για την αξιολόγηση του μοντέλου, καθώς επίσης θα πραγματοποιηθεί και μία εισαγωγή στο θεωρητικό τους υπόβραθρο.

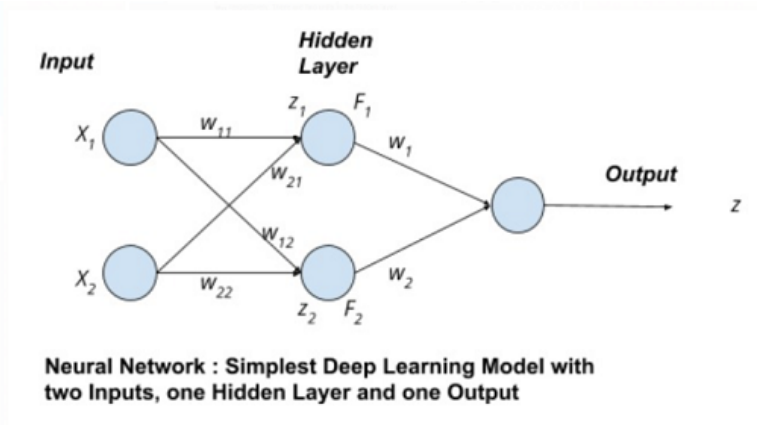
5.1 Θεωρητικό Υπόβαθρο

Σε αυτήν την ενότητα θα επεξηγηθούν κάποιες έννοιες, προκειμένου να γίνει κατανοητή η παρουσίαση των μεθόδων αξιολόγησης.

5.1.1 Τεχνητά νευρωνικά δίκτυα (ANN)

Σε αυτήν την ενότητα θα γίνει μία μικρή εισαγωγή στα τεχνητά νευρωνικά δίκτυα (ANN), προκειμένου να γίνει πιο κατανοητή η επεξήγηση του μοντέλου Word2vec. Τα ANN [47] αποτελούνται από τρία επίπεδα, το Input Layer, το Hidden Layer και το Output Layer. Το Hidden Layer μπορεί να αποτελείται από πολλά ενδιάμεσα Layers, αλλά στην προκειμένη περίπτωση θα περιοριστεί η επεξήγηση στην ύπαρξη μόνο ενός όπως φαίνεται στο Σχήμα 5.1.

Ουσιαστικά, ο βασικός στόχος ενός νευρωνικού δικτύου είναι η εκπαίδευση του, δηλαδή ο κατάλληλος προσδιορισμός των παραμέτρων του, προκειμένου για μία συγκεκριμένη είσοδο να παράγει μία επιθυμητή έξοδο. Οι παράμετροι αυτοί είναι τα βάρη w, το bias b και ο ρυθμός με τον οποίον αλλάζει το βάρος και το bias, α, που ονομάζεται και learning rate. Η διαδικασία με την οποία λειτουργεί θα εξηγηθεί με μία μαθηματική αναπαράσταση, όπου z,a,x,b είναι διανύσματα, W είναι πίνακας:



Σχήμα 5.1: Νευρωνικό Δίκτυο με ένα Hidden Layer [7]

► **Για την παραγωγή της εξόδου (Forward Propagation):**

$$\begin{aligned} z &= xW^{(1)} + b^{(1)} \\ a &= F(z) \\ output &= aW^{(2)} + b^{(2)} \\ error &= C(target, output) \end{aligned} \tag{5.1}$$

Ο $W^{(1)}$ είναι ο πίνακας βαρών που συνδέει το Input Layer με το Hidden Layer και $W^{(2)}$ ο πίνακας που συνδέει το Hidden Layer με το Output Layer. F είναι η συνάρτηση ενεργοποιήσης . Τα $b^{(1)}, b^{(2)}$ αφορούν το Hidden και Output Layer αντίστοιχα. Τέλος, C είναι μία συνάρτηση κόστους, που ποσοτικοποιεί τη διαφορά μεταξύ target και output.

► **Για την διόρθωση των βαρών και του bias (Backward Propagation):** Για ένα βάρος w και ένα bias b η διόρθωση γίνεται ως εξής:

$$\begin{aligned} w &= w - \alpha \frac{\partial C}{\partial w} \\ b &= b - \alpha \frac{\partial C}{\partial b} \end{aligned} \tag{5.2}$$

5.1.2 Μοντέλο Word2vec

Με τον όρο Ενσωμάτωση Λέξης (Word Embedding), νοείται η αναπαράσταση της λέξης ενός κειμένου με διάνυσματα, κατά τέτοιον τρόπο που οι όμοιες σημασιολογικά λέξεις αντιστοιχίζονται σε όμοια αριθμητικά διανύσματα [48]. Για παράδειγμα η λέξη “Instagram” θα αναπαριστάται με ένα όμοιο διάνυσμα με το διάνυσμα της λέξης “Facebook”. Ουσιαστικά, τα διανύσματα αυτά προσπαθούν να αφομοιώσουν οποιοδήποτε χαρακτηριστικό έχει αυτή

η λέξη που αναπαριστούν, και μπορεί να αντληθεί μέσα από το κείμενο.

Το Word2vec [48, 49, 8] είναι ένα από τα μοντέλα που χρησιμοποιούνται για την κατασκευή ενσωματωμένων λέξεων. Πιο συγκεκριμένα, είναι ένα μοντέλο νευρωνικού δικτύου ενός Hidden Layer, που προσπαθεί να μάθει τις συσχετίσεις μεταξύ των λέξεων, εισάγοντας στο διάνυσμα αναπαράστασης της λέξης το χαρακτηριστικό του πλαισίου λέξεων που περικλείουν τη λέξη στο κείμενο.

Αρχικά, τόσο η είσοδος όσο και η επιθυμητή έξοδος του νευρωνικού δικτύου είναι one-hot κωδικοποιημένα διανύσματα, τα οποία είναι διανύσματα με μέγεθος V, όσο και το μέγεθος του συνολικού λεξιλογίου, που αποτελούνται από μηδενικές τιμές, εκτός της θέσης που αντιστοιχίζεται στη λέξη ενδιαφέροντος που έχει την τιμή 1. Δηλαδή, αν το λεξιλόγιο είναι το [“Data”, “science”, “uses”, “machine”, “learning”, “algorithms”], τα one-hot κωδικοποιημένα διανύσματα για τις λέξεις θα είναι τα: Data:[1,0,0,0,0,0], science:[0,1,0,0,0,0], uses:[0,0,1,0,0,0] κ.λπ.

Για την υλοποίησή του, το Word2vec μπορεί να χρησιμοποιήσει δύο διαφορετικές τεχνικές, το Continuous Bag of Words (CBOW) και το Continuous Skip-gram. Πιο λεπτομερώς, για τις δύο τεχνικές:

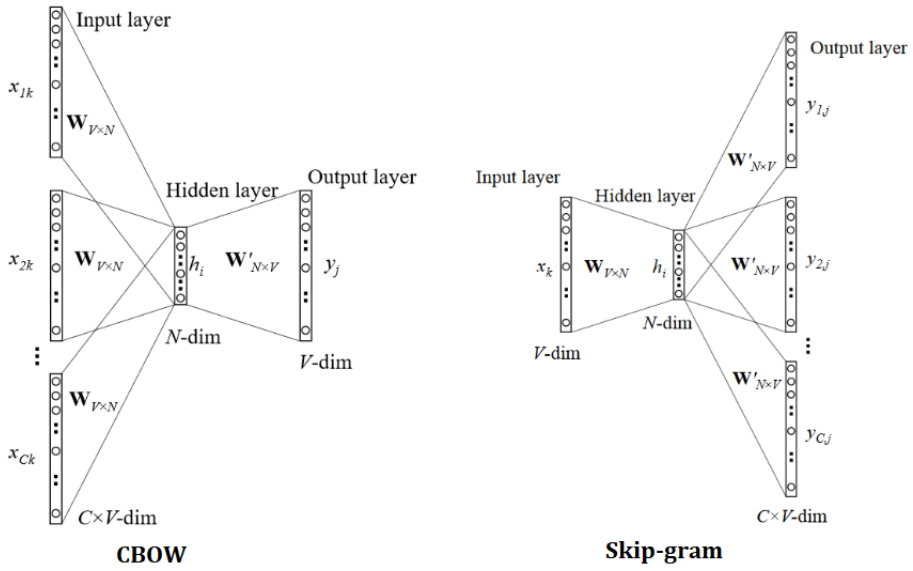
- **CBOW:**

  Η τεχνική CBOW χρησιμοποιεί ως είσοδο το σύνολο των λέξεων περιβάλλοντος X της λέξης Y, που θέλει να προβλέψει (βλπ. Σχήμα 5.2). Επομένως, δέχεται ως είσοδο C one-hot κωδικοποιημένα διανύσματα λέξεων, με V διάσταση το καθένα, όση και η διάσταση του λεξιλογίου. Κάθε είσοδος συνδέεται με τους νευρώνες του Hidden Layer μέσω βαρών, που αναπαρίστανται με τη μορφή πίνακα W_{VxN}, με N τον αριθμό των νευρώνων στο Hidden Layer. Στο Hidden Layer στο άθροισμα που δημιουργείται δεν εφαρμόζεται καμία συνάρτηση ενεργοποίησης. Στη συνέχεια, οι νευρώνες του Hidden Layer συνδέονται με το Output Layer μέσω βαρών, και σε μορφή πίνακα W'_{NxV}. Η έξοδος που παράγεται στο Output Layer εισάγεται σε μία συνάρτηση ενεργοποίησης softmax, που ουσιαστικά μετατρέπει ένα διάνυσμα K πραγματικών αριθμών σε κατανομή πιθανότητας K πιθανών αποτελεσμάτων, και στη συνέχεια συγκρίνεται με την one-hot επιθυμητή έξοδο της λέξης ενδιαφέροντος μέσω της συνάρτησης κόστους. Ο αλγόριθμος ή θα τερματίσει αν έχει έρθει σε σύγκλιση ή θα προχωρήσει στην αναπροσαρμογή των βαρών και στη συνέχεια στην επανάληψη της διαδικασίας, προκειμένου να προσεγγίσει καλύτερα την επιθυμητή έξοδο. Τελικώς, η Word Embedding αναπα-

ράσταση της επιθυμητής λέξης είναι η στήλη του πίνακα W'_{NxV} που αντιστοιχίζεται στη λέξη ενδιαφέροντος.

► **Continuous Skip-gram:**

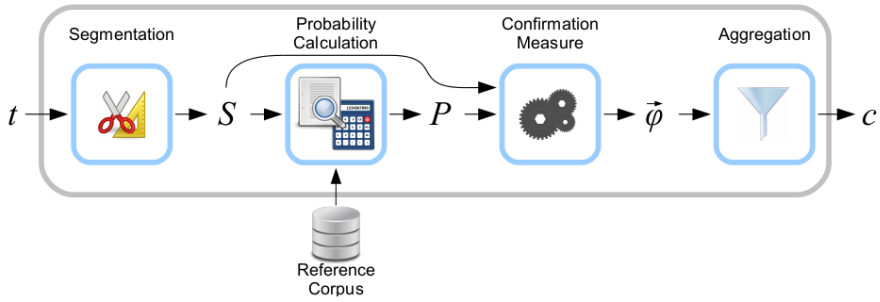
Η τεχνική Skip-gram ουσιαστικά επιτελεί την αντίστροφη διαδικασία από τη CBOW. Δηλαδή, ως είσοδο δέχεται μία λέξη και το νευρωνικό δίκτυο προσπαθεί να εκπαιδευτεί ώστε να βρίσκει τις λέξεις, που περιβάλλουν την λέξη εισόδου (βλπ. Σχήμα 5.2). Τελικώς, η Word Embedding αναπαράσταση της επιθυμητής λέξης είναι η γραμμή του πίνακα W_{VxN}, που αντιστοιχίζεται στη λέξη ενδιαφέροντος.



Σχήμα 5.2: CBOW (αριστερά) και Skip-gram (δεξιά) [8]

5.1.3 Συνάφεια θέματος (Topic Coherence)

Μία άμεσα συνυφασμένη έννοια με τα μοντέλα θέματος είναι η συνάφεια θέματος (Topic Coherence) ή αλλιώς συνοχή θέματος. Το Topic Coherence χρησιμοποιείται για την αξιολόγηση των θεματικών ενοτήτων που προκύπτουν από τα δεδομένα κειμένου. Συγκεκριμένα, όπως έχει ήδη προαναφερθεί, μία θεματική ενότητα είναι ένα σύνολο λέξεων, και ουσιαστικά προσπαθεί να αξιολογήσει τη συνάφεια των λέξεων αυτών, χρησιμοποιώντας πληροφορίες που αντλούνται είτε μέσα από τα δεδομένα κειμένου είτε από εξωτερικές πηγές. Για τον υπολογισμό της συνάφειας μίας θεματικής ενότητας απαιτείται η εκτέλεση και ο συνδυασμός τεσσάρων σταδίων (βλπ. Σχήμα 5.3) [9, 50]:



Σχήμα 5.3: Στάδια για τον υπολογισμό της Συνάφειας των Θεματικών ενοτήτων (Topic Coherence) [9]

1. **Κατάτμηση-Segmentation:**

   Σε αυτό το στάδιο γίνεται η ομαδοποίηση ενός συνόλου λέξεων σε μικρότερα υποσύνολα. Συγκεκριμένα, για μία θεματική ενότητα t με σύνολο n κορυφαίων λέξεων W = $w_1, w_2, \ldots, w_n$, το σύνολο των ζεύγων που μπορούν να αναπαραχθούν από τη διαδικασία της κατάτμησης συμβολίζεται με S και έχει την εξής γενική μορφή:

   $$S = \{(W', W^*), W', W^* \subseteq W\}, \tag{5.3}$$

   Υπάρχουν διάφοροι σχηματισμοί S, όπως:

   $$S^{one}_{one} = \{(W', W^*)|W' = \{w_i\}; W^* = \{w_j\}; w_i, w_j \in W; i \neq j\}, \tag{5.4}$$

   $$S^{one}_{pre} = \{(W', W^*)|W' = \{w_i\}; W^* = \{w_j\}; w_i, w_j \in W; i > j\}, \tag{5.5}$$

   $$S^{one}_{suc} = \{(W', W^*)|W' = \{w_i\}; W^* = \{w_j\}; w_i, w_j \in W; i < j\}, \tag{5.6}$$

   $$S^{one}_{all} = \{(W', W^*)|W' = \{w_i\}; w_i \in W; W^* = W \backslash \{w_i\}\}, \tag{5.7}$$

   $$S^{one}_{any} = \{(W', W^*)|W' = \{w_i\}; w_i \in W; W^* \subseteq W \backslash \{w_i\}\}, \tag{5.8}$$

   $$S^{any}_{any} = \{(W', W^*)|W', W^* \subset W; W' \cap W^* = \emptyset\}, \tag{5.9}$$

   $$S^{one}_{set} = \{(W', W^*)|W' = \{w_i\}; w_i \in W; W^* = W\} \tag{5.10}$$

2. **Υπολογισμός Πιθανότητας-Probability Calculation:**

   Η μέθοδος υπολογισμού πιθανότητας περιλαμβάνει τις διαφορετικές τεχνικές υπολογισμού των πιθανοτήτων ενδιαφέροντος αναφορικά με τα δεδομένα κειμένου, κάποιες από τις οποίες είναι οι εξής:

- **Boolean document (P_{bd}):** Υπολογίζει την πιθανότητα μίας λέξης $P(w_1)$, ως τον αριθμό των κειμένων/εγγράφων στα οποία εμφανίζεται η λέξη, διαιρεμένο με τον συνολικό αριθμό των κειμένων/εγγράφων. Αντίστοιχα και ο υπολογισμός της πιθανότητας εμφάνισης δύο λέξεων $P(w_1, w_2)$ ισούται με τον αριθμό των κειμένων στα οποία εμφανίζονται και οι δύο λέξεις προς τον αριθμό των συνολικών κειμένων. Υπάρχουν και δύο υποκατηγορίες, Boolean paragraph (P_{bp}), Boolean sentence (P_{bs}), που ουσιαστικά υπολογίζουν με τον ίδιο τρόπο, αλλά η καταμέτρηση γίνεται ως προς την παράγραφο ή πρόταση αντίστοιχα.

