



ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΥΛΟΠΟΙΗΣΗ ΕΝΟΣ CHATBOT ΓΙΑ ΤΗΝ ΥΠΟΣΤΗΡΙΞΗ ΕΝΟΣ
ΣΥΣΤΗΜΑΤΟΣ ΣΥΣΤΑΣΕΩΝ ΤΑΙΝΙΩΝ

ΕΥΘΥΜΙΟΠΟΥΛΟΣ ΔΗΜΗΤΡΙΟΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΥΠΕΥΘΥΝΟΣ

Κωνσταντίνος Κολομβάτσος

Επίκουρος Καθηγητής

Τμήματος Πληροφορικής & Τηλεπικοινωνιών Πανεπιστημίου Θεσσαλίας

Λαμία 2024



ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΥΛΟΠΟΙΗΣΗ ΕΝΟΣ CHATBOT ΓΙΑ ΤΗΝ ΥΠΟΣΤΗΡΙΞΗ ΕΝΟΣ
ΣΥΣΤΗΜΑΤΟΣ ΣΥΣΤΑΣΕΩΝ ΤΑΙΝΙΩΝ

ΕΥΘΥΜΙΟΠΟΥΛΟΣ ΔΗΜΗΤΡΙΟΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΥΠΕΥΘΥΝΟΣ

Κωνσταντίνος Κολομβάτσος

Επικουρος Καθηγητής

Τμήματος Πληροφορικής & Τηλεπικοινωνιών Πανεπιστημίου Θεσσαλίας

Λαμία 2024



UNIVERSITY OF
THESSALY

SCHOOL OF SCIENCE

DEPARTMENT OF COMPUTER SCIENCE & TELECOMMUNICATIONS

IMPLEMENTATION OF A CHATBOT TO SUPPORT A FILM
RECOMMENDATION SYSTEM

EFTHYMIOPOULOS DIMITRIOS

FINAL THESIS

ADVISOR

Konstantinos Kolomvatsos

Assistant Professor

Department of Informatics & Telecommunications, University of Thessaly

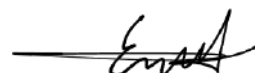
Lamia 2024

«Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις ⁽¹⁾, που προβλέπονται από της διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

1. Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί **χωρίς να τα περικλείω σε εισαγωγικά** και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάστηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.
2. Δέχομαι ότι η αυτολεξεί **παράθεση χωρίς εισαγωγικά**, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.
3. Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια
4. Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.

Ημερομηνία: 3/7/2024

Ο Δηλών



(1) «Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται ή αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση του άρθρου 8 παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων σκόπευε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκόπευε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ετών .»

Περίληψη

Στη σημερινή εποχή όπως γίνεται αντιληπτό η τεχνολογία εξελίσσεται ραγδαία, για τον λόγο αυτό η ανάγκη για αυτοματοποίηση διαδικασιών που αποτελούν κομμάτι της καθημερινότητας μας είναι ένα από τα ζητούμενα. Η τεχνητή νοημοσύνη και η μηχανική μάθηση έχουν πλέον φτάσει σε ένα επίπεδο που μας επιτρέπει να δημιουργήσουμε συστήματα που μπορούν να βελτιώσουν πολλές πτυχές της καθημερινότητας μας και της ζωής μας γενικότερα. Ένα από αυτά τα συστήματα είναι ένα σύστημα συστάσεων που χρησιμοποιείται για να προτείνει ταινίες και σειρές στους χρήστες με βάση τις προτιμήσεις και τη συμπεριφορά τους. Μεγάλες εταιρείες όπως η Netflix, η Amazon Prime Video και η Hulu έχουν ήδη κατασκευάσει τέτοια συστήματα στις πλατφόρμες τους, προσφέροντας στους χρήστες εξατομικευμένες προτάσεις ταινιών και σειρών.

Αυτή η εργασία έχει ως ζητούμενο την κατανόηση και ανάλυση του συστήματος συστάσεων ταινιών. Θα παρουσιαστούν και θα αναλυθούν αυτές οι εφαρμογές και οι κυριότεροι αλγόριθμοι που χρησιμοποιούνται για την υλοποίησή τους, όπως αλγόριθμοι συνεργατικού φιλτραρίσματος και αλγόριθμοι βασισμένοι στο περιεχόμενο. Θα αναλύσουμε όλες τις προκλήσεις που προκύπτουν τώρα αλλά και στο μέλλον από την ανάπτυξη αυτών των συστημάτων καθώς και τις τάσεις γύρω από τα συστήματα αυτά.

Στο πλαίσιο αυτής της εργασίας, αναπτύχθηκε ένα σύστημα συστάσεων ταινιών. Το σύστημα χρησιμοποιεί τεχνολογία μηχανικής μάθησης για να αναλύσει τις προτιμήσεις των χρηστών και να προτείνει ταινίες στον εκάστοτε χρήστη. Με αυτόν τον τρόπο, παρέχει μια βελτιωμένη, εξατομικευμένη εμπειρία προβολής που κάνει την εμπειρία των χρηστών στην πλατφόρμα καλύτερη και τους βοηθά να ανακαλύψουν νέο περιεχόμενο που ταιριάζει στα ενδιαφέροντά τους.

Με βάση τις σύγχρονες τεχνολογίες, το σύστημα συστάσεων που εφαρμόζεται στο πλαίσιο αυτής της εργασίας χρησιμοποιεί δεδομένα χρηστών όπως βαθμολογίες ταινιών, προτιμήσεις είδους και προηγούμενη συμπεριφορά προβολής, χρησιμοποιώντας τεχνολογίες όπως νευρωνικά δίκτυα και αλγόριθμους βαθιάς μάθησης. Το σύστημα που σχεδιάσαμε παρουσιάζει ιδιαίτερη ακρίβεια στις συστάσεις του. Ταυτόχρονα, η έμφαση δίνεται στη βελτίωση της εμπειρίας του χρήστη μέσω μιας φιλικής προς το χρήστη αλληλεπίδρασης και διαδραστικών χαρακτηριστικών. Οι χρήστες μπορούν να αξιολογήσουν και να βαθμολογήσουν τις συστάσεις, οι οποίες βοηθούν το σύστημα να βελτιώνεται συνεχώς και να προσαρμόζεται στις ανάγκες τους. Αυτή η συνεχής ανατροφοδότηση και προσαρμογή είναι το κλειδί για την επιτυχία και τη διατήρηση του συστήματος συστάσεων σε ένα συνεχώς μεταβαλλόμενο τοπίο προτιμήσεων των χρηστών.

Abstract

In today's world, as it is understood, technology is evolving rapidly, which is why the need for automation of processes that are part of our daily life is one of the demands. Artificial intelligence and machine learning have now reached a level that allows us to create systems that can improve many aspects of our daily lives and our lives in general. One such system is a recommendation system used to recommend movies and series to users based on their preferences and behaviour. Major companies such as Netflix, Amazon Prime Video and Hulu have already built such systems into their platforms, offering users personalized recommendations of movies and series.

This paper aims to understand and analyze the movie recommendation system. We will present and analyze these applications and the main algorithms used to implement them, such as collaborative filtering algorithms and content-based algorithms. We will analyze all the challenges that arise now and in the future from the development of these systems as well as the trends around these systems.

As part of this work, a film recommendation system was developed. The system uses machine learning technology to analyze user preferences and recommend movies to the individual user. In this way, it provides an improved, personalized viewing experience that makes users' experience on the platform better and helps them discover new content that matches their interests.