- **Boolean sliding window (P_{sw}):** Υπολογίζει με τον ίδιο τρόπο τις πιθανότητες όπως η τεχνική Boolean document (P_{bd}), απλά ως κείμενο ορίζεται το κομμάτι στο οποίο απευθύνεται το sliding window.

- **Word2vec (P_{w2v}):** Αυτή η τεχνική διαφέρει από τις υπολοιπές, καθώς απλά εκπαιδεύει ένα word2vec μοντέλο πάνω στα δεδομένα κειμένου που εισάχθηκαν, αν δεν ορίζεται ένα ήδη εκπαιδευμένο. Με αυτόν τον τρόπο δημιουργεί για κάθε λέξη μία διανυσματική αναπαράσταση, βασισμένη στο περιβάλλον λέξεων της καθεμίας.

3. **Μέτρο Επιβεβαίωσης-Confirmation Measure:**

   Αυτό το στάδιο, χρησιμοποιώντας τις πιθανότητες που υπολογίστηκαν στο δεύτερο στάδιο για κάθε $S_i = (W', W^*)$, υπολογίζει το πόσο καλά το υποσύνολο λέξεων W^* υποστηρίζει το υποσύνολο W'. Συμβολίζεται με m και προσπαθεί να ποσοτικοποιήσει τη σχέση μεταξύ W^* και W'. Το υψηλό m προδίδει καλή σχέση μεταξύ των υποσυνόλων. Υπάρχουν δύο είδη μέτρων επιβεβαίωσης που είναι τα εξής:

   - **Άμεσα μέτρα επιβεβαίωσης (Direct Confirmation Measures):** Υπολογίζουν το μέτρο επιβεβαίωσης άμεσα, χρησιμοποιώντας τις πιθανότητες που υπολογίστηκαν στο δεύτερο στάδιο για κάθε ζευγάρι από υποσύνολα. Μερικά παραδείγματα τέτοιων άμεσων μέτρων είναι τα:

$$m_r(S_i) = \frac{P(W', W^*)}{P(W')P(W^*)}, \tag{5.11}$$

$$m_{lr}(S_i) = log\frac{P(W', W^*) + \epsilon}{P(W')P(W^*)}, \tag{5.12}$$

$$m_{nlr}(S_i) = \frac{m_{lr}(S_i)}{-log(P(W', W^*) + \epsilon)}, \tag{5.13}$$

$$m_c(S_i) = \frac{P(W', W^*) + \epsilon}{P(W^*)}, \tag{5.14}$$

$$m_{lc}(S_i) = log\frac{P(W', W^*) + \epsilon}{P(W^*)}, \tag{5.15}$$

όπου ϵ μία μικρή τιμή ($\epsilon = 10^{-12}$).

► **Έμμεσα μέτρα επιβεβαίωσης (Indirect Confirmation Measures):**Ουσιαστικά, σε αυτήν την περίπτωση υπολογίζεται πρώτα το άμεσο μέτρο επιβεβαίωσης του υποσυνόλου λέξεων που ανήκουν στο W' με τις λέξεις που ανήκουν στο W και αντίστοιχα του υποσυνόλου λέξεων που ανήκουν στο W^* με τις λέξεις που ανήκουν στο W δημιουργώντας δύο διανύσματα:

$$\vec{u}_{m,\gamma}(W') = \left\{ \sum_{w_i \in W'} m(w_i, w_j)^\gamma \right\}_{j=1,\dots,|W|} \tag{5.16}$$

$$\vec{u}_{m,\gamma}(W^*) = \left\{ \sum_{w_i \in W^*} m(w_i, w_j)^\gamma \right\}_{j=1,\dots,|W|} \tag{5.17}$$

Στη συνέχεια υπολογίζεται η ομοιότητα μεταξύ των δύο αυτών διανυσμάτων, με μετρικές όπως ομοιότητα συνημιτόνου, και έτσι προκύπτει το μέτρο της επιβεβαίωσης:

$$m_{sim(m,\gamma)}(W', W^*) = s_{sim}(\vec{u}_{m,\gamma}(W'), \vec{u}_{m,\gamma}(W^*)) \tag{5.18}$$

Με αυτόν τον τρόπο, τονίζονται σχέσεις λέξεων που μπορεί να μην εμφανίζονται στο ίδιο κείμενο συχνά, αλλά να έχουν ισχυρή σύνδεση και αυτό να μπορεί να αποδειχτεί από το παρόμοιο πλαίσιο λέξεων που τοποθετούνται.

4. **Συνάθροισμα-Aggregation:**

   Το τελευταίο στάδιο είναι το συνάθροισμα, που ουσιαστικά εφαρμόζει μία συνάρτηση, όπως η αριθμητική μέση τιμή σ_α σε όλα τα αποτελέσματα που προήλθαν από το προηγούμενο βήμα για κάθε διαφορετικό συνδυασμό S_i, προκειμένου να προκύψει η συνάφεια-συνοχή θέματος.

Για τον υπολογισμό της συνολικής συνάφειας των θεματικών ενοτήτων του μοντέλου, ουσιαστικά εφαρμόζονται όλα τα παραπανω στάδια για κάθε θεματική ενότητα ξεχωριστά και αφού συλλεχθούν όλα τα αποτελέσματα για κάθε θεματική ενότητα στο στάδιο Συναθροίσματος, εφαρμόζεται μία συνάρτηση όπως η μέση αριθμητική τιμή.

5.2 Επεξήγηση μεθόδων αξιολόγησης

Σε αυτό το κεφάλαιο επεξηγούνται με λεπτομέρεια οι μέθοδοι αξιολόγησης που χρησιμοποιήθηκαν, καθώς και βήματα προετοιμασίας, απαραίτητα για την λειτουργία τους.

5.2.1 Υλοποίηση μοντέλου Word2vec

Ένα βήμα προετοιμασίας περιγράφεται σε αυτήν την ενότητα. Συγκεκριμένα, στην υλοποίηση των μεθόδων αξιολόγησης συμμετέχει και το μοντέλο Word2vec, καθώς η προσφορά του είναι σημαντική, αφού μπορεί και συσχετίζει σημασιολογικά λέξεις με την διανυσματική αναπαράσταση που ανακαλύπτει.

Στην προκειμένη εργασία, η υλοποίηση του μοντέλου Word2vec πραγματοποιήθηκε με χρήση της βιβλιοθήκης *gensim* [51]. Κάποιες από τις παραμέτρους που παρέχει η συνάρτηση *Word2vec* και χρησιμοποιήθηκαν για την υλοποίηση των μοντέλων είναι οι εξης:

- **window = X:** Αποτελεί το μέγεθος του πλαισίου λέξεων που χρησιμοποιείται, δηλαδή για μία λέξη ενδιαφέροντος το πλαίσιο λέξεων που την περικλείουν ορίζεται ως X λέξεις πριν και X λέξεις μετα από αυτήν.
- **min_count = X:** Αγνοεί τις λέξεις που εμφανίζονται λιγότερο από X φορές.
- **workers = X:** Για να τρέχει παράλληλα σε X επεξεργαστές και να επιταχυνθεί η διαδικασία.
- **sg= {0,1}:** Αν τεθεί 0 εφαρμόζεται η τεχνική CBOW, ενώ αν τεθεί ίσον με 1 εφαρμόζεται η τεχνική Skip-gram .
- **size = X:** Ο αριθμός X δηλώνει τον αριθμό των νευρώνων στο Hidden Layer και κατ'επέκταση το μέγεθος του αριθμητικού διανύσματος αναπαράστασης της λέξης.
- **epochs = X:** Ο αριθμός των φορών που εκτελείται το νευρωνικό δίκτυο χρησιμοποιώντας όλα τα δεδομένα εκπαίδευσης.

Δημιουργήθηκαν δύο μοντέλα Word2vec πάνω στα δεδομένα κειμένου των βιβλίων μετά την τελική προεπεξεργασία και οι παράμετροι που τέθηκαν ήταν sg=0, window=10 και 20 για τα δύο διαφορετικά μοντέλα, min_count=5, workers=4, size=200 και epochs=1000. Αυτά τα μοντέλα αποθηκεύτηκαν για την μετέπειτα ενσωμάτωση τους στις μετρικες.

5.2.2 Αυτόματη μετρική συνάφειας θέματος C_{umass} και C_{w2v}

Η βιβλιοθήκη *gensim* έχει έτοιμη υλοποιημένη συνάρτηση, την *CoherenceModel* [52], για τον υπολογισμό της συνολικής συνάφειας του μοντέλου, με διαφορετικές τεχνικές, ανάλογα με τις επιλογές στα διαφορετικά στάδια που περιγράφηκαν. Η συναρτηση αυτή παίρνει ως παραμέτρους το υλοποιημένο μοντέλο LDA για το οποίο θα μετρηθεί η συνάφεια θεμάτων, τα σύνολα κειμένων, το λεξιλόγιο που παράχθηκε από αυτά και τη μέθοδο συνάφειας θέματος που θα εφαρμοστεί. Στις παραμέτρους εισόδου προστίθενται και η εισαγωγή κάποιου προεκπαιδευμένου μοντέλου Word2vec ή η εισαγωγή του παραθύρου με το οποίο θα γίνει εσωτερικά η εκπαίδευση του μοντέλου Word2vec, σε περιπτώσεις που η μετρική συνάφειας είναι η C_{w2v}.

Συγκεκριμένα, στην προκειμένη εργασία υπολογίστηκε η συνάφεια χρησιμοποιώντας τα μέτρα C_{umass} και C_{w2v}. Η υλοποίηση που παρέχει η βιβλιοθήκη *gensim* ανά στάδιο, για τις δύο αυτές μετρικές θέματος αναγράφεται στον Πίνακα 5.1 [9, 50].

Πίνακας 5.1: Συνάφεια θέματος C_{umass} και C_{w2v}

Coherence:	$\mathbf{C_{umass}}$	$\mathbf{C_{w2v}}$
Segmentation:	S^{one}_{pre}	S^{one}_{set}
P. Calculation:	P_{bd}	P_{w2v}
C. Measure:	m_{lc}	$w2v\ similarity$
Aggregation:	σ_{α}	σ_{α}

Ο τρόπος υλοποίησης του μέτρου συνοχής θέματος C_{umass} είναι αρκετά κατανοητός καθότι αυτά που εφαρμόζονται έχουν εξηγηθεί στο 5.1.3. Για το μέτρο συνοχής C_{w2v} στο στάδιο P. Calculation είναι σημαντικό να τονιστεί πως χρησιμοποιήθηκε τόσο η εσωτερική εκπαίδευση που παρέχει η συνάρτηση(min_count=1, sg=1) [50] με window = 20, όσο και ήδη προεκπαιδευμένα μοντέλα. Τα προεκπαιδευμένα είναι τα δύο μοντέλα Word2vec που δημιουργήθηκαν πάνω στα δεδομένα κειμένου των βιβλίων, για το σκοπό της εργασίας, στην ενότητα 5.2.1, καθώς επίσης και ένα έτοιμο μοντέλο Word2vec που ήταν εκπαιδευμένο σε μία βάση δεδομένων με Google's News [53].

Επίσης να επισημανθεί πως στο μέτρο συνοχής C_{w2v}, υπολογίζεται έμμεσα το μέτρο επιβεβαίωσης μέσω της συνάρτησης *model.n_similarity* της βιβλιοθήκης *gensim*. Συγκεκρι-

μένα, βάσει της διανυσματικής αναπαράστασης των λέξεων από το Word2vec μοντέλο που έχει τεθεί στη μεταβλητή model, υπολογίζει πόσο όμοια είναι δύο σύνολα λέξεων.

Τέλος, όσο υψηλότερες οι τιμές των C_{umass} και C_{w2v}, τόσο πιο επιτυχημένο θεωρείται το μοντέλο.

5.2.3 Μετρική intra/inter ομοιότητα

Στη συνέχεια, η επόμενη μέθοδος που θα περιγραφεί, βασίζεται στην intra/inter συνάφεια, των θεματικών ενοτήτων που δημιουργούνται από το μοντέλο με χρήση του Word2vec μοντέλου [54, 55]. Πιο συγκεκριμένα, εξετάζει τόσο την συνάφεια μεταξύ των 20 κορυφαίων λέξεων εντός της κάθε θεματικής ενότητας (intra), όσο και την συνάφεια μεταξύ των θεματικών ενοτήτων (inter).

Για τη μετρική αυτή υλοποιήθηκαν δύο διαφορετικές εκδοχές, που και οι δύο χρησιμοποίησαν τα δύο μοντέλα Word2vec που εκπαιδεύτηκαν στα δεδομένα βιβλίων της εργασίας, καθώς επίσης και του έτοιμου προεκπαιδευμένου μοντέλου στα Google's News, προκειμένου να υπολογίσουν την συνάφεια μεταξύ των λέξεων.

Οι δύο εκδοχές για τη μετρική αυτή είναι οι εξής [56, 57]:

- $$Intra/Inter\,Ομοιοτητα = \frac{mean([intra\,Ομοιοτητα(t_i)]_{i=\{1,\dots,K\}})}{mean([inter\,Ομοιοτητα(t_j,t_l)]_{j,l=\{1,\dots,K\},j\neq l,j>l>0})} \tag{5.19}$$

  όπου ο αριθμητής και ο παρανομαστής του υπολογίζονται ξεχωριστά και θα εξηγηθούν παρακάτω.

- $$Intra/Inter\,Ομοιοτητα = mean\left(\left[\frac{\frac{intra\,Ομοιοτητα(t_i)+intra\,Ομοιοτητα(t_j)}{2}}{inter\,Ομοιοτητα(t_i,t_j)}\right]_{\substack{i,j=\{1,\dots,K\},\\ i\neq j,i>j>0}}\right) \tag{5.20}$$

  όπου το t συμβολίζει τη θεματική ενότητα και K είναι ο συνολικός αριθμός θεματικών ενοτήτων.

Για την intra Ομοιότητα:

Ουσιαστικά, είναι η αριθμητική μέση τιμή των ομοιοτήτων που υπολογίστηκαν για κάθε πιθανή δυάδα λέξεων της θεματικής ενότητας. Η ομοιότητα ανά δυάδα λέξεων υπολογίστηκε με τη συνάρτηση *model.similarity* της βιβλιοθήκης *gensim*, βασισμένο στην αναπαράσταση που προσφέρει για τις λέξεις το εκάστοτε Word2vec μοντέλο

Για την inter Ομοιότητα:

Υπολογίζει απλά την συνάφεια ως προς τις λέξεις των θεματικών ενοτήτων που τέθηκαν ως παράμετροι με χρήση της συνάρτησης *model.n_similarity* της ίδιας βιβλιοθήκης, που είναι υπεύθυνη για τον υπολογισμό των ομοιοτήτων μεταξύ συνόλων λέξεων βασισμένο στην αναπαράσταση που προσφέρει για τις λέξεις το εκάστοτε Word2vec μοντέλο

Είναι σημαντικό να σημειώθει ότι η συνάρτηση *model.similarity*, καθώς επίσης και η *model.n_similarity* υπολογίζουν την ομοιότητα των αριθμητικών διανυσμάτων των λέξεων με την ομοιότητα συνημιτόνου και η τιμή που μπορούν να επιστρέψουν κινείται από το [-1,1]. Για να μην επιστρέφονται αρνητικές τιμές, καθώς δε θα μπορεί να γίνει αντιληπτό αν προέρχονται από τον αριθμητή ή τον παρανομαστή, σε κάθε τιμή που υπολογίζεται με τη χρήση αυτών των συναρτήσεων προστίθεται η τιμή 1, για να υπάγεται σε ένα θετικό διάστημα, στο [0,2].

Τέλος, και για τις δύο εκδοχές ισχύει ότι όσο υψηλότερη η τιμή αυτού του κλάσματος τόση μεγαλύτερη σημασιολογική ευκρίνεια έχουν οι θεματικές ενότητες εσωτερικά και τόσο λιγότερη σημασιολογική συσχέτιση έχουν οι θεματικές ενότητες μεταξύ τους, που είναι και το επιθυμητό.

5.2.4 Μετρική απόστασης θεματικών ενοτήτων

Τελευταία μέθοδος αξιολόγησης του μοντέλου LDA αποτελεί η απόσταση θεματικών ενοτήτων. Διαφέρει από τις προαναφερθείσες μετρικές, καθώς δε βασίζεται ούτε στο περιβάλλον των λέξεων ούτε στις ίδιες τις λέξεις, προκειμένου να προκύψουν συναφείς θεματικές ενότητες, αλλά ασχολείται με την στατιστική απόσταση των θεματικών ενοτήτων [55, 58, 59].

Συγκεκριμένα, στο μοντέλο Λανθάνουσας Κατανομής Dirichlet, δημιουργείται μία Dirichlet κατανομή θεματικών ενοτήτων και λέξεων. Επομένως, κάθε θεματική ενότητα μπορεί να αντιστοιχηθεί με τις λέξεις με μία κατανομή.

Πρώτο βήμα λοιπόν για τη δημιουργία της μετρικής, είναι η δημιουργία ενός πίνακα ΛxK, όπου Λ ο αριθμός των λέξεων και K ο αριθμός των θεμάτων. Κάθε στήλη αυτού του πίνακα περιέχει την κατανομή του θέματος προς το σύνολο των λέξεων. Στην προκειμένη εργασία, το σύνολο των λέξεων συμπεριλαμβάνει τις 50 πιο αντιπροσωπευτικές λέξεις για κάθε θεματική ενότητα, χωρίς διπλότυπες εγγραφές λέξεων.

Στη συνέχεια, για κάθε πιθανή δυάδα θεματικών ενοτήτων υπολογίζεται η απόσταση των κατανομών τους ως προς το σύνολο λέξεων του πίνακα, με χρήση της απόστασης Jensen-Shannon. Τέλος, επιστρέφεται η μέση αριθμητική τιμή των απόστάσεων που υπολογίστηκαν. Μια μεγάλη τιμή υποδεικνύει υψηλή διαφορά στις κατανομές των θεματικών ενοτήτων ως προς το σύνολο των λέξεων, που είναι και το επιθυμητό, καθώς δηλώνει την ευκρίνεια των θεματικών ενοτήτων.

5.3 Σύνοψη μεθόδων αξιολόγησης

Δεδομένου όσα προαναφέρθηκαν οι τελικές μετρικές αξιολόγησης είναι οι παρακάτω:

1. Αυτόματη μετρική Συνάφειας θέματος C_{umass}.
2. Αυτόματη μετρική Συνάφειας θέματος C_{w2v}, χωρίς προεκπαιδευμένο μοντέλο.
3. Αυτόματη μετρική Συνάφειας θέματος C_{w2v}, με χρήση των τριών προεκπαιδευμένων μοντέλων Word2vec.
4. Μετρική Intra/Inter Ομοιότητα πρώτης εκδοχής με χρήση των τριών προεκπαιδευμένων μοντέλων Word2vec.
5. Μετρική Intra/Inter Ομοιότητα δεύτερης εκδοχής με χρήση των τριών προεκπαιδευμένων μοντέλων Word2vec.
6. Μετρική απόστασης θεματικών ενοτήτων

Επομένως, συνολικά το μοντέλο θα αξιολογηθεί από δώδεκα μετρικές.

Κεφάλαιο 6

Ανάπτυξη Μοντέλου LDA και Συντονισμός Παραμέτρων

Στο παρόν κεφάλαιο θα επεξηγηθεί η διαδικασία ανάπτυξης και υλοποίησης του μοντέλου LDA προγραμματιστικά, πάνω στα προεπεξεργασμένα δεδομένα. Στη συνέχεια, θα γίνει μία περιγραφή της διαδικασίας συντονισμού των παραμέτρων του μοντέλου. Τέλος, θα εξαχθούν τα καλύτερα αποτελέσματα μέσα από το συντονισμό και θα επιλεχθεί το μοντέλο, με το οποίο θα πραγματοποιηθεί τελικά η σύσταση.

6.1 Εφαρμογή της υπόθεσης BoW στα προεπεξεργασμένα δεδομένα

Όπως προαναφέρθηκε στην ενότητα 2.3.1, η υπόθεση BoW είναι ένας τρόπος να αναπαρασταθούν τα δεδομένα κειμένου σε μορφή διαχειρίσιμη για τα διάφορα μοντέλα. Στην προκειμένη εργασία θα χρησιμοποιηθεί αυτή η υπόθεση, έτσι ώστε τα δεδομένα να μπορούν να εισαχθούν στο μοντέλο LDA. Η διαδικασία δημιουργίας λεξιλογίου και η καταμέτρηση εμφάνισης των λέξεων επιτεύχθηκε μέσω της βιβλιοθήκης *gensim*.

Πιο συγκεκριμένα, το λεξιλόγιο δημιουργήθηκε με τη συνάρτηση *corpora.Dictionary()* [60], που ως είσοδο παίρνει τα δεδομένα κειμένου σε μορφή λέξεων, και στην περίπτωση αυτής τη εργασίας τα προεπεξεργασμένα δεδομένα κειμένου, και κάθε λέξη την αντιστοιχίζει με ένα ακέραιο αναγνωριστικό (βλπ. Σχήμα 6.1). Έπειτα, προκειμένου να μην υπάρχουν σπάνιες λέξεις στο λεξιλόγιο, καθώς και συχνά εμφανιζόμενες, εφαρμόστηκε η συνάρτηση *filter_extremes*, στο λεξιλόγιο και αφαιρέθηκαν οι λέξεις που εμφανίζονται λιγότερο από

15 φορές συνολικά στα κείμενα, όπως επίσης και οι λέξεις που εμφανίζονται στο 50% των κειμένων.

```
Out[15]: {'action': 0,
          'annotation': 1,
          'apache': 2,
          'application': 3,
          'base': 4,
          'business': 5,
          'centric': 6,
          'change': 7,
          'chapter': 8,
          'choice': 9,
          'code': 10,
          'component': 11,
          'configuration': 12,
          'configure': 13,
          'connect': 14,
          'control': 15,
          'create': 16,
          'database': 17,
          'deployment': 18,
```

Σχήμα 6.1: Απεικόνιση Λεξιλογίου

Στη συνέχεια, με βάση αυτό το λεξιλόγιο και με τη χρήση της συνάρτησης *doc2bow*, δημιουργήθηκε η αναπαράσταση κάθε κειμένου από ένα σώμα που περιέχει τα αναγνωριστικά των λέξεων του κειμένου μαζί με τον αριθμό της συχνότητας εμφάνισης τους στο κείμενο. Στο Σχήμα 6.2 παρουσιάζεται πώς αναπαριστάται ένα κείμενο, όπου (0,1) δηλώνει ότι η λέξη με το αναγνωριστικό 0 εμφανίζεται 1 φορά στο κείμενο. Τελικώς, όλες οι περιγραφές των βιβλίων έχοντας ενσωματώσει και τον τίτλο τους, αναπαρίστανται κατά αυτόν τον τρόπο, προκειμένου να εισαχθούν στο μοντέλο LDA.

```
[(0, 1), (1, 1), (2, 1), (3, 6), (4, 1), (5, 3), (6, 1), (7, 1), (8, 1), (9, 3), (10, 2), (11, 1), (12, 1), (13, 1), (14, 1),
(15, 1), (16, 3), (17, 2), (18, 1), (19, 1), (20, 2), (21, 1), (22, 1), (23, 1), (24, 1), (25, 5), (26, 2), (27, 1), (28, 1),
(29, 1), (30, 1), (31, 1), (32, 1), (33, 1), (34, 1), (35, 1), (36, 1), (37, 1), (38, 1), (39, 1), (40, 3), (41, 10), (42, 1),
(43, 1), (44, 1), (45, 1), (46, 1), (47, 2), (48, 2), (49, 1), (50, 1), (51, 1), (52, 1), (53, 1), (54, 1), (55, 1), (56, 1),
(57, 2), (58, 1), (59, 1), (60, 5), (61, 4), (62, 2)]
```

Σχήμα 6.2: Απεικόνιση κειμένου ως BoW

6.2 Υλοποίηση Λανθάνουσας Κατανομής Dirichlet (LDA)

Η υλοποίηση του μοντέλου LDA πραγματοποιήθηκε μέσω της βιβλιοθήκης *gensim*, με χρήση του μοντέλου *LdaMulticore* [61], που χρησιμοποιεί τον αλγόριθμο Online Variational Bayes, η οποία είναι μία πιο γρήγορη εκδοχή του Variational Bayes Inference [62] για τον υπολογισμό της μεταγενέστερης πιθανότητας. Η συνάρτηση αυτή δημιουργεί ένα μοντέλο LDA αναφορικά με τα δεδομένα, που δέχεται ως είσοδο, μέσα από το οποίο μπορούν να εξαχθούν το πώς τελικά κατανέμονται οι λέξεις στις θεματικές ενότητες, καθώς επίσης το πώς οι θεματικές ενότητες κατανέμονται στα κείμενα.

Μερικές από τις εισόδους που δέχεται η συνάρτηση αυτή και που χρησιμοποιήθηκαν σε αυτήν την εργασία προκειμένου να παραχθεί το μοντέλο LDA:

- **corpus:** το σύνολο των κειμένων σε μορφή BoW,
- **num_topics:** ο αριθμός των θεματικών ενοτήτων που θα ανακαλύψει από το σύνολο των κειμένων,
- **id2word:** το λεξιλόγιο των κειμένων, δηλαδή η αντιστοίχιση των λέξεων σε ακέραια αναγνωριστικά,
- **chunksize:** ο αριθμός κειμένων που χρησιμοποιούνται σε κάθε επανάληψη εκπαίδευσης,
- **passes:** ο αριθμός φορών που διατρέχεται όλο το σύνολο κειμένων κατά την εκπαίδευση,
- **iterations:** ο μέγιστος αριθμός επαναλήψεων στο σύνολο κειμένων όταν γίνεται η ανάθεση κατανομής θεματικών ενοτήτων σε ένα κείμενο.
- **alpha:** η παράμετρος Dirichlet Κατανομής, που ρυθμίζει το πώς κατανέμονται οι θεματικές ενότητες στα κείμενα. Οι τιμές που μπορεί να δεχτεί είναι οι εξης:
  - έναν αριθμό, για μία προεπιλεγμένη συμμετρική κατανομή,
  - ένας 1D πίνακας, μεγέθους ίσου με num_topics, για μία προεπιλεγμένη ασύμμετρη κατανομή,
  - “symmetric”, που ουσιαστικά είναι ένας 1D πίνακας μεγέθους num_topics, με ίδιες τιμές ίσες με $\frac{1.0}{num\_topics}$,
  - “asymmetric”, που ουσιαστικά είναι ένας 1D πίνακας μεγέθους num_topics, με διαφορετικές τιμές σύμφωνα με τον τύπο $\frac{1.0}{topic\_index+\sqrt{num\_topics}}$.
- **eta:** η παράμετρος Dirichlet Κατανομής, που ρυθμίζει το πώς κατανέμονται οι λέξεις στις θεματικές ενότητες. Οι τιμές που παίρνει είναι οι παρακάτω:
  - έναν αριθμό, για μία προεπιλεγμένη συμμετρική κατανομή,
  - ένας 1D πίνακας, μεγέθους ίσου με num_topics, για μία προεπιλεγμένη ασύμμε-

τρη κατανομή,

- έναν πίνακα μεγέθους num_topicsXnum_words, για να οριστεί μία πιθανότητα για κάθε συνδυασμό θεματικής ενότητας-λέξης,
- "symmetric", που ουσιαστικά είναι ένας 1D πίνακας μεγέθους num_topics, με ίδιες τιμές ίσες με $\frac{1.0}{num\_topics}$
- "auto", που ουσιαστικά εκπαιδεύεται αυτόματα μία ασύμμετρη κατανομή

► **minimum_probability:** ο αριθμός της μικρότερης πιθανότητας στην κατανομή θεματική ενότητα-κείμενο, που είναι επιτρεπτός για να επιστραφεί.