Based on state-of-the-art technologies, the recommendation system implemented in this work uses user data such as movie ratings, genre preferences and previous viewing behavior using technologies such as neural networks and deep learning algorithms. The system we have designed exhibits high accuracy in its recommendations. At the same time, emphasis is placed on improving the user experience through user-friendly interaction and interactive features. Users can evaluate and rate the recommendations, which helps the system to continuously improve and adapt to their needs. This constant feedback and adaptation is key to the success and maintenance of the recommendation system in an ever-changing landscape of user preferences

Πίνακας Περιεχομένων

Περιεχόμενα

Περίληψη	7
Abstract	8
Πίνακας Περιεχομένων	9
Κεφάλαιο 1 - Εισαγωγή	10
1.1 Σκοπός της διπλωματικής	10
1.2 Γιατί σύστημα πρότασης ταινιών	11
1.3 Δομή της Πτυχιακής	12
Κεφάλαιο 2 - Chatbots	13
2.1 Τι είναι τα Chatbots και πως χρησιμοποιούνται	13
2.2 Ιστορική αναδρομή	14
2.3. Chatbots που άλλαξαν τον κόσμο	15
2.4 Πως λειτουργούν τα chatbots	16
2.5 Τι θα μας φέρει το μέλλον των Chatbot	16
Κεφάλαιο 3 - Recommendation Systems	18
3.1 Τρόπος λειτουργίας	18
3.2 Τεχνική ανάλυση των Recommendation Systems	19
3.3 Αδυναμίες και βελτιώσεις	22
Κεφάλαιο 4 - Ανάπτυξη του συστήματος προ επεξεργασίας	23
4.1 Εισαγωγή	23
4.2 Προ επεξεργασία δεδομένων	23
4.3 Το σύνολο δεδομένων	25
4.4 Επεξεργασία Φυσικής Γλώσσας (NLP)	28
4.5 Μετατροπή των χαρακτήρων σε πεζούς (Lowercase)	29
4.6 Αφαίρεση Συμβόλων και Αριθμών	29
4.7 Αφαίρεση των τερματικών όρων (Stopwords)	29
4.8 Εφαρμογή αποκοπής καταλήξεων (Stemming)	31
4.9 Εφαρμογή λημματοποίησης (Lemmatization)	33
4.10 Term Frequency - Inverse Document Frequency (TF-IDF)	35
4.11 Διανυσματική αναπαράσταση λέξεων (Word Vectorization)	37
4.12 Global Vectors for Word Representation (GloVe)	40
4.13 Διανυσματική αναπαράσταση κειμένων (Document Vectorization)	45
4.13.1 Άθροισμα των επιμέρους διανυσμάτων	47
4.13.2 Μέσος Όρος των επιμέρους διανυσμάτων	47

4.14 Διανυσματική αναπαράσταση του αιτήματος του χρήστη (Query Vectorization).....	51
4.15 Μέθοδοι σύγκρισης απόστασης διανυσμάτων.....	53
4.15.1 Ευκλείδεια Απόσταση (Euclidean Distance)	53
4.15.2 Ομοιότητα Συνημιτόνων (Cosine Similarity).....	55
4.15.3 Απόσταση Manhattan (Manhattan Distance ή L1 Distance)	57
4.15.4 Ομοιότητα Jaccard (Jaccard Similarity).....	57
4.16 Επιλογή κατάλληλης μεθόδου υπολογισμού απόστασης διανυσμάτων	58
4.17 Αποθήκευση των προ επεξεργασμένων δεδομένων	60
4.18 Ανασκόπηση Προεπεξεργασίας Δεδομένων	60
Κεφάλαιο 5 - Ανάπτυξη Διαδικτυακής Εφαρμογής για Προτάσεις Ταινιών	63
5.1 Εισαγωγή.....	63
5.2 Περιγραφή λειτουργίας της εφαρμογής	63
5.3 Αρχιτεκτονική της εφαρμογής διαδικτύου.....	64
5.4 Ενοποίηση διαδοχικών αιτημάτων του χρήστη	67
Κεφάλαιο 6 - Συμπεράσματα.....	78
Κεφάλαιο 7 - Βελτιώσεις και μελλοντικές προσθήκες.....	79
Βιβλιογραφία	80

Κεφάλαιο 1 - Εισαγωγή

1.1 Σκοπός της διπλωματικής

Σε αυτή την εργασία συνθέσαμε ένα σύστημα πρότασης ταινιών μέσω προσωπικού βοηθού. Το σύστημα που υλοποιήσαμε μέσω του βοηθού θα συλλέγει τα δεδομένα του χρήστη σε συνδυασμό των ερωτήσεων και της ανατροφοδότησης του, θα μπορεί να κατανοήσει τις προτιμήσεις και τα ενδιαφέροντά του, ώστε να του προτείνει σειρές βάση των προτιμήσεων του. Τα συστήματα πρότασης ταινιών είναι ιδιαίτερα δημοφιλή στην εποχή μας και είναι σημαντικά για την εμπειρία του χρήστη στις διάφορες πλατφόρμες προβολής ταινιών. Για το λόγο αυτό υλοποιήσαμε ένα τέτοιο σύστημα το οποίο μαζί με τα chatbots θα παρουσιαστούν αναλυτικότερα παρακάτω.

1.2 Γιατί σύστημα πρότασης ταινιών

Επιλέχθηκε η υλοποίηση ενός συστήματος που προτείνει ταινίες γιατί είναι ένα δημοφιλές θέμα στον κόσμο των υπολογιστών και της τεχνολογίας. Αυτά τα συστήματα βοηθούν τους χρήστες να βρίσκουν ταινίες που θα τους αρέσουν και να βελτιώνουν την εμπειρία χρήσης των υπηρεσιών streaming.

Η τεχνολογία έχει διευκολύνει τους ανθρώπους να παρακολουθούν ταινίες και σειρές, αλλά υπάρχουν τόσες πολλές επιλογές που μπορεί να είναι δύσκολο να επιλέξουν. Ένα καλό σύστημα συστάσεων μπορεί να βοηθήσει προτείνοντας πράγματα που είναι πιθανό να αρέσουν στο χρήστη, πάντα βάση αυτών που έχει παρακολουθήσει στο παρελθόν ή έχει δηλώσει πως του αρέσουν.

Η δημιουργία αυτού του συστήματος δίνει την ευκαιρία να μάθουμε και να χρησιμοποιήσουμε προηγμένα προγράμματα υπολογιστών για να το κάνουμε καλύτερο. Κάνοντας αυτό, μπορούμε να αποκτήσουμε σημαντικές γνώσεις και δεξιότητες που θα μας βοηθήσουν στο μέλλον με την τεχνητή νοημοσύνη. Τα συστήματα προτάσεων είναι το μέλλον όπως φαίνεται καθημερινά σε όλες τις πτυχές της ζωής μας. Από την επιλογή του κατάλληλου κινητού μέχρι την επιλογή ταινίας. Ιδιαίτερα τα συστήματα επιλογής ταινιών είναι ένα εργαλείο που μπορεί να βελτιώσει την εμπειρία χρήσης σε μια συνήθεια της καθημερινότητας βοηθώντας τον κλάδο της ψυχαγωγίας ακόμα και αν με την πρώτη ματιά δεν φαίνεται ο σημαντικός ρόλος του.

Φτιάξαμε ένα σύστημα που θα βοηθήσει τους χρήστες να βρουν πιο εύκολα, γρήγορα και αποτελεσματικά την ταινία που θέλουν την στιγμή που θα την θέλουν. Έτσι, με το σύστημα πρότασης ταινιών η ψυχαγωγία μέσα από τις διάφορες πλατφόρμες θα γίνει ακόμα πιο διασκεδαστική και εύκολη καθώς ο κάθε χρήστης δεν θα αναλώνεται με τις ώρες στην αναζήτηση της σωστής επιλογής ανάμεσα στην πληθώρα επιλογών που συναντά κάθε μέρα.

1.3 Δομή της Πτυχιακής

Η διπλωματική υλοποιήθηκε και διαρθρώνεται σε 7 κεφάλαια μαζί με την εισαγωγή. Στα κεφάλαια 2 και 3 αναλύεται η έννοια καθώς και η ιστορική διαδρομή των chatbots και των συστημάτων πρότασης ενώ αναλύονται και οι μελλοντικές εξελίξεις των συγκεκριμένων κλάδων.

Στα επόμενα κεφάλαια παρουσιάζεται αναλυτικά η υλοποίηση του συστήματος και όλα τα εργαλεία που χρησιμοποιήθηκαν. Στα τελευταία δύο κεφάλαια θα παρουσιαστούν οι προτάσεις και οι μελλοντικές εξελίξεις. Αναλυτικότερα:

Στο κεφάλαιο 2 επίκεντρο είναι τα chatbots και η τεχνική παρουσίαση τους, η ιστορική αναδρομή μέχρι το σήμερα καθώς και τι περιμένουμε στο μέλλον από αυτά. Στην ίδια φιλοσοφία έχει γραφεί και το κεφάλαιο 3 παρουσιάζοντας τα αντίστοιχα στοιχεία γύρω από τα συστήματα προτάσεων.