Στην προκειμένη εργασία, στις παραμέτρους corpus και id2word, εισήχθησαν τα κείμενα σε μορφή BoW και το λεξιλόγιο των κειμένων, η δημιουργία των οποίων συζητήθηκε στην προηγούμενη ενότητα. Επίσης, η παράμετρος minimum_probability τέθηκε ίσον με 0 προκειμένου να επιστραφούν όλες οι πιθανότητες, καθώς είναι σημαντικό για τη σύσταση των κειμένων να υπάρχει όλη η κατανομή των θεματικών ενοτήτων για κάθε κείμενο.

Για τις παραμέτρους passes, chunksize και iteration, πραγματοποιήθηκαν αρκετές δοκιμές, και με έλεγχο των μηνυμάτων που επιστρέφει το μοντέλο, κατά την εκτέλεσή του, σχετικά με τη σύγκλιση, τέθηκαν τελικά οι τιμές passes = 30, chunksize = 1000, iteration = 200. Οι τιμές αυτές δεν αποτελούν τις καλύτερες δυνατές τιμές που μπορεί να θέσει κάποιος προκειμένου να επιτύχει ένα καλό μοντέλο και σε γρήγορο χρόνο με αυτά τα δεδομένα, αλλά η προκειμένη εργασία επικεντρώθηκε στον συντονισμό των παραμέτρων num_topics, alpha και eta. Η προσοχή στράφηκε προς τα εκεί καθώς είναι ξεκάθαρο οτι επηρεάζουν άμεσα και την ποιότητα και τον αριθμό των θεματικών ενοτήτων, που θα ανακαλυφθούν μέσα από το σύνολο κειμένων.

6.3 Συντονισμός παραμέτρων και επιλογή καλύτερου μοντέλου

Ο συντονισμός παραμέτρων υπήρξε μία αρκετά χρονοβόρα διαδικασία, καθώς εκτελέσθηκε σαν μία μεγάλη επανάληψη για κάθε πιθανή τριάδα των παραμέτρων num_topics, alpha και eta. Πιο λεπτομερώς, οι τιμές για κάθε παράμετρο ήταν οι παρακάτω:

► **num_topics:** [16,64] με βήμα 2

- **alpha:** {0.01, 0.31, 0.61, "symmetric", "asymmetric"}
- **eta:** {0.01, 0.31, 0.61, "symmetric"}

Οι αριθμητικές τιμές στο alpha και eta επιλέχθηκαν να είναι μικρότερες του 1, καθώς από το Σχήμα 2.5 έγινε κατανοητό πως έτσι επιτυγχάνεται μεγαλύτερη διακριτοποίηση.

Στη συνέχεια, για κάθε διαφορετικό συνδυασμό των τριών αυτών παραμέτρων εκπαιδευόταν ένα μοντέλο LDA, το οποίο και αποθηκευόταν. Με βάση τις κορυφαίες λέξεις των θεματικών ενοτήτων του δεδομένου, κάθε φορά, μοντέλου, υπολογίζονταν και οι δώδεκα μετρικές αξιολόγησης που υλοποιήθηκαν για αυτήν την εργασία, και οι οποίες επεξηγήθηκαν λεπτομερώς στο Κεφάλαιο 5.

	Topics	Alpha	Beta	Auto Coherence Umass	Intra/Inter Similarity word2vec10 1st version	Intra/Inter Similarity word2vec20 1st version	Intra/Inter Similarity word2vecGoogle 1st version	Intra/Inter Similarity word2vec10 2nd version	Intra/Inter Similarity word2vec20 2nd version	Intra/Inter Similarity word2vecGoogle 2nd version
0	16	0.01	0.01	-2.704128	1.218417	1.212401	0.722715	1.101374	1.095362	1.008432
1	16	0.01	0.31	-2.604525	1.198338	1.192703	0.717730	1.097292	1.091642	1.007582
2	16	0.01	0.61	-2.558049	1.204514	1.197006	0.721481	1.099346	1.093183	1.007874
3	16	0.01	symmetric	-2.677200	1.207011	1.203086	0.716035	1.100449	1.094700	1.007074
4	16	0.31	0.01	-2.723212	1.242661	1.234717	0.717912	1.107527	1.101567	1.007298
...	...	...	...	...	...	...	...	...	...	...
495	64	symmetric	symmetric	-5.594661	1.162849	1.150306	0.731904	1.078796	1.072782	1.009707
496	64	asymmetric	0.01	-4.851628	1.168581	1.158129	0.730980	1.082444	1.076744	1.009179
497	64	asymmetric	0.31	-5.678161	1.174731	1.159449	0.738081	1.084107	1.077017	1.011565
498	64	asymmetric	0.61	-7.232952	1.163719	1.147136	0.738656	1.077627	1.070593	1.013159
499	64	asymmetric	symmetric	-5.014887	1.171610	1.160281	0.728651	1.083066	1.077134	1.009251

500 rows × 16 columns

Σχήμα 6.3: Απεικόνιση ενός τμήματος των αποτελεσμάτων των μετρικών αξιολόγησης κατά τον συντονισμό των παραμέτρων num_topics, alpha και eta

Επειδή για κάθε μετρική αξιολόγησης ισχύει πως όσο μεγαλύτερη αριθμητικά είναι η τιμή της, τόσο καλύτερο το μοντέλο LDA, εξάχθηκαν από τα συνολικά αποτελέσματα οι καλύτεροι συνδυασμοί παραμέτρων σύμφωνα με κάθε μετρική, οι οποίοι παρουσιάζονται στον Πίνακα 6.1.

Πίνακας 6.1: Οι καλύτεροι συνδυασμοί παραμέτρων βάσει των μετρικών αξιολόγησης

	num_topics	alpha	eta	Score	OOV	Ανθρώπινη κρίση
C_{umass}	16	asymmetric	0.61	-2.535	-	0.75
Intra/Inter Ομ./1η εκδ./W2v στα βιβλία με window=10	18	0.61	0.01	1.253	-	0.72
Intra/Inter Ομ./1η εκδ./W2v στα βιβλία με window=20	18	0.61	0.01	1.247	-	0.72
Intra/Inter Ομ./1η εκδ./W2v Google's news	50	asymmetric	0.61	0.748	0.9%	0.78
Intra/Inter Ομ./2η εκδ./W2v στα βιβλία με window=10	16	0.61	0.61	1.112	-	0.75
Intra/Inter Ομ./2η εκδ./W2v στα βιβλία με window=20	16	0.61	0.61	1.107	-	0.75
Intra/Inter Ομ./2η εκδ./W2v Google's news	62	asymmetric	0.61	1.016	0.89%	0.68
C_{w2v} χωρίς προεκπαιδευμένο W2v	16	0.61	0.01	0.746	-	0.75
C_{w2v}/W2v στα βιβλία με window=10	16	0.61	0.61	0.525	-	0.75
C_{w2v}/W2v στα βιβλία με window=20	16	0.61	0.61	0.51	-	0.75
C_{w2v}/W2v Google's news	32	0.61	0.61	0.636	0.2%	0.69
Απόσταση θεματικών ενοτήτων	60	0.61	0.01	0.831	-	0.75

Η στήλη OOV(Out Of Vocabulary) αναγράφει το ποσοστό των λέξεων που υπήρχαν στις θεματικές ενότητες αλλά δεν αναγνωρίζονταν από το έτοιμο προεκπαιδευμένο μοντέλο Word2vec στα Google's news.

Η στήλη "Ανθρώπινη κρίση", περιέχει την ανθρώπινη αξιολόγηση ως προς τις θεματικές ενότητες που προέκυψαν. Πιο λεπτομερώς, για καθένα από τα καλύτερα μοντέλα του Πίνακα 6.1, αξιολογήθηκαν οι θεματικές ενότητες που δημιουργήθηκαν ως έγκυρες ή άκυρες, και υπολογίστηκε το ποσοστό των έγκυρων ($\frac{number\_right\_topics}{number\_topics}$). Σημαντικό είναι να επισημανθεί ότι η ανθρώπινη κρίση βασίστηκε μόνο στις 10, πιο υψηλα συσχετισμένες με την θεματική ενότητα, λέξεις.

Τελικώς, το μοντέλο που κρίνεται καλύτερο σύμφωνα και με την ανθρώπινη κρίση, είναι το μοντέλο με τις παραμέτρους num_topics = 50, alpha = "asymmetric" και eta = 0.61, το οποίο παρουσιάζεται στην επόμενη ενότητα.

6.3.1 Παρουσίαση καλύτερου μοντέλου LDA

Το τελικό μοντέλο LDA που θα χρησιμοποιηθεί και για το σύστημα σύστασης, αποτελείται από 50 θεματικές ενότητες, κάποιες από τις οποίες παρουσιάζονται στον Πίνακα 6.2. Για την οπτικοποίηση του μοντέλου η python παρέχει τη βιβλιοθήκη *pyLDAvis* [63]. Εισά-

Πίνακας 6.2: 7 από τις 50 θεματικές ενότητες (Topics) με τις 10 κορυφαίες λέξεις

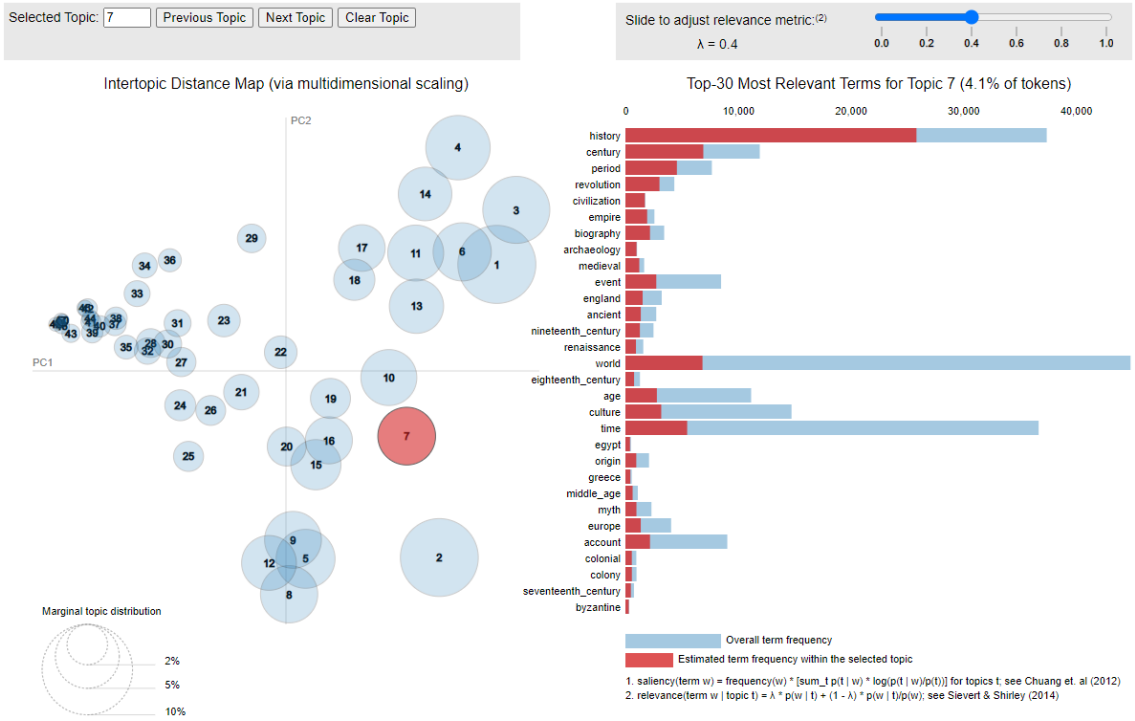
Topic 0	**Topic 9**	**Topic 11**	**Topic 14**	**Topic 17**	**Topic 21**	**Topic 33**
murder	plant	god	history	health	business	science
crime	water	church	century	disease	technology	university
death	bird	religion	world	treatment	company	professor
mystery	specie	prayer	time	care	system	scientist
police	animal	theology	period	drug	management	department
victim	conservation	christianity	life	medicine	information	intelligence
killer	ecology	life	people	therapy	network	research
life	environment	study	year	disorder	design	college
investigation	soil	scripture	culture	patient	industry	institute
body	nature	word	revolution	cancer	service	cambridge

γοντας στη συνάρτηση *prepare*, το μοντέλο, καθώς και το σύνολο των κειμένων σε μορφή BoW και το λεξιλόγιο, υλοποιείται αυτόματα μία διαδραστική οπτικοποίηση των θεματικών ενοτήτων, που απεικονίζεται στο στο Σχήμα 6.4.

Στο αριστερο τμήμα που ονομαζετα και Intertopic Distance Map οι κύκλοι αναπαριστούν τις θεματικές ενότητες. Το μέγεθος των κύκλων αντικατοπτρίζει το πόσο συχνά αυτή η θεματική ενότητα εμφανίζεται στα κείμενα. Επίσης ο τρόπος με τον οποίο οι θεματικές ενότητες είναι τοποθετημένες στον χάρτη έχει να κάνει με τη συσχέτιση μεταξύ τους, τα παρόμοια θέματα εμφανίζονται κοντά ενώ τα ανόμοια πιο μακρια. Επίσης, είναι σημαντικό να τονιστεί πως η αρίθμηση των θεματικών ενοτήτων δεν ταυτίζεται με την αρίθμηση που προκύπτει από το μοντέλο.

Το δεξί τμήμα εμφανίζει τις 30 πιο σχετικές λέξεις με το θέμα που έχει επιλεγεί, ενώ αν δεν έχει επιλεγεί κάποια θεματική ενότητα εμφανίζει τις 30 πιο συχνά εμφανιζόμενες λέξεις του συνόλου των κειμένων. Οι λέξεις που παρουσιάζονται μπορούν να μεταβληθούν μετακι-

νώντας την μπάρα στο πάνω μέρος, η οποία ορίζει τη συνάφεια μέσω της παραμέτρου λ. Πιο λεπτομερώς, όσο ελαττώνεται η τιμή του λ τείνει να εμφανίζει λέξεις που έχουν μικρότερη συχνότητα εμφάνισης, αλλά εμφανίζονται κατά βάση σε αυτή τη θεματική ενότητα.