Παράλληλα γίνεται και μια παρουσίαση της περίπτωσης του Amazon Prime Video μία από τις τελευταίες πλατφόρμες streaming.

Στο επόμενο κεφάλαιο ξεκινάει η τεχνική παρουσίαση του σχεδιασμού του συστήματος και το πως αυτό υλοποιήθηκε. Το κεφάλαιο 5 αποτελεί ουσιαστικά συνέχεια του κεφαλαίου 4 καθώς παρουσιάζονται λεπτομερώς οι αλγόριθμοι και πως αυτοί χρησιμοποιήθηκαν για το τελικό αποτέλεσμα.

Το έκτο και προτελευταίο κεφάλαιο είναι το σημείο όπου συνοψίζονται τα σημαντικά σημεία αυτής της διπλωματικής. Στο έβδομο και τελευταίο κεφάλαιο παρουσιάζονται οι προτάσεις μέσα από την εμπειρία που αποκομίσαμε από την υλοποίηση του συστήματος και δείχνουμε τι πρόκειται να δούμε στο μέλλον και τι να περιμένουμε από τα συστήματα αυτά και τον συγκεκριμένο κλάδο γενικότερα.

Κεφάλαιο 2 - Chatbots

2.1 Τι είναι τα Chatbots και πως χρησιμοποιούνται

Σύμφωνα με τον επίσημο ορισμό τα chatbot αποτελούν λογισμικά τα οποία μπορούν να μιμηθούν ανθρώπινες συζητήσεις με έναν άνθρωπο, στην περίπτωση μας τον χρήστη. Δεν είναι πάντα εργαλεία που λειτουργούν με την βοήθεια της τεχνητής νοημοσύνης όσο και αν αυτό μπορεί να προκαλέσει εντύπωση σε κάποιον. Βέβαια μετά την ραγδαία εξέλιξη της μηχανικής μάθησης και της τεχνητής νοημοσύνης γενικότερα, η πλειοψηφία των σύγχρονων chatbot εκμεταλλεύονται τις δυνατότητες της τεχνητής νοημοσύνης για να προσφέρουν ακόμα πιο αληθοφανείς συνομιλίες και πιο σωστές και επιτυχημένες συζητήσεις με τον κάθε χρήστη.

Ένα μεγάλο πλεονέκτημα τους είναι πως έχουν μεγάλο εύρος εφαρμογής. Μπορούν να χρησιμοποιηθούν σε εφαρμογές ανταλλαγής μηνυμάτων, σε ιστότοπους και σε τηλεφωνικές κλήσεις για να απαντήσουν σε ερωτήσεις ή να παρέχουν πληροφορίες. Τα chatbots έχουν σχεδιαστεί ώστε να προσαρμόζονται στις ανάγκες της κάθε συνομιλίας ξεχωριστά και αυτό γιατί μπορούν να κατανοούν και να μιλούν διαφορετικές γλώσσες.

Θεωρώ πως δεν γίνεται να αναφερθούμε στα chatbots χωρίς να παρουσιάσουμε το Turing test . Το τεστ αυτό ουσιαστικά μετρά την «νοημοσύνη» των εργαλείων μηχανικής νοημοσύνης και η λογική του τεστ είναι απλή. Υπάρχει ένας αξιολογητής ο οποίος παρατηρεί μια συνομιλία μεταξύ ανθρώπου και chatbot χωρίς να γνωρίζει ποιος είναι το chatbot. Αν παρακολουθώντας την συνομιλία δεν μπορεί να ξεχωρίσει ποιος από τους συνομιλητές είναι το chatbot τότε το chatbot είναι «ευφυές» και έχει επιτύχει στο τεστ. Αξίζει όμως να σημειωθεί πως κανένα chatbot δεν τα έχει καταφέρει μέχρι τώρα.

Αυτή τη στιγμή βρισκόμαστε στο μεταίχμιο της αλλαγής των chatbot καθώς με τη χρήση της τεχνητής νοημοσύνης οι δυνατότητες τους αυξάνονται και εξελίσσονται με μεγάλα και γρήγορα βήματα και χρησιμοποιούνται όλο και περισσότερο.

2.2 Ιστορική αναδρομή

Τα chatbots αν και μπορεί να μην το περίμενε κάποιος έχουν εμφανιστεί εδώ και 50 χρόνια, πολύ πριν γίνει γνωστή η τεχνητή νοημοσύνη και η μηχανική μάθηση. Μπορεί σήμερα να είναι μέρος της καθημερινότητας καθώς χρησιμοποιούνται ευρέως ιδίως από εταιρείες παροχής υπηρεσιών, δεν ήταν όμως πάντα έτσι.

Το 1966, ένας επιστήμονας έφτιαξε ένα πρόγραμμα υπολογιστή που ονομαζόταν ELIZA, το οποίο μπορούσε να μιλήσει στους ανθρώπους σαν πραγματικό άτομο. Προσποιήθηκε ότι ήταν ψυχίατρος κάνοντας ερωτήσεις και παρακολουθώντας τη συνομιλία. Η ELIZA χρησιμοποίησε ορισμένες λέξεις από αυτά που έλεγαν οι άνθρωποι για να βρει απαντήσεις, κάνοντάς το να φαίνεται ότι το κατανοούσε και απαντούσε. Αυτό έκανε τους ανθρώπους να αισθάνονται ότι μιλούσαν πραγματικά με κάποιον, έτσι μοιράστηκαν τα συναισθήματα και τις σκέψεις τους με το πρόγραμμα. Η επιτυχία του συγκεκριμένου chatbot ήταν τόσο μεγάλη που ο δημιουργός του δυσκολεύτηκε να πείσει τους χρήστες πως δεν υπήρχε άνθρωπος πίσω από την ELIZA και τις απαντήσεις της. Το chatbot αυτό έδειξε τον δρόμο και έθεσε τις βάσεις για την μετέπειτα εξέλιξη του συγκεκριμένου κλάδου.

Μετά το ELIZA, φτιάχτηκαν κάποια άλλα σημαντικά chatbot. Αυτά περιλαμβάνουν το έργο PARRY, το οποίο λειτουργεί σαν κάποιος με ψυχική ασθένεια, το A.L.I.C.E, το οποίο συνομιλεί στο διαδίκτυο και έχει κερδίσει βραβεία, το Siri της Apple, το οποίο βοηθά με εργασίες, το Cortana της Microsoft, που χρησιμοποιεί μηχανή αναζήτησης και το Amazon Alexa, που μπορεί να βοηθήσει με πολλά πράγματα όπως η αναζήτηση στο διαδίκτυο, η αναπαραγωγή μουσικής και ο έλεγχος συσκευών.

Δεν θα μπορούσαμε να μην αναφέρουμε και το ChatGPT το οποίο αποτελεί την πιο πρόσφατη και εξελιγμένη προσθήκη στην οικογένεια των chatbot, δημιουργώντας μια ολοκαίνουρια κατηγορία και εξελίσσοντας ακόμα περισσότερο τον τρόπο που λειτουργούσαν ως τώρα τα chatbots και το τι μπορούσαν να κάνουν.

Τα chatbots έχουν προχωρήσει πολύ από τότε που πρώτο δημιουργήθηκαν. Έχουν γίνει πολύ καλύτερα στο να μιλούν με ανθρώπους και να κάνουν διαφορετικά πράγματα. Τώρα, μπορούν να βοηθήσουν σε πολλές εργασίες, όπως η αγορά πραγμάτων, η απάντηση σε ερωτήσεις και η κράτηση εισιτηρίων. Ένα ακόμα ιδιαίτερα σημαντικό χαρακτηριστικό των πιο εξελιγμένων chatbot είναι η παραγωγή δικού τους περιεχομένου βάση των δεδομένων που θα τους παρέχει ο συνομιλητής τους, όπως στο ChatGPT που αναφέρθηκε και παραπάνω.

2.3. Chatbots που άλλαξαν τον κόσμο

Όπως είναι εύκολα αντιληπτό μέσα στα χρόνια εμφανίστηκαν πολλά και διαφορετικά chatbot, μερικά από αυτά όμως αποτέλεσαν σταθμούς στην εξέλιξη του κλάδου τους και γι' αυτό και θα αναφέρουμε παρακάτω μερικά από αυτά:

Furby – 1998. Αν και δεν είναι το ίδιο ένα chatbot, αυτό το δημοφιλές παιχνίδι της δεκαετίας του '90 κατέχει σημαντικό ρόλο στην ιστορία των chatbot. Το Furby εισήγαγε στο ευρύ κοινό την έννοια της συνομιλίας με μηχανές.

Όπως και τα chatbots που αναπτύσσονταν εκείνη την εποχή, ο Furby ήταν σε θέση να δημιουργήσει την ψευδαίσθηση της συνομιλίας και της νοημοσύνης. Ως αποτέλεσμα, ο γούνινος σύντροφος ήταν μια φιλική, μη απειλητική εισαγωγή στη μορφή αλληλεπίδρασης ανθρώπου-μηχανής που απαιτούν τα chatbots.

Η **Mitsuku** είναι ένα chatbot - σχεδιασμένο για διασκέδαση και ψυχαγωγία. Η ανάπτυξή της ξεκίνησε το 2005, πολύ πριν από μερικούς από τους άλλους χαρακτήρες αυτής της συλλογής. Η Mitsuku γνωρίζει σταθερή επιτυχία από τότε που κέρδισε για πρώτη φορά το βραβείο Leobner το 2013. Μετά τη βράβευση της, αυτή και ο δημιουργός της, ο Steve Worswick, έχουν κερδίσει το βραβείο άλλες τρεις φορές, το 2016, το 2017 και το 2018.

Επόμενα στη λίστα δεν θα μπορούσαν να είναι άλλα από τις φωνές που άλλαξαν την ιστορία στη χρήση των κινητών και της ζωής μας ακόμα και μέσα στο ίδιο μας το σπίτι.

Η **Siri (2010)** και η **Alexa (2015)** αποτελούν παραδείγματα επιτυχημένων, φωνητικά ελεγχόμενων chatbots. Αντιπροσωπεύουν επίσης μια απομάκρυνση από τα παραδοσιακά, βασισμένα σε κείμενο bots.

Από την κυκλοφορία τους, και οι δύο φωνητικά ελεγχόμενοι βοηθοί απολαμβάνουν ευρεία δημοτικότητα. Η επιτυχία αυτή έχει τροφοδοτήσει την αποδοχή των φωνητικά ελεγχόμενων, συνομιλιακών αλληλεπιδράσεων.

Η **Siri** και η **Alexa** αντιπροσωπεύουν την τεχνολογία αναγνώρισης φωνής ως ένα σταθερό, εξαιρετικά επιτυχημένο μέσο αλληλεπίδρασης με τους βοηθούς μας **chatbot**.

Τέλος, το πιο πρόσφατο και πραγματικά ριζοσπαστικό chatbot που έχει ήδη αλλάξει τον τρόπο που χρησιμοποιούμε το ίντερνετ και τον κόσμο που ζούμε είναι το **ChatGPT**.

Το **ChatGPT** είναι ένα chatbot και ένας εικονικός βοηθός που αναπτύχθηκε από την OpenAI και ξεκίνησε στις 30 Νοεμβρίου 2022. Βασισμένο σε μεγάλα γλωσσικά μοντέλα (LLM), επιτρέπει στους χρήστες να βελτιώσουν και να κατευθύνουν μια συζήτηση προς την επιθυμητή διάρκεια, μορφή, στυλ, επίπεδο λεπτομέρειας και γλώσσα. Οι διαδοχικές προτροπές και απαντήσεις του χρήστη λαμβάνονται υπόψη σε κάθε στάδιο της συνομιλίας ως πλαίσιο.

Από την εμφάνιση του πρώτου chatbot ο κλάδος έχει εξελιχθεί και έχει αλλάξει τόσο πολύ που πιστεύω κανείς δεν θα μπορούσε να φανταστεί. Τα chatbot θα γίνονται ολοένα και περισσότερο μέρος της ζωής μας και ίσως αποτελέσουν και παράγοντα αλλαγής της καθημερινότητάς μας.

2.4 Πως λειτουργούν τα chatbots

Τα chatbots μπορούν να ομαδοποιηθούν σε τρεις τύπους: βασισμένα σε κανόνες, έξυπνα και βασισμένα σε τεχνητή νοημοσύνη. Τα chatbot που βασίζονται σε κανόνες ακολουθούν καθορισμένους κανόνες, τα έξυπνα chatbot χρησιμοποιούν μηχανική εκμάθηση για να κατανοήσουν τι θέλουν οι χρήστες και τα chatbot που βασίζονται σε ΑΙ συνδυάζουν και τις δύο προσεγγίσεις με επεξεργασία φυσικής γλώσσας. Οι τεχνικές που χρησιμοποιούνται εξαρτώνται από τον τύπο του chatbot και από το τι προορίζεται να κάνει.

Τα chatbots είναι σαν διαφορετικοί τύποι ρομπότ που μπορούν να μιλήσουν σε ανθρώπους. Μερικά ρομπότ ακολουθούν αυστηρούς κανόνες για το τι μπορούν να πουν, άλλα μπορούν να μάθουν και να κάνουν πιο φυσικές συζητήσεις και μερικά είναι πραγματικά έξυπνα και μπορούν να καταλάβουν και να αλλάξουν τον τρόπο που μιλούν στους ανθρώπους. Είναι σημαντικό να επιλέξετε το σωστό είδος ρομπότ για το τι θέλετε να κάνει και πώς θέλετε να ενεργεί.

Τα chatbots είναι διαφορετικοί τύποι ρομπότ που μιλούν σε ανθρώπους. Κάποιοι ακολουθούν αυστηρούς κανόνες και δίνουν επιλογές, ενώ άλλοι μαθαίνουν και κάνουν πιο φυσικές συζητήσεις. Τα πιο προηγμένα χρησιμοποιούν προηγμένη τεχνολογία για να εξατομικεύουν τις απαντήσεις. Ο τύπος του chatbot που χρησιμοποιείται εξαρτάται από το πόσο περίπλοκη πρέπει να είναι η συνομιλία.

2.5 Τι θα μας φέρει το μέλλον των Chatbot

Τα chatbots έχουν εξελιχθεί σημαντικά τα τελευταία χρόνια και φαίνεται να έχουν ακόμα μεγαλύτερα περιθώρια βελτίωσης και εξέλιξης, παρέχοντας αποτελέσματα που ξεπερνούν τις αρχικές προσδοκίες. Η επικοινωνία με αυτές τις εφαρμογές είναι κυρίως μέσω κειμένου ή προφορικού λόγου, αλλά υπάρχουν πλέον επιλογές για αλληλεπίδραση με πολλούς και διαφορετικούς τρόπους. Τα chatbots που προσφέρουν αυτού του είδους την αλληλεπίδραση είναι πιο εύελικτα, επιτρέποντας στους χρήστες να επιλέξουν τον τρόπο εισόδου που τους εξυπηρετεί περισσότερο. Αυτό ανοίγει νέες προοπτικές για τη βελτίωση της εμπειρίας χρήστη και την αποτελεσματικότητα των εφαρμογών.

Η επόμενη προοπτική είναι η grounded language learning, όπου τα chatbots μπορούν να κατανοήσουν τις εκφράσεις μέσω οπτικών δεδομένων. Μέσω της συνδυασμένης τεχνολογίας της όρασης των υπολογιστών και της συνομιλητικής ΑΙ, τα chatbots μπορούν να επικοινωνήσουν με τους χρήστες σχετικά με περιεχόμενο που αντιλαμβάνονται οπτικά. Αυτή η προσέγγιση ανοίγει νέους ορίζοντες για την ανάπτυξη εφαρμογών που θα μπορούν να ανταποκρίνονται στις ανάγκες των χρηστών με πιο φυσικό τρόπο.

Μια αξιοσημείωτη έρευνα που πραγματοποιήθηκε το 2016 αφορά το παιχνίδι "Μάντεψε τι;", όπου ο στόχος είναι να αναγνωριστεί ένα αντικείμενο σε μια εικόνα μέσω μιας σειράς ερωτήσεων. Το παιχνίδι περιλαμβάνει δύο συμμετέχοντες, έναν που γνωρίζει το αντικείμενο και έναν που κάνει τις ερωτήσεις για να περιορίσει τη λίστα των αντικειμένων. Και οι δύο συμμετέχοντες

εκπαιδεύτηκαν χρησιμοποιώντας νευρωνικά δίκτυα, με τον έναν να απαντάει σε ερωτήσεις ναι/όχι και τον άλλο να δημιουργεί ερωτήσεις και να προβλέπει το σωστό αντικείμενο. Αυτό επιτεύχθηκε με τη χρήση ενός ιεραρχικού αναδρομικού κωδικοποιητή-αποκωδικοποιητή (hierarchical recurrent encoder-decoder - HRED).