Σχήμα 6.4: Οπτικοποίηση θεματικών ενοτήτων καλύτερου μοντέλου LDA μέσω της βιβλιοθήκης pyLDAvis

Κεφάλαιο 7

Σύστημα Σύστασης

Το ΣΣ που υλοποιήθηκε στην παρούσα διπλωματική εργασία, αποτελείται από τις εξής λειτουργίες:

- Εισαγωγή του τίτλου ή του ISBN ενός βιβλίου και η σύσταση παρόμοιων θεματικά βιβλίων, αλλά και βιβλίων παρόμοιων θεματικά που έχουν διαβάσει και βαθμολογήσει θετικά οι χρήστες
- Εισαγωγή ενος κειμένου περιγραφής από οποιαδήποτε πηγή και η σύσταση παρόμοιων θεματικά βιβλίων
- Εισαγωγή ενός ερωτήματος και η σύσταση με βάση τις λέξεις εισαγωγής

Με αυτές τις λειτουργίες το σύστημα δεν υπάγεται ξεκάθαρα σε καμία από τις κατηγορίες που περιγράφηκαν στο θεωρητικό υπόβαθρο της εργασίας. Συγκεκριμένα, πρόκειται για ένα σύστημα, το οποίο αφενώς στηρίζεται στο περιεχόμενο για να πραγματοποιήσει τη σύσταση, αφετέρου δεν χρησιμοποιεί παρελθοντικές προτιμήσεις του ίδιου του χρήστη προκειμένου να πραγματοποιήσει κάποια σύσταση. Επομένως, διαφέρει σε αυτό το σημείο από τα ΣΣ βασισμένα στο περιεχόμενο. Επίσης όσον αφορά τα συστήματα σύστασης συνεργατικού φιλτραρίσματος επειδή ακριβώς δε συσχετίζει χρήστες δεν υπάγεται σε αυτήν την κατηγορία. Η βάση με τους χρήστες αξιοποιείται, για την επιπλέον πρόταση βιβλίων, σχετικών με το βιβλίο αναζήτησης, για τα οποία υπάρχει και η θετική βαθμολόγηση από τους χρήστες. Βέβαια, επειδή για να λειτουργήσει δε βασίζεται στο προφίλ του χρήστη, δεν έρχεται αντιμέτωπο με το πρόβλημα της Ψυχρής Εκκίνησης, αλλά αντιθέτως αποτελεί και έναν τρόπο αποφυγής του. Στις επόμενες ενότητες, θα πραγματοποιηθεί μία πιο αναλυτική επεξήγηση του ΣΣ.

7.1 Υλοποίηση λειτουργιών συστήματος σύστασης

Πιο συγκεκριμένα για τις λειτουργίες του ΣΣ της εργασίας, ουσιαστικά πρωταρχικό βήμα ήταν η ανάθεση μίας κατανομής σε κάθε βιβλίο της βάσης δεδομένων, ως προς τις θεματικές ενότητες που ανακαλύφθηκαν. Αυτό υλοποιήθηκε με χρήση του τελικού μοντέλου LDA και της *get_document_topics*. Η αρχική ιδέα βέβαια ήταν η συσχέτιση των κειμένων να γίνεται με την κυρίαρχη πιθανότητα ως προς τις θεματικές ενότητες, αλλά κάτι τέτοιο αποδήχτηκε ατελέσφορο, καθώς ένα κείμενο με τις λέξεις που περιέχει κατατάσσεται σε παραπάνω από μία ενότητες, και πολλές φορές με το ίδιο ποσοστό, επομένως η συσχέτιση θα ήταν αποτυχημένη. Συνεπώς, σε κάθε βιβλίο αποδόθηκε μία κατανομή ως προς όλες τις θεματικές ενότητες και βάσει αυτής πραγματοποιήθηκαν οι λειτουργίες.

7.1.1 Σύσταση με εισαγωγή του τίτλου/ISBN του βιβλίου

Η πρώτη λειτουργία που θα αναλυθεί είναι αυτή της εισαγωγής του τίτλου ενός βιβλίου ή του αναγνωριστικού του ISBN και η επεξήγηση της θα διακριθεί σε περιπτώσεις.

Εισαγωγή τίτλου/ISBN που δεν υπάρχει στη βάση δεδομένων με τα βιβλία:

Στην περίπτωση που γίνει εισαγωγή ενός τίτλου βιβλίου που δεν υπάρχει στη βάση, έχει δημιουργηθεί ένα σύστασημα σύστασης βάσει της ομοιότητας του τίτλου εισαγωγής με τους τίτλους που υπάρχουν στη βάση. Για να επιτευχθεί αυτό έγινε χρήση της συνάρτησης *TfidfVectorizer* της βιβλιοθήκης *sklearn*, που ουσιαστικά σκοπός της είναι να μετατρέπει το κείμενο σε μία αναπαράσταση αριθμών ως προς το συνολικό λεξιλόγιο των κειμένων. Συγκεκριμένα, ο αριθμός που αντιστοιχίζεται σε κάθε λέξη i του j-οστου κειμένου, που ονομάζεται και βάρος της λέξης, υπολογίζεται από τον τύπο [64]:

$$w_{i,j} = tf_{i,j} \cdot log(\frac{N}{df_i}), \tag{7.1}$$

όπου tf σημαίνει term frequency και είναι η συχνότητα της i-οστής λέξης στο j-οστό κείμενο, ενώ ο δεύτερος όρος του πολλαπλασιασμού ονομάζεται idf και σημαίνει η αντίστοφη συχνότητα κειμένου, όπου το df ποσοτικοποιεί σε πόσα κείμενα εμφανίζεται η i-οστή λέξη.

Επομένως, όλα τα ονόματα των κειμένων μετατρέπονται σύμφωνα με αυτήν την αναπαράσταση, καθώς και ο τίτλος αναζήτησης του χρήστη με χρήση του ίδιου λεξιλογίου, και στη συνέχεια συγκρίνοντας τα διανύσματα αναπαράστασης, με τη μετρική ομοιότητας συνημι-

τόνου, εξάγονται τα 20 ονόματα βιβλίων που εμφανίζουν μεγαλύτερη ομοιότητα με αυτό που αναζητήθηκε.

Στην περίπτωση εισαγωγής ενός αναγνωριστικού ISBN που δεν περιέχετε στη βάση, εκτυπώνεται ένα μήνυμα ενημέρωσης ότι δεν υπάρχει αυτό που αναζητήθηκε.

Εισαγωγή τίτλου/ISBN που υπάρχει στη βάση δεδομένων με τα βιβλία:

Στην περίπτωση που ο τίτλος ή το ISBN που εισήχθηκαν υπάρχει στη βάση, γίνεται μία σύγκριση της κατανομής ως προς τις θεματικές ενότητες του βιβλίου που εισήχθει, με όλες τις υπόλοιπες κατανομές των βιβλίων προκειμένου να βρεθούν οι πιο όμοιες. Η σύγκριση υλοποιήθηκε μέσω της απόστασης Jensen-Shannon, που περιγράφηκε στο 2.4. Δεν χρησιμοποιήθηκε κάποια έτοιμη συνάρτηση, απλά έγινε εφαρμογή του τύπου (2.8). Σημαντικό είναι να επισημανθεί, πως δοκιμάστηκαν και άλλες μετρικές απόστασης, όπως η Hellinger Distance, καθώς επίσης και η Ομοιότητα Συνημιτόνου, αλλά το σύστημα φάνηκε να είναι πιο επιτυχημένο με την απόσταση Jensen-Shannon. Τελικώς, συστήνονται έως 15 βιβλία, τα οποία θα πρέπει να έχουν απόσταση, με το βιβλίο αναζήτησης, μικρότερη από 0.35, ως προς τις κατανομές τους.

Επίσης στην περίπτωση που το βιβλίο ή το ISBN που εισήχθηκε από τον χρήστη υπάρχει στη βάση με τις βαθμολογήσεις των χρηστών, γίνεται μία επιπλέον σύσταση. Συγκεκριμένα, αυτή η σύσταση αποτελείται από τα παρακάτω βήματα:

- Αναζήτηση χρηστών στη βάση, που έχουν βαθμολογήσει θετικά αυτό το βιβλιο και αποθήκευση των ID τους.
- Με βάση αυτά τα ID αναζήτηση των υπόλοιπων βιβλίων που έχουν βαθμολογήσει και αποθήκευση των ονομάτων και των ISBN τους.
- Αντιστοίχιση με τη βάση με τα δεδομένα με τα βιβλία, ως προς τα ISBN, προκειμένου να παρθεί για καθένα από αυτά η κατανομή που έχουν ως προς τις θεματικές ενότητες.
- Σύσταση 15 βιβλίων με σειρά φθίνουσα ως προς την ομοιότητα της κατανομης των βιβλίων με την κατανομή του βιβλίου αναζήτησης. Και σε αυτό το σημείο ελέγχεται η απόσταση μεταξύ των κατανομών να είναι μικρότερη του 0.35

Με την προσθήκη και αυτής της σύστασης, στον χρήστη της εφαρμογής προσφέρονται αφενώς βιβλία που είναι όμοια θεματικά με το βιβλίο αναζήτησης, αφετέρου βιβλία που έχουν αρέσει στους χρήστες. Επίσης με αυτόν τον τρόπο έρχονται στην επιφάνεια βιβλία που η

ομοιότητα μπορεί να τα είχε κρατήσει πίσω στη λίστα σύστασης, λόγω της λίγο πιο χαμηλής ομοιότητας.

7.1.2 Σύσταση με εισαγωγή ενός κειμένου περιγραφής

Η δεύτερη λειτουργία είναι η σύσταση με βάση ενός κειμένου περιγραφής. Υλοποιήθηκαν δύο στάδια για την σωστή εκτέλεση αυτής της λειτουργίας. Συγκεκριμένα, το πρώτο στάδιο ήταν η προεπεξεργασία του κειμένου εισαγωγής, που περιλάμβανε όλα τα στάδια προεπεξεργασίας που αναφέρθηκαν στην ενότητα 4.4.

Στη συνέχεια, αφού το κείμενο έχει καθαριστεί και αποτελείται από λέξεις που έχουν αναγνωριστεί ως ουσιαστικά, με χρήση του λεξιλογίου με το οποίο εκπαιδεύτηκε το μοντέλο LDA, μετατρέπεται σε μορφή BoW. Αφού μετατραπεί στην κατάλληλη μορφή, με χρήση της συνάρτησης *get_document_topics* και αναφορικά με το ήδη εκπαιδευμένο μοντέλο LDA, του ανατίθεται μία κατανομή ως προς τις προϋπάρχουσες θεματικές ενότητες. Με βάση αυτήν την κατανομή και με τη χρήση της απόστασης Jensen-Shannon γίνεται η εύρεση των βιβλίων με τις πιο όμοιες κατανομές και συστήνονται τα 15 πιο όμοια βιβλία, ελέγχοντας πάντα η απόσταση Jensen-Shannon να έχει τιμή μικρότερη του 0.35.

7.1.3 Σύσταση με εισαγωγή ενός ερωτήματος

Η τελευταία λειτουργία που παρέχει η εργασία είναι μία μηχανή αναζήτησης, όπου ο χρήστης μπορεί να εισάγει ένα ερώτημα(query), όπως “how to make a successful parenting” και να του προταθούν βιβλία σύμφωνα με τις λέξεις αυτής της αναζήτησης. Στην προκειμένη λειτουργία, δε συμμετέχει το μοντέλο LDA και διακρίνεται σε τρία τμήματα.

Το πρώτο τμήμα περιλαμβάνει τη χρήση της συνάρτησης *MatrixSimilarity*, που παρέχεται από τη βιβλιοθήκη *gensim* [65]. Ουσιαστικά, σε αυτή τη συνάρτηση δίνονται ως είσοδο το σύνολο των κειμένων σε μορφή BoW, καθώς επίσης και το λεξιλόγιο και αυτή επιστρέφει ένα αντικείμενο MatrixSimilarity, που είναι ένας πίνακας ευρετηρίου ομοιότητας.

Το δεύτερο τμήμα περιλαμβάνει τον καθαρισμό της εισαγωγής του χρήστη μέσω της διαδικασίας της προεπεξεργασίας της ενότητας 4.4, καθώς επίσης και η μετατροπή του κειμένου εισαγωγής σε μορφή BoW αναφορικά με το λεξιλογίο.

Χρησιμοποιώντας το αντικείμενο MatrixSimilarity και εισάγοντας το καθαρισμένο πλέον query του χρήστη, υπολογίζει και επιστρέφει την ομοιότητα αυτού που εισήγαγε ο χρήστης

με κάθε κείμενο από το σύνολο κειμένων. Την ομοιότητα την υπολογίζει εσωτερικά, χρησιμοποιώντας την ομοιότητα συνημιτόνου και επιστρέφει τιμές στο εύρος [0,1]. Αφού ταξινομηθεί ο πίνακας με τις ομοιότητες και γίνει η αντιστοίχιση με τα βιβλία, παρουσιάζονται τα 15 πιο όμοια βιβλία, κατά φθίνουσα σειρά ομοιότητας.

7.2 Έλεγχος αποτελεσματικότητας συστήματος σύστασης

Τέλος, θα εξεταστεί αν οι συστάσεις του συστήματος στις διαφορετικές λειτουργίες είναι επιτυχημένες με χρήση μερικών παραδειγμάτων.