Με βάση αυτήν την έρευνα, το 2020 παρουσιάστηκε ένα πλαίσιο εργασίας για την εκμάθηση γλωσσών με χαρακτηριστικά (Grounded Language Learning with Attributes - GroLLA). Αυτό το πλαίσιο επιτρέπει την αναπαράσταση αντικειμένων με αφηρημένα και συγκεκριμένα χαρακτηριστικά, όπως "is_appliance" και "is_white", για να διευκολύνει την διάκρισή τους σε εικόνες. Παρόλο που τα αποτελέσματα δεν ήταν ιδιαίτερα ικανοποιητικά, οι συγγραφείς πρότειναν τη χρήση πιο προηγμένων μηχανισμών μάθησης που θα επιτρέπουν την κωδικοποίηση των χαρακτηριστικών με χρήση ενός co-grounding operator για τη συγχώνευση του κειμένου και της εικόνας.

Για να είναι ένα chatbot σε θέση να αλληλεπιδρά με συνεργατικό τρόπο σε μια συνομιλία, πρέπει να είναι σε θέση να αιτιολογεί τις δικές του ενέργειες και να κατανοεί τις ενέργειες του συνομιλητή του. Αυτό απαιτεί την κατανόηση του νοήματος των εκφράσεων του συνομιλητή και των προθέσεων του, καθώς και την παροχή μιας λογικής απάντησης. Αν και έχουν γίνει ορισμένες έρευνες σε αυτόν τον τομέα, όπως μια πρόσφατη μελέτη του Cohen το 2019, δεν έχει δοθεί ακόμα μεγάλη προσοχή σε αυτά τα ζητήματα, εκτός από μια συζήτηση μιας επιτροπής που εξέταζε προκλήσεις για συστήματα πολύτροπων διαλόγων.

Τα τελευταία χρόνια παρατηρούμε μια αυξανόμενη διάδοση έξυπνων συσκευών, με έξυπνα σπίτια, γραφεία και άλλα περιβάλλοντα που γίνονται πραγματικότητα. Οι συσκευές αυτές ελέγχονται συνήθως μέσω εφαρμογών, ενώ ορισμένες φορές μπορεί να υπάρχει ένας κεντρικός ελεγκτής που ενσωματώνει τις λειτουργίες. Μέχρι στιγμής, όμως, λίγες εφαρμογές επιτρέπουν την αλληλεπίδραση μέσω διαλόγου.

Μια προσέγγιση που παρουσίασε η Samsung Electronics είναι η εφαρμογή ενός συστήματος που απαντά σε αιτήματα χρηστών, όπως "άναψε το φως" ή "κλείσε την τηλεόραση σε 30 λεπτά", μέσω διαλόγου. Ο χρήστης μπορεί να διατυπώνει τα αιτήματά του με φυσική φρασεολογία, και το σύστημα εκτελεί τις αντίστοιχες ενέργειες. Αυτές οι διαδικασίες έχουν ενσωματωθεί στο Action Planner σε μορφή δέντρου εργασιών, όπου κάθε σχέδιο ορίζει μια συγκεκριμένη κατάσταση και τις ενέργειες που πρέπει να εκτελεστούν για αυτήν.

Ένας κεντρικός διαχειριστής συσκευών επικοινωνεί με τις υπόλοιπες συσκευές και αναθέτει τις εργασίες σε αυτές, με στόχο την ολοκλήρωση του σχεδίου. Επίσης, μπορεί να ενημερώνει τον χρήστη για την εξέλιξη των διαδικασιών, όπως ότι το κλιματιστικό αναμένεται να ενεργοποιηθεί σε 10 λεπτά. Αυτό παρουσιάζει ορισμένες δυσκολίες, καθώς το σύστημα πιθανόν να παρερμηνεύσει τα αιτήματα του χρήστη, όπως στο παράδειγμα που δίνεται.

Αυτές οι τεχνολογίες και οι προκλήσεις που προκύπτουν από αυτές είναι σημαντικές για την εξέλιξη των chatbots. Παρόλο που έχει γίνει αξιοσημείωτη πρόοδος, η δουλειά για τη δημιουργία ενός απόλυτου chatbot, όπως ο JARVIS του Iron Man, είναι ακόμα πολύ μπροστά. Ωστόσο, είναι βέβαιο ότι τα chatbots θα ενσωματωθούν σε όλους τους τομείς και θα γίνουν ακόμα πιο έξυπνα με την πάροδο του χρόνου, αντικαθιστώντας ανθρώπους σε επαναλαμβανόμενες εργασίες.

Κεφαλαίο 3 - Recommendation Systems

3.1 Τρόπος Λειτουργίας

Η ανάπτυξη του Παγκόσμιου Ιστού και η συσσώρευση πληροφοριών έχουν καταστήσει τα συστήματα συστάσεων ολοένα πιο σημαντικά για επιχειρήσεις και πελάτες. Αυτά τα συστήματα προτείνουν νέα προϊόντα και υπηρεσίες βασισμένα σε διάφορους παράγοντες, βοηθώντας τους χρήστες να βρουν αυτό που ψάχνουν και αυξάνοντας τις πωλήσεις για τις επιχειρήσεις.

Αυτά τα συστήματα λειτουργούν μέσω της επεξεργασίας των δεδομένων των χρηστών, των επιλογών τους και των προτιμήσεών τους. Βρίσκουν αντιστοιχίες μεταξύ του χρήστη και του αντικείμενου που εξετάζεται και προτείνουν συναφή προϊόντα ή υπηρεσίες.

Η υλοποίηση αυτών των συστημάτων είναι απαιτητική και απαιτεί συνδυασμό γνώσεων και δεξιοτήτων από διάφορους τομείς της επιστήμης των υπολογιστών.

Τα συστήματα συστάσεων ασχολούνται με τρεις βασικές οντότητες: τους χρήστες, τα αντικείμενα σύστασης και τις ενέργειες. Οι χρήστες είναι οι πελάτες που λαμβάνουν προτάσεις από το σύστημα. Οι πληροφορίες που συλλέγονται για τον κάθε χρήστη δημιουργούν ένα προφίλ, περιλαμβάνοντας το ιστορικό αναζήτησής του και τη συμπεριφορά του. Τα αντικείμενα είναι οι υπηρεσίες ή τα προϊόντα που προτείνονται, λαμβάνοντας υπόψη τα χαρακτηριστικά τους και τη συμπεριφορά των χρηστών για αυτά. Οι ενέργειες είναι οι δραστηριότητες που εκτελεί ο κάθε χρήστης για ένα συγκεκριμένο αντικείμενο, καταγράφονται και χρησιμοποιούνται για να αναδειχθεί το ενδιαφέρον του.

Χρήστες (Πελάτες):

Οι χρήστες αποτελούν τον πυρήνα του συστήματος συστάσεων. Αυτοί είναι οι άνθρωποι που χρησιμοποιούν την πλατφόρμα ή την εφαρμογή για την οποία αναπτύσσεται το σύστημα συστάσεων. Το σύστημα συλλέγει πληροφορίες σχετικά με τη συμπεριφορά των χρηστών, όπως τα αντικείμενα που αναζητούν, τα αντικείμενα που αγοράζουν, ή τις αξιολογήσεις που δίνουν σε αντικείμενα. Αυτές οι πληροφορίες χρησιμοποιούνται για τη δημιουργία ενός προφίλ χρήστη, το οποίο αναπαριστά τα ενδιαφέροντα, τις προτιμήσεις και τη συμπεριφορά του κάθε χρήστη.

Αντικείμενα Σύστασης:

Τα αντικείμενα είναι τα προϊόντα, οι υπηρεσίες ή οποιοδήποτε άλλο στοιχείο προτείνεται στους χρήστες από το σύστημα συστάσεων. Κάθε αντικείμενο έχει ένα σύνολο χαρακτηριστικών που το περιγράφουν. Αυτά τα χαρακτηριστικά μπορεί να είναι σταθερά (όπως η κατηγορία ή ο τύπος) ή μεταβαλλόμενα (όπως οι αξιολογήσεις ή οι κριτικές). Το σύστημα λαμβάνει υπόψη την αντίδραση των χρηστών σε κάθε αντικείμενο (θετική ή αρνητική) και τις διάφορες δραστηριότητες που σχετίζονται με αυτό (όπως ο χρόνος που αφιέρωσαν στην αναζήτησή του).

Ενέργειες:

Οι ενέργειες αναπαριστούν τις δραστηριότητες που εκτελούν οι χρήστες για ένα συγκεκριμένο αντικείμενο. Μπορεί να περιλαμβάνουν ενέργειες όπως η αναζήτηση, η αγορά, η αξιολόγηση ή η αποθήκευση ενός αντικειμένου. Αυτές οι ενέργειες καταγράφονται και αναλύονται από το σύστημα, προκειμένου να κατανοήσει τα ενδιαφέροντα του κάθε χρήστη και να προτείνει αντικείμενα που ταιριάζουν με αυτά.

3.2 Τεχνική ανάλυση των Recommendation Systems

Στη βάση του περιεχομένου, το σύστημα συστάσεων χρησιμοποιεί την ομοιότητα μεταξύ των αντικειμένων για να προτείνει στον χρήστη ένα αντικείμενο που ταιριάζει με τα ενδιαφέροντά του. Αυτό γίνεται μέσω της ανάλυσης των χαρακτηριστικών των αντικειμένων και του γενικού συνόλου αντικειμένων που έχουν ήδη βαθμολογηθεί θετικά από τον χρήστη, καθώς και του προφίλ του χρήστη. Αρχικά, δημιουργείται ένα προφίλ για κάθε αντικείμενο, με βάση τις αξιολογήσεις που έχει λάβει ή τις ενέργειες που έχει πραγματοποιήσει ο χρήστης, όπως παρακολούθηση ταινιών, αγορά προϊόντων, ακρόαση μουσικής κλπ. Έπειτα, δημιουργείται το προφίλ του χρήστη, το οποίο καταγράφει πληροφορίες όπως τα συχνά αγοραζόμενα προϊόντα, οι προτιμήσεις σε ταινίες ή μουσική, κλπ.

Μετά τη δημιουργία των προφίλ, πραγματοποιείται έλεγχος μεταξύ των αντικειμένων που δεν έχουν βαθμολογηθεί και του προφίλ του χρήστη για να υπολογιστεί η ομοιότητα μεταξύ τους. Αυτός ο έλεγχος μπορεί να βασίζεται σε απλούς αλγορίθμους, όπως η Ευκλείδεια απόσταση ή η συνημιτονοειδής ομοιότητα, ή σε πιο προηγμένους. Τέλος, αν η ομοιότητα υπερβαίνει ένα καθορισμένο όριο, το αντικείμενο προτείνεται στον χρήστη.

Οι δύο κύριοι τύποι συστημάτων σύστασης, δηλαδή οι βασισμένοι στο περιεχόμενο και οι με συνεργατικό φιλτράρισμα, μπορούν να χρησιμοποιηθούν ανεξάρτητα ή σε συνδυασμό, ανάλογα με τους στόχους, τους πελάτες, την πλατφόρμα και τα αντικείμενα που προτείνονται. Ωστόσο, συνήθως, οι καλύτερες επιδόσεις προέρχονται από υβριδικά μοντέλα που συνδυάζουν τις δύο προσεγγίσεις.

Το content-based σύστημα συστάσεων ακολουθεί ένα συγκεκριμένο σύνολο σταδίων για να προτείνει αντικείμενα σε έναν χρήστη. Ας εξετάσουμε πιο αναλυτικά αυτά τα στάδια:

1. **Ανάλυση Περιεχομένου:** Το πρώτο στάδιο είναι η ανάλυση του περιεχομένου των αντικειμένων. Εδώ εξάγονται και επιλέγονται οι πληροφορίες που θα χρησιμοποιηθούν για την

πρόταση των συστάσεων. Συνήθως, οι πληροφορίες αυτές προέρχονται από τα δεδομένα που υπάρχουν σε ιστοσελίδες ή άλλες πηγές και μπορεί να είναι σε μη δομημένη μορφή. Αυτές οι πληροφορίες πρέπει να υποστούν κάποια μορφοποίηση ώστε να γίνουν καταλληλότερες για επεξεργασία από το σύστημα.

2. Εκμάθηση Προφίλ Χρήστη: Σε αυτό το στάδιο, δημιουργείται το προφίλ του χρήστη. Μέσω της εξαγωγής πληροφοριών από το προηγούμενο βήμα, το σύστημα μαθαίνει το μοτίβο των αντικειμένων που ενδιαφέρουν τον χρήστη. Αυτό συνήθως γίνεται μέσω ενός μοντέλου μηχανικής μάθησης. Στη συνέχεια, τυχόν πληροφορίες που δεν παρέχουν αξία για την πρόταση των συστάσεων αφαιρούνται.

3. Δημιουργία Συστάσεων: Στο τρίτο στάδιο, γίνεται η πραγματική δημιουργία των συστάσεων. Εδώ, ο αλγόριθμος συγκρίνει τα προφίλ του χρήστη με τα χαρακτηριστικά των αντικειμένων και αντιστοιχεί το προφίλ του χρήστη με τα προτιμώμενα αντικείμενα. Τα αποτελέσματα αυτής της σύγκρισης μπορεί να είναι δυαδικά ή δεκαδικά. Κατά την προσθήκη ενός νέου αντικειμένου στην πλατφόρμα, ο αλγόριθμος υπολογίζει το βαθμό προτίμησης για κάθε χρήστη και προτείνει τα αντικείμενα σε φθίνουσα κατάταξη. Επίσης, παρακολουθεί τη δράση του χρήστη σχετικά με τις προτεινόμενες συστάσεις και, αν χρειαστεί, μπορεί να μετεκπαιδευτεί. Ορισμένες μέθοδοι που χρησιμοποιούνται για την εκπαίδευση αυτών των αλγορίθμων είναι η πιθανολογική μέθοδος, οι μέθοδοι Rocchio και Relevance Feedback, τα δέντρα αποφάσεων, ο κοντινότερος γείτονας και άλλοι αλγόριθμοι μηχανικής μάθησης.

Με αυτόν τον τρόπο, ένα σύστημα συστάσεων με βάση το περιεχόμενο λειτουργεί εξετάζοντας τις πληροφορίες των αντικειμένων και το προφίλ του χρήστη για να προτείνει αντικείμενα που πιθανόν να τον ενδιαφέρουν. Επιπλέον, το σύστημα παρακολουθεί τη συμπεριφορά του χρήστη και μπορεί να προσαρμόσει τις συστάσεις του ανάλογα με τις αλλαγές στις προτιμήσεις του.

Το συνεργατικό φιλτράρισμα (collaborative filtering) είναι μια τεχνική που εκμεταλλεύεται τις συστάσεις διαφορετικών ανθρώπων για να προτείνει αντικείμενα σε άλλους χρήστες. Αυτή η τεχνική βασίζεται στην υπόθεση ότι χρήστες με παρόμοια ενδιαφέροντα είναι πιθανό να αρέσουν και σε παρόμοια αντικείμενα. Υπάρχουν δύο κύριες προσεγγίσεις στο συνεργατικό φιλτράρισμα:

1. Προσέγγιση μνήμης (Memory-based Approach):

- Φιλτράρισμα με βάση τη γειτονιά (Neighborhood-based filtering):** Αυτός ο αλγόριθμος υπολογίζει την ομοιότητα μεταξύ χρηστών ή αντικειμένων και προβλέπει τις προτιμήσεις του χρήστη λαμβάνοντας υπόψη το βάρος των αξιολογήσεων των γειτόνων του.
- Φιλτράρισμα με βάση τον χρήστη (User-based filtering): Σε αυτή την προσέγγιση, εξετάζονται παρόμοιοι χρήστες. Δημιουργούνται ομάδες χρηστών και συγκρίνονται οι δραστηριότητές τους για να παραχθούν προτάσεις.