7.2.1 Έλεγχος αποτελεσματικότητας λειτουργιών

Παραδείγματα πρώτης λειτουργίας

Βρίσκονται στούς Πίνακες 7.1, 7.2 και 7.3

Πίνακας 7.1: Έλεγχός λειτουργίας εισαγωγής τίτλου βιβλίου (1)

Τίτλος Βιβλίου Εισαγωγής
The Encyclopedia Of Christianity
Τίτλοι Βιβλίων Σύστασης
The Text Of The Earliest New Testament Greek Manuscripts: A Corrected, Enlarged Edition Of The Complete Text Of The Earliest New Testament Manuscripts
Luther'S Works, Volume 55: Index
The Navarre Bible: Saint John
Living The Troth (Our Troth, #2)
Commentary On Hebrews: Exegetical & Expository - Vol. 1 (Pb)

Πίνακας 7.2: Έλεγχός λειτουργίας εισαγωγής τίτλου βιβλίου (2)

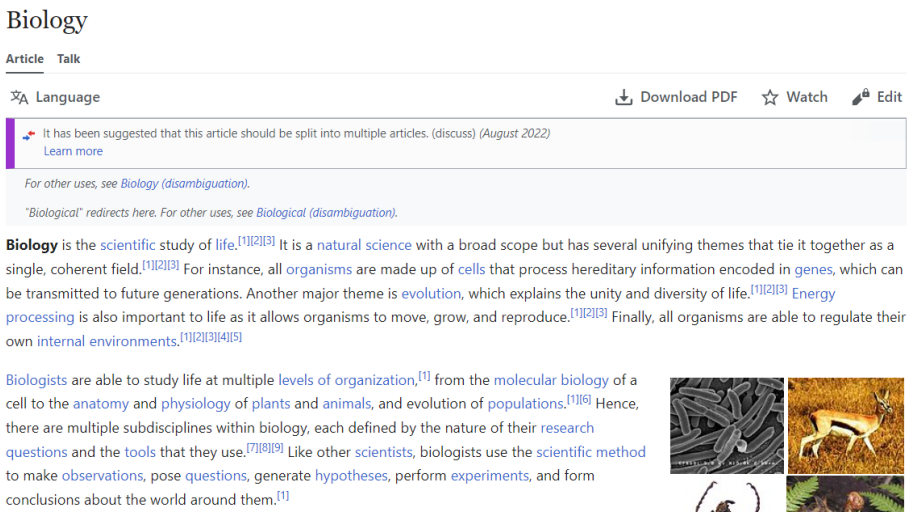
Εισαγωγή	
Τίτλος Βιβλίου Εισαγωγής	**Είδος**
The Fellowship Of The Ring (The Lord Of The Rings, #1)	Fantasy, Classics, Fiction, Adventure, High fantasy, Epic fantasy, Young adult, Novels
Συστάσεις	
Τίτλος Βιβλίου Σύστασης	**Είδος**
The Lord Of The Rings	Fantasy, Classics,Young adult, Fiction, Adventure, High fantasy, Epic fantasy
Death On Naboo (Star Wars: The Last Of The Jedi, #4)	Science fiction, Fiction, Young adult, Space, Fantasy, Children, Adventure
Beast Chooses Sides (X-Men: The Last Stand)	Science fiction, Children, literature, Fiction, Comics
Sachem'S Daughter (White Indian, #21)	Historical Fiction, Fiction
Splinter Of The Mind'S Eye	Science Fiction, Fiction, Fantasy, Novels Adventure

Πίνακας 7.3: Έλεγχός λειτουργίας εισαγωγής τίτλου βιβλίου (3)

Εισαγωγή	
Τίτλος Βιβλίου Εισαγωγής	**Είδος**
Saboteur (Star Wars: Darth Maul, #1)	Science fiction, Fiction, Fantasy, Adventure, Novels, Space
Συστάσεις	
Τίτλος Βιβλίου Σύστασης	**Είδος**
Dark Lord: The Rise Of Darth Vader	Science fiction, Fiction, Fantasy, Adventure, Novels, Space
The Way Of The Apprentice (Star Wars: Jedi Quest, #1)	Science fiction, Fiction, Fantasy, Young Adult, Novels, Space
Shadow Hunter (Star Wars: Darth Maul, #2)	Science fiction, Fiction, Fantasy, Adventure, Novels, Space
Heir To The Empire (Star Wars: The Thrawn Trilogy, #1)	Science fiction, Fiction, Fantasy, Adventure, Novels, Space
Death On Naboo (Star Wars: The Last Of The Jedi, #4)	Science fiction, Fiction, Young Adult, Fantasy, Adventure, War

Παραδείγματα δεύτερης λειτουργίας

Σε αυτήν την ενότητα θα αξιολογηθεί το πώς αντιδράει το σύστημα σε δεδομένα που δεν γνωρίζει και αν οι συστάσεις του εμφανίζουν κάποια συσχέτιση με το κείμενο εισαγωγής. Αυτό θα επιτευχθεί με τη μελέτη τριών παραδειγμάτων. Στο πρώτο παράδειγμα εισάγεται το κείμενο που απεικονίζεται στο Σχήμα 7.1. Πρόκειται για την επεξήγηση του όρου "Biology" από την ιστοσελίδα της Wikipedia.



Biology

Article Talk

Language Download PDF Watch Edit

It has been suggested that this article should be split into multiple articles. (discuss) *(August 2022)*
Learn more

For other uses, see Biology (disambiguation).

"Biological" redirects here. For other uses, see Biological (disambiguation).

Biology is the scientific study of life.[1][2][3] It is a natural science with a broad scope but has several unifying themes that tie it together as a single, coherent field.[1][2][3] For instance, all organisms are made up of cells that process hereditary information encoded in genes, which can be transmitted to future generations. Another major theme is evolution, which explains the unity and diversity of life.[1][2][3] Energy processing is also important to life as it allows organisms to move, grow, and reproduce.[1][2][3] Finally, all organisms are able to regulate their own internal environments.[1][2][3][4][5]

Biologists are able to study life at multiple levels of organization,[1] from the molecular biology of a cell to the anatomy and physiology of plants and animals, and evolution of populations.[1][6] Hence, there are multiple subdisciplines within biology, each defined by the nature of their research questions and the tools that they use.[7][8][9] Like other scientists, biologists use the scientific method to make observations, pose questions, generate hypotheses, perform experiments, and form conclusions about the world around them.[1]

Σχήμα 7.1: Κείμενο που τέθηκε ως είσοδος σχετικό με τη Βιολογία. [10]

Τα αποτελέσματα που επέφερε το σύστημα σύστασης αναγράφονται στον Πίνακα 7.4.

Πίνακας 7.4: Έλεγχός λειτουργίας εισαγωγής κειμένου περιγραφής (1)

Τίτλοι Βιβλίων Σύστασης
Non-Neutral Evolution: Theories And Molecular Data
Robustness And Evolvability In Living Systems
Symbiosis: Mechanisms And Model Systems (Cellular Origin, Life In Extreme Habitats And Astrobiology)
The Dynamics Of Arthopod Predator-Prey Systems. (Mpb-13), Volume 13
Algal Cultures, Analogues Of Blooms And Applications

Στο δεύτερο παράδειγμα εισάχθηκε το κείμενο περιγραφής του βιβλίου "The girl with the dragon tattoo", αφού ελέγχθηκε πως δεν υπάρχει στη βάση με τα βιβλία. Η περίληψη

πάρθηκε από την ιστοσελίδα goodreads και το κείμενο απεικονίζεται στο Σχήμα 7.2, ενώ οι συστάσεις στον Πίνακα 7.5 .

Millennium #1

The Girl with the Dragon Tattoo

Stieg Larsson, Reg Keeland (Translator)

★★★★★ 4.16 2,940,946 ratings · 73,005 reviews

Harriet Vanger, a scion of one of Sweden's wealthiest families disappeared over forty years ago. All these years later, her aged uncle continues to seek the truth. He hires Mikael Blomkvist, a crusading journalist recently trapped by a libel conviction, to investigate. He is aided by the pierced and tattooed punk prodigy Lisbeth Salander. Together they tap into a vein of unfathomable iniquity and astonishing corruption.

An international publishing sensation, Stieg Larsson's The Girl with the Dragon Tattoo combines murder mystery, family saga, love story, and financial intrigue into one satisfyingly complex and entertainingly atmospheric novel.

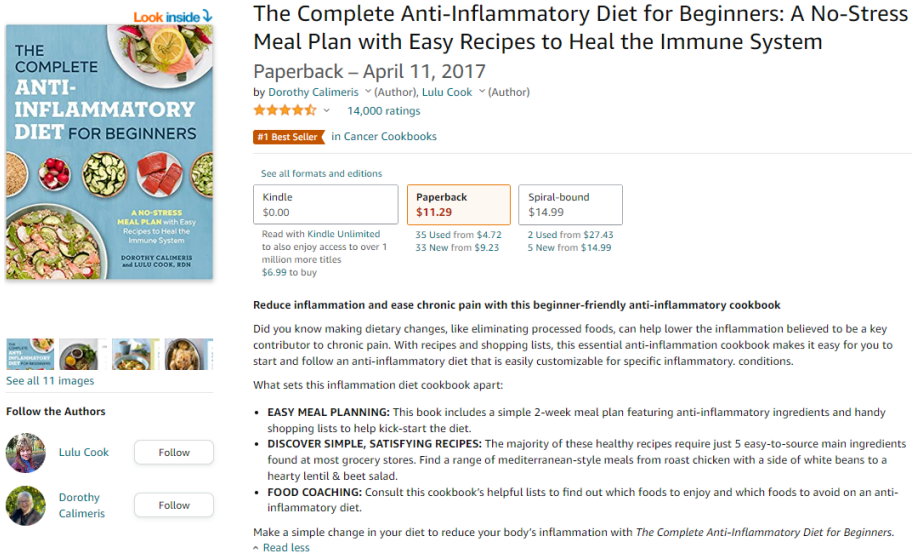
Σχήμα 7.2: Κείμενο περιγραφής βιβλίου "The girl with the dragon tattoo" [11]

Πίνακας 7.5: Έλεγχός λειτουργίας εισαγωγής κειμένου περιγραφής (2)

Εισαγωγή	
Τίτλος Βιβλίου Εισαγωγής	**Είδος**
The Girl With The Dragon Tattoo	Fiction, Mystery, Thriller, Crime,Suspense
Συστάσεις	
Τίτλος Βιβλίου Σύστασης	**Είδος**
Shades Of Grey: A Gaslight Gothic Mystery	Mystery, Suspense, Fiction, Romance novel
God Save The Mark	Mystery, Fiction, Humor, Crime, Thriller, Suspence
A Death In Vienna	Fiction, Thriller, Mystery, Espionage, Suspense, Spy thriller
Never Count Out The Dead	Crime, Fiction, Mystery, Thriller
The Meaning Of Night (The Meaning Of Night, #1)	Historical fiction, Fiction, Mystery, Thriller, Crime

Ενώ στο τρίτο παράδειγμα έγινε εισαγωγή του κειμένου περιγραφής του βιβλίου "The Complete Anti-Inflammatory Diet for Beginners: A No-Stress Meal Plan with Easy Recipes

to Heal the Immune System", και συγκεκριμένα το κείμενο απεικονίζεται στο Σχήμα 7.3, ενώ τα αποτελέσματα στον Πίνακα 7.6.



Σχήμα 7.3: Περιγραφή κειμένου του βιβλίου "The Complete Anti-Inflammatory Diet for Beginners: A No-Stress Meal Plan with Easy Recipes to Heal the Immune System" [12]

Πίνακας 7.6: Έλεγχός λειτουργίας εισαγωγής κειμένου περιγραφής (3)

Τίτλοι Βιβλίων Σύστασης
Cholesterol Control 3-Week Plan Handbook And Cookbook
Eat Well, Live Well With High Cholesterol: Low Cholesterol Recipes And Tips
The Antioxidant Save-Your-Life Cookbook: 150 Nutritious, High Fiber, Low-Fat Recipes To Protect You Against The Damaging Effects Of Free Radicals
Superfeast: Foods & Juices For Health & Healing
The T-Factor Fat Gram Counter

Παραδείγματα τρίτης λειτουργίας

Τέλος στους Πίνακες 7.7, 7.8 και 7.9 απεικονίζονται οι εισαγωγές και οι αποκρίσεις του συστήματος σε διαφορετικά ερωτήματα.

Πίνακας 7.7: Έλεγχος λειτουργίας εισαγωγής ερωτήματος (1)

Ερώτημα
How to make a successful parenting
Ονόματα Βιβλίων Σύστασης
The Ten Basic Principles Of Good Parenting
Parent Talk: 50 Quick, Effective Solutions To The Most Common Parenting Challenges
101 Ways To Tell Your Child “I Love You”
365 Positive Strategies For Single Parenting
Parenting For A Peaceful World

Πίνακας 7.8: Έλεγχος λειτουργίας εισαγωγής ερωτήματος (2)

Ερώτημα
How to improve the company’s strategy for better marketing
Ονόματα Βιβλίων Σύστασης
Total Global Strategy: Managing For Worldwide Competitive Advantage
Counterintuitive Marketing: Achieving Great Results Using Uncommon Sense
Key Marketing Skills: A Complete Action Kit Of Professional Marketing Concepts, Tools And Methods
Marketing Plans That Work
The Portable Mba In Marketing

Πίνακας 7.9: Έλεγχος λειτουργίας εισαγωγής ερωτήματος (3)

Ερώτημα
Literary mystery book where a murder case unfolds
Ονόματα Βιβλίων Σύστασης
Murder Through The Ages: A Bumper Anthology Of Historical Mysteries
Murders & Mysteries Of The North York Moors
Five-Minute Mysteries 3: Another 40 Cases Of Murder And Mayhem For You To Solve
The Mystery Of Edwin Drood
The Tokyo Zodiac Murders (#1)

7.2.2 Αξιολόγηση λειτουργιών

Η πρωτη λειτουργία σε γενικές γραμμές, ανταποκρίνεται σωστά, και συγκεκριμένα αντιλαμβάνεται το είδος του βιβλίου που εισέρχεται. Βέβαια, όπως παρατηρείται στον Πίνακα 7.2 εισάγοντας το “The Fellowship Of The Ring (The Lord Of The Rings, #1)” στις πέντε πρώτες επιλογές της σύστασης δεν υπάρχουν τα άλλα δύο βιβλία της τριλογίας του “Lord of the Rings”, το “The Tow Towers” και το “The Return Of The King”, αλλά βρίσκονται κατώτερα στη λίστα σύστασης. Αυτό είναι κάτι που δε θα έπρεπε να συμβαίνει αν αναλογιστεί κανείς ότι αυτά τα βιβλία στο κείμενο περιγραφής τους εμφανίζουν κοινές λέξεις που μπορεί να μην εμφανίζει κανένα άλλο κείμενο. Παρόλα αυτά μετά την επεξεργασία και το φιλτράρισμα οι κοινές λέξεις μεταξύ αυτών των κειμένων μπορεί να είναι πιο λίγες έναντι άλλων με αποτέλεσμα τα κείμενα να έχουν διαφορετική κατανομή ως προς αυτές κάτι που συνεπάγεται και διαφορετική κατανομή ως προς τις θεματικές ενότητες. Αυτό το πρόβλημα παρατηρήθηκε και σε άλλες περιπτώσεις, που βιβλία που ο άνθρωπος θα προσέδιδε ένα χαμηλό ποσοστό συσχέτισης, το μοντέλο να τα συστήνει έναντι άλλων.

Σημαντικό είναι να επισημανθεί, πως υπάρχουν περιπτώσεις που η λάθος, για την ανθρώπινη κρίση, σύσταση οφείλεται στην έντονη παρουσία μέσα στο κείμενο, θεματικών ενοτήτων που χαρακτηρίζονται από ασάφεια. Συγκεκριμένα, με την εισαγωγή ενός βιβλίου που περιέχει μία κατανομή, που κυριαρχούν σε υψηλό ποσοστό μη ευκρινείς θεματικές ενότητες, είναι πολύ πιθανόν τα βιβλία σύστασης να είναι μπερδεμένα και να μη φαίνεται λογική

η σύσταση.