2. Προσέγγιση βάσει μοντέλου (Model-based Approach):

- Σε αυτή την προσέγγιση, χρησιμοποιούνται μοντέλα μηχανικής μάθησης για να προβλέψουν τις προτιμήσεις του χρήστη. Αυτά τα μοντέλα εκπαιδεύονται σε δεδομένα αξιολογήσεων και αντικειμένων ώστε να μπορούν να κάνουν προβλέψεις για τις προτιμήσεις των χρηστών.

Οι δύο προσεγγίσεις έχουν τα πλεονεκτήματά τους και επιλέγονται ανάλογα με τις ανάγκες και τα χαρακτηριστικά του συστήματος συστάσεων.

Τα υβριδικά συστήματα σύστασης προσπαθούν να συνδυάσουν τα πλεονεκτήματα της τεχνικής με βάση το περιεχόμενο και του συνεργατικού φιλτραρίσματος για να βελτιώσουν την ακρίβεια και την απόδοση των συστάσεων. Μέσω της συνδυασμένης χρήσης αυτών των τεχνικών, τα υβριδικά συστήματα μπορούν να αντιμετωπίσουν ορισμένα από τα προβλήματα που αντιμετωπίζουν τα μοντέλα κάθε μιας τεχνικής μόνα τους.

Η στρατηγική των υβριδικών συστημάτων είναι να χρησιμοποιούν την καλύτερη προσέγγιση για κάθε περίπτωση. Για παράδειγμα, αν ένας νέος χρήστης δεν έχει ακόμα ιστορικό αξιολογήσεων για να του κάνει συστάσεις με βάση τη συμπεριφορά άλλων χρηστών, το σύστημα μπορεί να ξεκινήσει πρώτα με μια σύσταση περιεχομένου. Στη συνέχεια, μπορεί να ενσωματώσει τις προτιμήσεις του χρήστη σε μια προσέγγιση συνεργατικού φιλτραρίσματος για να παράγει πιο προσαρμοσμένες συστάσεις.

Τα υβριδικά συστήματα συνήθως παρουσιάζουν καλύτερη απόδοση από τα μοντέλα που βασίζονται μόνο σε μία τεχνική, καθώς αποφεύγουν τα μειονεκτήματα και εκμεταλλεύονται τα πλεονεκτήματα κάθε προσέγγισης. Ωστόσο, παραμένουν προκλήσεις, όπως η διαχείριση της ποικιλίας των χρηστών και η αλλαγή στις προτιμήσεις τους που αντιμετωπίζουν τα συστήματα σύστασης. Με την ανάπτυξη πιο προηγμένων μοντέλων και τη χρήση πιο πλούσιων δεδομένων, οι υβριδικές προσεγγίσεις συνεχίζουν να βελτιώνονται για να παρέχουν ακόμα καλύτερες συστάσεις.

3.3 Αδυναμίες και βελτιώσεις

Οι προκλήσεις που αντιμετωπίζουν τα συστήματα συστάσεων είναι πράγματι πολλές και σημαντικές. Ορισμένες φορές τα δεδομένα υπάρχουν διάσπαρτα στο ίντερνετ και χωρίς ικανοποιητικά στοιχεία για την εξαγωγή συμπερασμάτων, αυτό αποτελεί ένα από τα κύρια προβλήματα, καθώς η έλλειψη αξιολογήσεων και πληροφοριών καθιστά δύσκολη την παραγωγή ακριβών συστάσεων. Η δυσκολία αυτή είναι ιδιαίτερα έντονη στο συνεργατικό φιλτράρισμα, όπου η αναλογία των αξιολογήσεων προς τον αριθμό των χρηστών και των αντικειμένων είναι συνήθως αραιή.

Η πρόκληση της αύξησης του όγκου δεδομένων είναι επίσης σημαντική, καθώς η υπερβολική πολυπλοκότητα του πίνακα χρήστη-αντικειμένου μπορεί να μειώσει την απόδοση των αλγορίθμων. Αυτό οφείλεται στο γεγονός ότι η επεξεργαστική ικανότητα των υπολογιστικών συστημάτων μπορεί να φτάσει σε όρια όταν πρέπει να διαχειριστούν τεράστια σύνολα δεδομένων.

Η "ψυχρή εκκίνηση" αναφέρεται στο πρόβλημα της παροχής συστάσεων σε νέους χρήστες που δεν έχουν ακόμα ιστορικό αξιολογήσεων. Αυτό μπορεί να αποτελέσει πρόκληση για τα συστήματα συστάσεων, αλλά οι προτεινόμενες λύσεις όπως η συλλογή πληροφοριών από διάφορες πηγές ή η χρήση δημογραφικών δεδομένων μπορούν να βοηθήσουν στην αντιμετώπισή του.

Τέλος, η προστασία της ιδιωτικότητας και η αντιμετώπιση των ψεύτικων ή κακόβουλων προφίλ είναι άλλες σημαντικές προκλήσεις για τα συστήματα συστάσεων. Οι τεχνικές για την ανίχνευση και την αντιμετώπιση αυτών των προβλημάτων είναι απαραίτητες για τη διατήρηση της αξιοπιστίας και της αποτελεσματικότητας των συστημάτων συστάσεων.

Κεφάλαιο 4 - Ανάπτυξη του συστήματος προ επεξεργασίας

4.1 Εισαγωγή

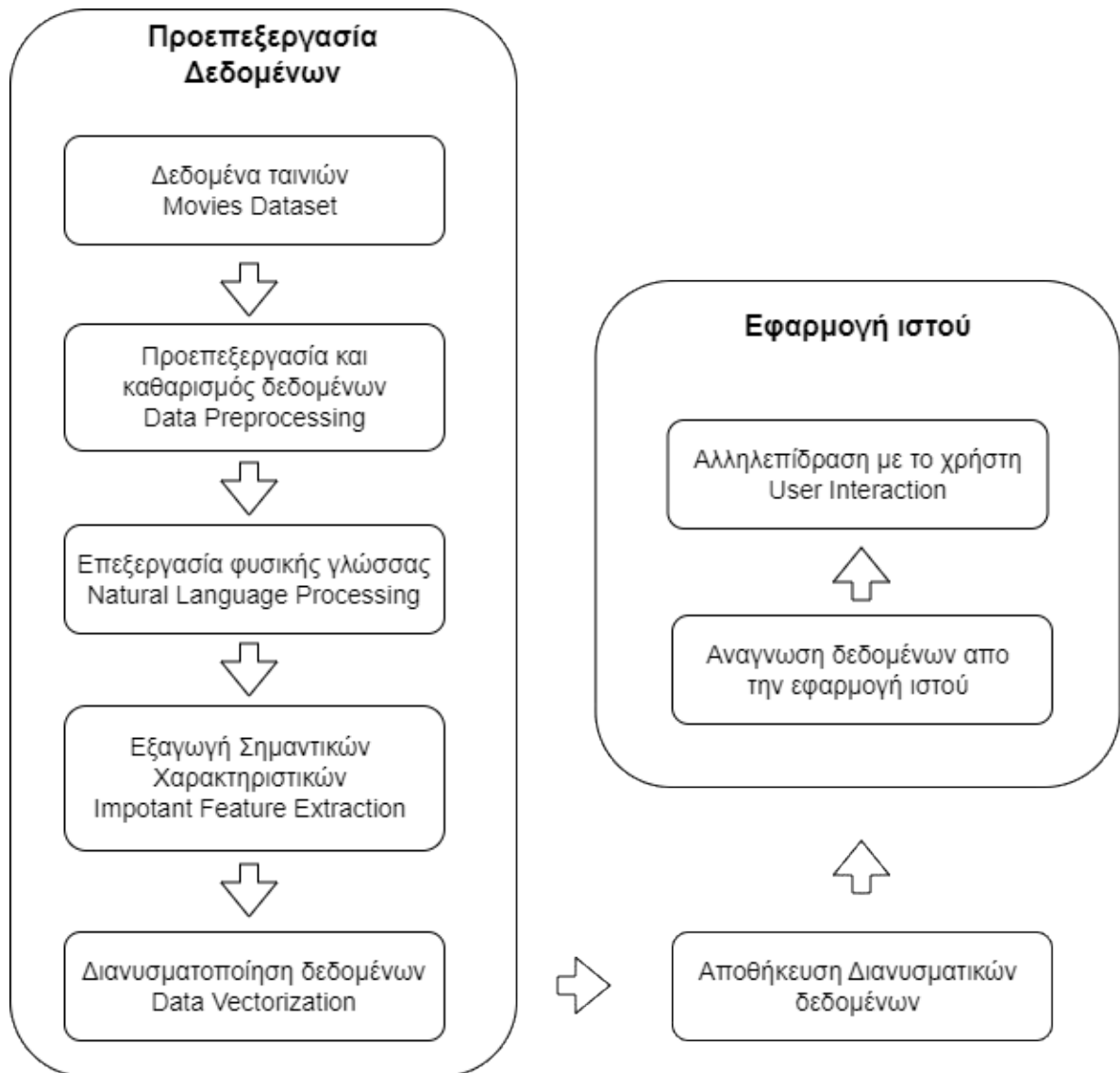
Σε αυτό το κεφάλαιο θα πραγματοποιηθεί μία αναλυτική περιγραφή που αφορά τόσο το σχεδιασμό όσο και την υλοποίηση του συστήματος πρότασης ταινιών που αναπτύχθηκε για τη συγκεκριμένη εργασία. Θα γίνει προσέγγιση και ανάλυση των βημάτων που πραγματοποιήθηκαν για την ανάπτυξη του συστήματος πρότασης ταινιών της εργασίας καθώς επίσης τους αλγορίθμους και τις μεθόδους που χρησιμοποιήθηκαν για την υλοποίηση του συστήματος. Τα βήματα αυτά μπορούν να διαχωριστούν σε δύο πιο βασικά μέρη τα οποία περιλαμβάνουν την προεπεξεργασία και το μετασχηματισμό των δεδομένων, καθώς επίσης και την ανάπτυξη της εφαρμογής με την οποία αλληλεπιδρά ο τελικός χρήστης.