Σχετικά με τη δεύτερη λειτουργία, τόσο από τα παραδείγματα στους Πίνακες 7.4, 7.5 και 7.6 όσο και από δοκιμές, παρατηρείται ένα ικανοποιητικό αποτέλεσμα. Συγκεκριμένα, το σύστημα σύστασης φαίνεται αποτελεσματικό ως προς τον τρόπο που γίνεται ανάθεση κατανομής θεματικών ενοτήτων στα ξένα δεδομένα που εισέρχονται στην εφαρμογή. Βέβαια, ισχύουν και εδώ τα προβλήματα που αναφέρθηκαν στην πρώτη λειτουργία.

Τέλος, στη σύσταση με βάση το ερώτημα και κατ' επέκτασιν με βάση τις λέξεις, επιφέρει θετικά αποτελέσματα, καθώς οι συστάσεις του μπορούν να κριθούν μόνο εύστοχες.

7.3 Παρουσίαση εφαρμογής

Τέλος σε αυτήν την ενότητα γίνεται μία παρουσίαση με στιγμιότυπα οθόνης του γραφικού περιβάλλοντος της εφαρμογής, που όπως ήδη αναφέρθηκε υλοποιήθηκε με χρήση του *tkinter* πακέτου της python.

Σχήμα 7.4: Αρχική σελίδα εφαρμογής

Στο Σχήμα 7.4 απεικονίζεται η αρχική σελίδα της εφαρμογής που ουσιαστικά αποτελείται από τέσσερα κουμπία όσες και οι λειτουργίες που μπορούν να εκτελεστούν.

Στη συνέχεια επιλέγοντας το κουμπί "Suggestions by book title", ο χρήστης κατευθύνεται προς τη σελίδα στην οποία μπορεί να εισάγει τον τίτλο του βιβλίου που θέλει να αναζητήσει ή

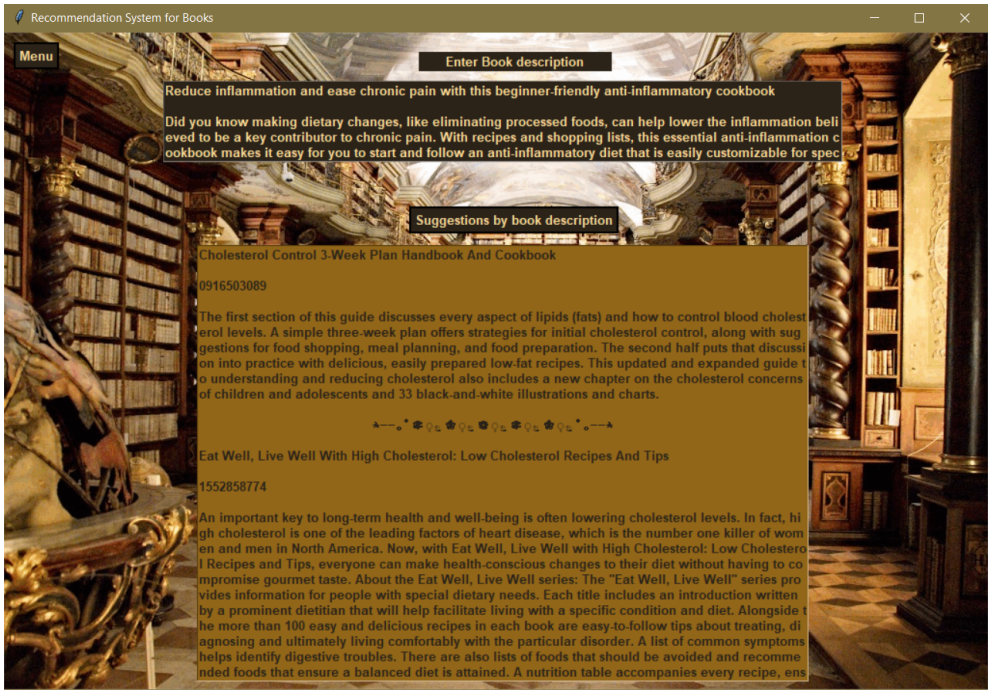
κάποιον παρόμοιο προκειμένου να εμφανιστούν τα αποτελέσματα. Όπως απεικονίζεται στο Σχήμα 7.5, υπάρχουν δύο πλαίσια στα οποία μπορούν να εμφανιστούν αποτελέσματα. Στο αριστερό παρουσιάζονται τα αποτελέσματα με βάση την ομοιότητα ως προς τις κατανομές του βιβλίου εισαγωγής και των υπόλοιπων βιβλίων, ενώ στο δεξί πλαίσιο εμφανίζονται οι προτάσεις που έχουν επέλθει από τους χρήστες. Σε ένα αντίστοιχο περιβάλλον κατευθύνεται ο χρήστης αν επιλέξει το κουμπί "Suggestions by book ISBN".

Σχήμα 7.5: Περιβάλλον λειτουργίας εισαγωγής τίτλου

Στη συνέχεια, πατώντας το κουμπί "Suggestions by book description", κατευθύνεται στη λειτουργία που ο χρήστης εισάγει ένα κείμενο περίληψης προκειμένου να του προταθούν βιβλία (βλπ. Σχήμα 7.6).

Τέλος, επιλέγοντας το "Suggestions by user query", εισέρχεται στην τελευταία λειτουργία στην οποία μπορεί να διατυπώσει κάποιο ερώτημα και να του συστηθούν βιβλία ανάλογα με αυτό που εισήγαγε (βλπ. Σχήμα 7.7).

Επίσης σε κάθε περιβάλλον υπάρχει το κουμπί "Menu", προκειμένου ο χρήστης να μπορεί να επιστρέφει στην αρχική σελίδα.



Σχήμα 7.6: Περιβάλλον λειτουργίας εισαγωγής περίληψης κειμένου

Σχήμα 7.7: Περιβάλλον λειτουργίας εισαγωγής ερωτήματος (query)

Κεφάλαιο 8

Επίλογος

Σε αυτό το κεφάλαιο θα πραγματοποιηθεί μία σύνοψη όσων διερευνήθηκαν στην παρούσα διπλωματική εργασία, καθώς επίσης θα πραγματοποιηθεί μία αξιολόγηση όσον αφορά το πόσο ικανοποιητικά επιτέλεσε το σκοπό της, βάσει των αποτελεσμάτων της. Τέλος, παρατίθενται κάποιες μελλοντικές επεκτάσεις και λύσεις σε προβλήματα που παρουσιάστηκαν.

8.1 Σύνοψη και συμπεράσματα

Σκοπός αυτής της εργασίας αποτέλεσε η δημιουργία ενός συστήματος σύστασης για βιβλία, βασισμένο στο κείμενο περιγραφής του βιβλίου. Η σύσταση αυτή πραγματοποιήθηκε με χρήση του μοντέλου Λανθάνουσας Κατανομής Dirichlet (LDA). Ο ρόλος αυτού του μοντέλου ήταν η ανακάλυψη θεματικών ενοτήτων, που είναι κρυμμένες μέσα στα κείμενα περιγραφής των βιβλίων σε συνδυασμό με τον τίτλο. Σημαντικό ήταν, πέρα από μία καλή προεπεξεργασία των δεδομένων, η δημιουργία ενός ικανοποιητικού μοντέλου. Αυτό επιτεύχθηκε μέσα από τη διαδικασία του συντονισμού παραμέτρων του μοντέλου, που τελικά προέκυψε πως σύμφωνα με τις μετρικές αξιολόγησης και την ανθρώπινη κρίση ήταν αυτό με τις παραμέτρους:

- **Αριθμός θεματικών ενοτήτων:** 50
- **Παράμετρος Dirichlet για την κατανομή κείμενα-θεματικές ενότητες:** asymmetric
- **Παράμετρος Dirichlet για την κατανομή θεματικές ενότητες-λέξεις:** 0.61

Με βάση, λοιπόν, τις θεματικές ενότητες που προέκυψαν από το πιο ικανοποιητικό μοντέλο και με χρήση της απόστασης Jensen-Shannon, υλοποιήθηκε το σύστημα σύστασης, το

οποίο στην πλειοψηφία των λειτουργιών του φαίνεται να δρα αρκετά ικανοποιητικά. Παρόλα αυτά, εξακολουθούν να υπάρχουν συστάσεις που δεν αξιολογούνται ως σωστές με βάση την ανθρώπινη κρίση. Αυτό όμως δεν αναιρεί το γεγονός ότι το μοντέλο λειτουργεί σωστά δεδομένου των λέξεων και την κατανομή τους στις θεματικές ενότητες. Επομένως, τόσο τα ενθαρρυντικά αποτελέσματα που προέκυψαν, όσο και οι αστοχίες αποτελούν έναυσμα για περαιτέρω έρευνα.

8.2 Μελλοντικές έρευνες και επεκτάσεις

Σύμφωνα με όσα ειπώθηκαν παραπάνω κάποιες μελλοντικές έρευνες που θα μπορούσαν να πραγματοποιηθούν προκειμένου να διορθωθούν κάποιες αστοχίες, που παρουσιάζονται στη σύσταση, είναι οι παρακάτω:

- Μια δοκιμή θα ήταν το μοντέλο να δέχεται ένα σύνολο κειμένων που δε θα είναι της μορφής BoW, αλλά η αναπαράσταση των κειμένων να γίνεται με χρήση της τεχνικής tf-idf διανυσματοποίησης. Συγκεκριμένα, μέσω αυτής της τεχνικής οι λέξεις στα κείμενα, δεν αποκτούν βάρος που εξαρτάται μόνο από τη συχνότητα εμφάνισης της λέξης στο εκάστοτε κείμενο, όπως στο BoW, αλλά εξαρτάται και από το πόσο συχνά εμφανίζεται αυτή η λέξη στο σύνολο κειμένων. Κατά αυτόν τον τρόπο τονίζονται περισσότερο οι λέξεις, που δεν είναι συχνά εμφανιζόμενες και ίσως έχουν περισσότερη σημασία. Αυτή η τεχνική βέβαια θα πρέπει να ρυθμιστεί και με τα φίλτρα που έχουν εφαρμοστεί κατά τη διάρκεια αυτής της διπλωματικής εργασίας προκειμένου να αφαιρεθούν οι συχνά εμφανιζόμενες λέξεις.

- Μελέτη του πώς επηρεάζει το φίλτρο συχνότητας των λέξεων την έκβαση του μοντέλου LDA.

- Υλοποίηση ενός συντονισμού όλων των παραμέτρων που είναι διαθέσιμες, για τη δημιουργία καλύτερων θεματικών ενοτήτων. Αυτό βέβαια απαιτεί και ισχυρά υπολογιστικά συστήματα για την ταχύτερη διεκπεραίωση του.

Στη συνέχεια, θα μπορούσε να ενισχυθεί και η εφαρμογή με ποικίλους τρόπους, όπως:

- Αξιοποίηση της βάσης με τους χρήστες προκειμένου να υλοποιηθεί σύστημα σύστασης συνεργατικού φιλτραρίσματος, καθώς επίσης και σύστημα σύστασης βασισμένο στο περιεχόμενο, που θα είναι προσωποποιημένο στον χρήστη.

- Ενίσχυση των βάσεων με τα βιβλία και τους χρήστες προκειμένου να υπάρχει περισσότερη πληροφορία διαθέσιμη, για την υλοποίηση του συστήματος σύστασης.

Βιβλιογραφία

[1] Figure 2.5: Content based filtering vs Collaborative filtering (... https://www.researchgate.net/figure/Content-based-filtering-vs-Collaborative-filtering-Source_fig5_323726564.

[2] A bag of words: Levels of language. https://sep.com/blog/a-bag-of-words-levels-of-language/. Ημερομηνία πρόσβασης: 24-8-2022.

[3] Multinomial distribution. https://kr.mathworks.com/help/stats/mnpdf.html. Ημερομηνία πρόσβασης: 24-8-2022.

[4] Dirichlet Distribution - ppt video online download. https://slideplayer.com/slide/9732893/. Ημερομηνία πρόσβασης: 24-8-2022.

[5] Latent dirichlet allocation. https://medium.com/codechef-vit/latent-dirichlet-allocation-812bb8099518. Ημερομηνία πρόσβασης: 24-8-2022.

[6] Figure 1. Graphical model of latent Dirichlet allocation (LDA).

[7] Neural networks. https://www.includehelp.com/python/one-hidden-layer-simplest-neural-network.aspx. Ημερομηνία πρόσβασης: 24-8-2022.

[8] arvindpdmn. Word2vec. https://devopedia.org/word2vec, October 2019. Ημερομηνία πρόσβασης: 24-8-2022.

[9] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 399–408, Shanghai China, February 2015. ACM.

[10] Biology. https://en.wikipedia.org/wiki/Biology. Ημερομηνία πρόσβασης: 24-8-2022.

[11] The girl with the dragon tattoo. https://www.goodreads.com/book/show/6081368-the-girl-with-the-dragon-tattoo. Ημερομηνία πρόσβασης: 24-8-2022.

[12] Meal plan. https://www.amazon.com/Complete-Anti-Inflammatory-Diet-Beginners-No-Stress/dp/1623159040. Ημερομηνία πρόσβασης: 24-8-2022.

[13] Konstantinos Christidis and Gregoris Mentzas. A topic-based recommender system for electronic marketplace platforms. *Expert Systems with Applications*, 40(11):4370–4379, September 2013.

[14] Rel Guzman Apaza, Elizabeth Vera Cervantes, Laura Cruz Quispe, and Jose Ochoa Luna. Online Courses Recommendation based on LDA. page 7, 2014.

[15] Zhiqiang He, Zhongyi Wu, Bochong Zhou, Lei Xu, and Weifeng Zhang. Tourist Routs Recommendation Based on Latent Dirichlet Allocation Model. In *2015 12th Web Information System and Application Conference (WISA)*, pages 201–206, September 2015.

[16] Dhiraj Vaibhav Bagul and Sunita Barve. A novel content-based recommendation approach based on LDA topic modeling for literature recommendation. In *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, pages 954–961, January 2021.

[17] Recommender system. https://en.wikipedia.org/w/index.php?title=Recommender_system&oldid=1106185544, August 2022. Page Version ID: 1106185544.

[18] Introduction to recommender systems. https://towardsdatascience.com/introduction-to-recommender-systems-6c66cf15ada. Ημερομηνία πρόσβασης: 24-8-2022.

[19] Collaborative filtering. https://en.wikipedia.org/w/index.php?title=Collaborative_filtering&oldid=1092795851, June 2022. Ημερομηνία πρόσβασης: 24-8-2022.

[20] Knowledge-based recommender system. https://en.wikipedia.org/wiki/Knowledge-based_recommender_system. Ημερομηνία πρόσβασης: 24-8-2022.

[21] Knowledge-based recommender systems: An overview. https://medium.com/@jwu2/knowledge-based-recommender-systems-an-overview-536b63721dba. Ημερομηνία πρόσβασης: 24-8-2022.

[22] Machine learning. https://www.ibm.com/cloud/learn/machine-learning. Ημερομηνία πρόσβασης: 24-8-2022.

[23] Supervised vs unsupervised learning explained. https://www.seldon.io/supervised-vs-unsupervised-learning-explained. Ημερομηνία πρόσβασης: 24-8-2022.

[24] Topic model. https://en.wikipedia.org/wiki/Topic_model. Ημερομηνία πρόσβασης: 24-8-2022.

[25] Topic modeling: An introduction. https://monkeylearn.com/blog/introduction-to-topic-modeling/. Ημερομηνία πρόσβασης: 24-8-2022.

[26] Jason Brownlee. A Gentle Introduction to the Bag-of-Words Model. https://machinelearningmastery.com/gentle-introduction-bag-words-model/, October 2017. Ημερομηνία πρόσβασης: 24-8-2022.

[27] Bayes' theorem. https://en.wikipedia.org/w/index.php?title=Bayes%27_theorem&oldid=1105964812, August 2022. Ημερομηνία πρόσβασης: 24-8-2022.

[28] S. Sinharay. Discrete Probability Distributions. In Penelope Peterson, Eva Baker, and Barry McGaw, editors, *International Encyclopedia of Education (Third Edition)*, pages 132–134. Elsevier, Oxford, January 2010.

[29] Multinomial distribution. https://en.wikipedia.org/wiki/Multinomial_distribution. Ημερομηνία πρόσβασης: 24-8-2022.

[30] Dirichlet distribution. https://en.wikipedia.org/wiki/Dirichlet_distribution. Ημερομηνία πρόσβασης: 24-8-2022.

[31] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3(null):993–1022, November 2003.

[32] Topic Modelling using LDA | Guide to Master NLP (Part 18). https://www.analyticsvidhya.com/blog/2021/06/part-18-step-by-step-guide-to-master-nlp-topic-modelling-using-lda-probabilistic-approach/, June 2021. Ημερομηνία πρόσβασης: 24-8-2022.

[33] Thushan Ganegedara. Intuitive Guide to Latent Dirichlet Allocation. https://towardsdatascience.com/light-on-math-machine-learning-intuitive-guide-to-latent-dirichlet-allocation-437c81220158, July 2022. Ημερομηνία πρόσβασης: 24-8-2022.

[34] Haaya Naushan. Topic Modeling with Latent Dirichlet Allocation. https://towardsdatascience.com/topic-modeling-with-latent-dirichlet-allocation-e7ff75290f8, December 2020. Ημερομηνία πρόσβασης: 24-8-2022.

[35] Ioana. Latent Dirichlet Allocation: Intuition, math, implementation and visualisation. https://towardsdatascience.com/latent-dirichlet-allocation-intuition-math-implementation-and-visualisation-63ccb616e094, September 2020. Ημερομηνία πρόσβασης: 24-8-2022.

[36] William M Darling. A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling. page 10.

[37] Thomas Griffiths and Mark Steyvers. Finding Scientific Topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101 Suppl 1:5228–35, April 2004.

[38] Jensen–shannon divergence. https://en.wikipedia.org/wiki/Jensen%E2%80%93Shannon_divergence. Ημερομηνία πρόσβασης: 24-8-2022.

[39] Cosine similarity. https://en.wikipedia.org/w/index.php?title=Cosine_similarity&oldid=1105991142, August 2022. Ημερομηνία πρόσβασης: 24-8-2022.

[40] Python (programming language). https://en.wikipedia.org/w/index.php?title=Python_(programming_language)&oldid=1104576011, August 2022. Ημερομηνία πρόσβασης: 24-8-2022.

[41] Anaconda | The World's Most Popular Data Science Platform. https://www.anaconda.com/. Ημερομηνία πρόσβασης: 24-8-2022.

[42] goodreads. https://www.goodreads.com/. Ημερομηνία πρόσβασης: 24-8-2022.

[43] Web Scraping With Python - Full Guide to Python Web Scraping. https://www.edureka.co/blog/web-scraping-with-python/, November 2018. Ημερομηνία πρόσβασης: 24-8-2022.

[44] A Guide to Text Preprocessing Techniques for NLP - Blog. https://exchange.scale.com/public/blogs/preprocessing-techniques-in-nlp-a-guide. Ημερομηνία πρόσβασης: 24-8-2022.

[45] Lemmatisation. https://en.wikipedia.org/w/index.php?title=Lemmatisation&oldid=1100992412, July 2022. Ημερομηνία πρόσβασης: 24-8-2022.

[46] Tanmoy Chakraborty, Santanu Pal, Tapabrata Mondal, Tanik Saikh, and Sivaju Bandyopadhyay. Shared Task System Description: Measuring the Compositionality of Bigrams using Statistical Methodologies. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, pages 38–42, Portland, Oregon, USA, March 2011. Association for Computational Linguistics. Ημερομηνία πρόσβασης: 24-8-2022.

[47] Dasaradh S. K. A Gentle Introduction To Math Behind Neural Networks. https://towardsdatascience.com/introduction-to-math-behind-neural-networks-e8b60dbbdeba, October 2020. Ημερομηνία πρόσβασης: 24-8-2022.

[48] Dhruvil Karani. Introduction to Word Embedding and Word2Vec. https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa, September 2020. Ημερομηνία πρόσβασης: 24-8-2022.

[49] Word2Vec For Word Embeddings -A Beginner's Guide. https://www.analyticsvidhya.com/blog/2021/07/word2vec-for-word-embeddings-a-beginners-guide/, July 2021. Ημερομηνία πρόσβασης: 24-8-2022.

[50] Gensim:implementation on github. https://github.com/RaRe-Technologies/gensim/tree/develop/gensim/topic_coherence. Ημερομηνία πρόσβασης: 24-8-2022.

[51] Gensim: Word2vec. https://radimrehurek.com/gensim/models/word2vec.html. Ημερομηνία πρόσβασης: 24-8-2022.

[52] Gensim: coherencemodel. https://radimrehurek.com/gensim/models/coherencemodel.html. Ημερομηνία πρόσβασης: 24-8-2022.

[53] Google Code Archive - Long-term storage for Google Code Project Hosting. https://code.google.com/archive/p/word2vec/. Ημερομηνία πρόσβασης: 24-8-2022.

[54] Sergey Nikolenko. Topic Quality Metrics Based on Distributed Word Representations. pages 1029–1032, July 2016.

[55] Derek O'Callaghan, Derek Greene, Joe Carthy, and Pádraig Cunningham. An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, 42(13):5645–5657, August 2015.

[56] Henri Trenquier. Improving Semantic Quality of Topic Models for Forensic Investigations. Technical report, University of Amsterdam, MSc System and Network Engineering, Amsterdam, Netherlands, Aug 2018.

[57] Enes Zvornicanin. When Coherence Score is Good or Bad in Topic Modeling? | Baeldung on Computer Science. https://www.baeldung.com/cs/topic-modeling-coherence-score, December 2021. Ημερομηνία πρόσβασης: 24-8-2022.

[58] Romain Deveaud, Eric Sanjuan, and Patrice Bellot. Accurate and Effective Latent Concept Modeling for Ad Hoc Information Retrieval. *Document Numérique*, pages 61–84, March 2014. Publisher: Lavoisier.

[59] Juan Cao, Tian Xia, Jintao Li, Yongdong Zhang, and Sheng Tang. A density-based method for adaptive LDA model selection. *Neurocomputing*, 72(7):1775–1781, March 2009.

[60] Gensim: topic modelling for humans. `https://radimrehurek.com/gensim/corpora/dictionary.html`. Ημερομηνία πρόσβασης: 24-8-2022.

[61] Gensim: ldamulticore. `https://radimrehurek.com/gensim/models/ldamulticore.html#module-gensim.models.ldamulticore`. Ημερομηνία πρόσβασης: 24-8-2022.

[62] Matthew D. Hoffman, David M. Blei, and Francis Bach. Online learning for Latent Dirichlet Allocation. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1*, NIPS'10, pages 856–864, Red Hook, NY, USA, September 2010. Curran Associates Inc.

[63] Carson Sievert and Kenneth Shirley. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 63–70, Baltimore, Maryland, USA, 2014. Association for Computational Linguistics.

[64] Udeshika Sewwandi. BoW vs TF-IDF in Information Retrieval. `https://medium.com/@sewwandikaus.13/bow-vs-tf-idf-in-information-retrieval-a325b5e61984`, May 2019. Ημερομηνία πρόσβασης: 24-8-2022.

[65] Gensim: Matrixsimilarity. `https://radimrehurek.com/gensim/similarities/docsim.html#gensim.similarities.docsim.MatrixSimilarity`. Ημερομηνία πρόσβασης: 24-8-2022.

Παράρτημα Α

Αρχεία Εφαρμογής

Το προγραμματιστικό κομμάτι της παρούσας διπλωματικής εργασίας, βρίσκεται αποθηκευμένο στην ιστοσελίδα του Github στον παρακάτω σύνδεσμο `https://github.com/Kourouniotou/RecommenderSystemForBooks_diplomaThesis`. Συγκεκριμένα, αποτελείται απο τους παρακάτω φακέλους:

- **App:** Σε αυτόν τον φάκελο υπάρχει το αρχείο *App.ipynb*, που είναι και το μοναδικό αρχείο που πρέπει να τρέξει κάποιος προκειμένου να του εμφανιστεί το παράθυρο της εφαρμογής.

- **BuiltWord2vecModel:** Περιέχει το αρχείο *Built word2vec model.ipynb*, στο οποίο βρίσκεται ο κώδικας υλοποίησης των μοντέλων Word2vec της παρούσας διπλωματικής εργασίας.

- **CombineDatasets_LanguageDetection_Selenium:** Σε αυτό το φάκελο είναι αποθηκευμένο το αρχείο *CombineDatasets_LanguageDetection_Selenium.ipynb*, το οποίο περιέχει όλη τη συνένωση των αρχικών δεδομένων, την αναγνώριση της γλώσσας, καθώς και τη διαδικασία Web Scraping

- **Lda_HyperparameterTuning:** Σε αυτόν τον φάκελο βρίσκεται ένα αρχείο *LDA on descriptions and titles and evaluation.ipynb* που περιέχει την ανάπτυξη του μοντέλου LDA πάνω στα δεδομένα της εργασίας, τη διαδικασία συντονισμού των παραμέτρων μαζί με τις μετρικές αξιολόγησης, την τελική απόφαση και κάποιες συναρτήσεις για οπτικοποίηση. Επίσης παρέχει τη διαδραστική οπτικοποίηση του τελικού μοντέλου στο αρχείο *lda.html*. Τέλος, υπάρχει και μία βάση δεδομένων με όνομα *tuningResults.csv* με τα αποτελέσματα απο το συντονισμό παραμέτρων.

- **LibrariesAndFunctions:** Στο αρχείο *LibrariesFunctions.ipynb* περιέχονται κάποιες συχνά χρησιμοποιούμενες συναρτήσεις που δημιουργήθηκαν για την επεξεργασία των δεδομένων, καθώς επίσης και οι πιο συχνά χρησιμοποιούμενες βιβλιοθήκες, για να μην επαναλαμβάνονται συνέχεια στα αρχεία.

- **Preprocessing:** Περιέχει όλα τα αρχεία που σχετίζονται με τη διαδικασία της προεπεξεργασίας των δεδομένων.

- **RecommenderSystemFunctions:** Περιέχει το αρχείο *RecommenderSystemFunctions .ipynb*, στο οποίο είναι υλοποιημένες όλες οι λειτουργίες σύστασης της εφαρμογής.

Στη συνέχεια, τα δεδομένα της εργασίας, καθώς και τα μοντέλα που υλοποιήθηκαν, συμπεριλαμβανομένου και του μοντέλου που αποφασίστηκε ως καλύτερο βρίσκονται ανεβασμένα στον σύνδεσμό `https://uthnoc-my.sharepoint.com/:f:/g/personal/kourouniotou_o365_uth_gr/ErXSiu9gFl1NnWfFUIAJYVEBWQUMD1AZ4GwbnXsKBuJVOw` και ο απαραίτητος, για την πρόσβαση, κωδικός είναι ο "RecommenderSystemForBooks-DiplomaThesis2022". Πιο λεπτομερώς, περιέχει τους παρακάτω φακέλους

- **Data:** Σε αυτό το φάκελο αρχικά είναι αποθηκευμένα τα δεδομένα βιβλίων *bookDataset.csv* και χρηστών *ratingsFinal* της εργασίας, με βάση τα οποία λειτουργεί η εφαρμογή. Επίσης περιέχονται βάσεις, οι οποίες περιέχουν τα bigrams και trigrams που δημιουργήθηκαν κατά την προεπεξεργασία των κειμένων. Στη συνέχεια, υπάρχει η βάση δεδομένων που περιέχει αγγλικά ονομάτα *englishNames.txt*. Τέλος, περιέχονται και τα δεδομένα των βιβλίων μετά το στάδιο της προεπεξεργασίας.

- **Data from Web Scraping:** Αυτός ο φάκελος περιέχει μία βάση, στην οποία αποθηκεύτηκαν οι περιλήψεις που πάρθηκαν από την ιστοσελίδα της goodreads, μέσω της διαδικασίας Web Scraping.

- **Initial Data:** Σε αυτόν τον φάκελο είναι αποθηκευμένες όλες οι αρχικές βάσεις βιβλίων και χρηστών, πριν την κατηγοριοποίησή τους σε μία βάση αποκλειστικά με χρήστες και τις βαθμολογήσεις τους σε διάφορα βιβλία και σε μία βάση με βιβλία και τις πληροφορίες τους.

- **Lda Models:** Αυτός ο φάκελος περιέχει όλα τα μοντέλα LDA, που αναπτύχθηκαν κατά τη διαδικασία συντονισμού παραμέτρων.

- **Word2Vec Models:** Τέλος, σε αυτόν τον φάκελο βρίσκονται όλα τα μοντέλα Word2vec, που χρησιμοποιήθηκαν για το σκοπό αυτής της εργασίας, καθώς και το αρχείο δημιουργίας τους.