Τα δεδομένα που αναφέρονται παραπάνω, αποτελούνται από μία μεγάλη συλλογή δεδομένων ταινιών, που περιλαμβάνουν στοιχεία όπως, τους τίτλους των ταινιών, την χρονολογία, τους σκηνοθέτες, τους ηθοποιούς, τις βαθμολογίες όπως επίσης και την περιγραφή του σεναρίου για κάθε ταινία που υπάρχει στη συγκεκριμένη συλλογή δεδομένων.

4.2 Προ επεξεργασία δεδομένων

Ως προ επεξεργασία δεδομένων ορίζεται η διαδικασία ανάλυσης και κατάλληλης μορφοποίησης των δεδομένων που αναφέρονται παραπάνω. Στόχος αυτού του βήματος είναι ο μετασχηματισμός και η μετατροπή των δεδομένων σε μια απλούστερη μορφή η οποία είναι πιο εύκολα διαχειρίσιμη και αξιοποιήσιμη από το σύστημα. Με άλλα λόγια, η προεπεξεργασία δεδομένων διευκολύνει τη χρήση των δεδομένων εισόδου από το σύστημα πρότασης ταινιών, καθιστώντας τα δεδομένα πιο εύκολα διαχειρίσιμα. Η προεπεξεργασία δεδομένων περιλαμβάνει μεθόδους όπως την αφαίρεση ελλιπών δεδομένων, τη μετατροπή κατηγορικών δεδομένων σε αριθμητικά, την επεξεργασία φυσικής γλώσσας, την εξαγωγή σημαντικών χαρακτηριστικών από τα δεδομένα καθώς και άλλες μεθόδους που θα παρουσιαστούν παρακάτω.

Η εφαρμογή που αναπτύσσεται στην εργασία είναι ένα Chatbot, το οποίο λειτουργεί ως διαδικτυακή εφαρμογή. Η συγκεκριμένη διαδικτυακή εφαρμογή χρησιμοποιεί τα δεδομένα των κινηματογραφικών ταινιών που περιγράφονται παραπάνω μετά την προεπεξεργασία τους. Στη συνέχεια επιτρέπει στον χρήστη να δώσει ως είσοδο σε μορφή φυσικής γλώσσας τι είδους ταινίες επιθυμεί, και το σύστημα πρότασης ταινιών του προτείνει τις κινηματογραφικές ταινίες που ταιριάζουν περισσότερο στο αίτημα που έδωσε ο χρήστης. Πιο απλά, με βάση το αίτημα του χρήστη, η εφαρμογή προτείνει κινηματογραφικές ταινίες από τα διαθέσιμα δεδομένα που έχει, οι οποίες ταιριάζουν περισσότερο στις προτιμήσεις του.



Σχ1. Στο παραπάνω σχήμα παρατηρούμε συνοπτικά τη δομή και τα επιμέρους βασικά τμήματα της εφαρμογής του συστήματος πρότασης κινηματογραφικών ταινιών. Διακρίνεται πως το σύστημα είναι χωρισμένο σε 2 κύρια τμήματα. Πρώτο τμήμα είναι το σύστημα της προ επεξεργασίας του συνόλου δεδομένων των κινηματογραφικών ταινιών, ενώ το δεύτερο αποτελείται από την εφαρμογή με την οποία αλληλεπιδρούν οι χρήστες.

4.3 Το σύνολο δεδομένων

Τα δεδομένα που χρησιμοποιούνται για το σύστημα προτάσεων αυτής της εργασίας προέρχονται από ένα σύνολο ανοιχτών δεδομένων το οποίο είναι διαθέσιμο στην ιστοσελίδα [Kaggle](#). Το σύνολο δεδομένων που χρησιμοποιούμε αντλείται από τις κινηματογραφικές ταινίες που υπάρχουν και είναι διαθέσιμες στην ιστοσελίδα [IMDB](#). Η ιστοσελίδα [IMDB](#) αποτελεί την πιο δημοφιλή και αξιόπιστη πηγή παγκοσμίως για πληροφορίες και περιεχόμενο κινηματογραφικών ταινιών. Ο σύνδεσμος από τον οποίο τα δεδομένα αυτά είναι προσπελάσιμα είναι ο ακόλουθος: [IMDB Movies Dataset Based on Genre](#)

Το σύνολο δεδομένων που χρησιμοποιείται στην εργασία περιέχει γενικές πληροφορίες για 368,297 κινηματογραφικές ταινίες, οι οποίες παρέχονται ταξινομημένες με βάση το είδος τους ή αλλιώς την κατηγορία τους.

Οι βασικές πληροφορίες που περιέχει το σύνολο δεδομένων που χρησιμοποιούμε για κάθε ταινία είναι οι ακόλουθες:

- **movie_id**: Αναγνωριστικό ID της κάθε ταινίας
- **movie_name**: Ο τίτλος της ταινίας
- **year**: Το έτος κυκλοφορίας της ταινίας
- **runtime**: Η διάρκεια της ταινίας σε λεπτά
- **genre**: Το είδος της ταινίας
- **rating**: Η μέση βαθμολογία της ταινίας
- **description**: Η περιγραφή του σεναρίου της ταινίας
- **director**: Ο σκηνοθέτης της ταινίας
- **star**: Οι κύριοι ηθοποιοί της ταινίας
- **votes**: Ο συνολικός αριθμός αξιολογήσεων που έχει λάβει η ταινία

Όλα τα πεδία του συνόλου δεδομένων που περιλαμβάνουν κάποιο κείμενο, όπως για παράδειγμα ο τίτλος (**movie_name**), η κατηγορία (**genre**) καθώς και η περιγραφή (**description**) κάθε κινηματογραφικής ταινίας παρέχεται στην αγγλική γλώσσα

Η μορφή του συνόλου δεδομένων παρουσιάζεται ενδεικτικά στις εικόνες παρακάτω:

	movie_id	movie_name	year	certificate	runtime	genre	rating	description
0	tt10640346	Babylon	2022	R	189 min	Comedy, Drama, History	7.4	A tale of outsized ambition and outrageous exc...
1	tt10304142	Amsterdam	2022	R	134 min	Comedy, Drama, History	6.1	In the 1930s, three friends witness a murder, ...
2	tt11138512	The Northman	2022	R	137 min	Action, Adventure, Drama	7.1	A young Viking prince is on a quest to avenge ...
3	tt9737876	Narvik: Hitler's First Defeat	2022	TV-14	108 min	Drama, History, War	6.6	April 1940. The eyes of the world are on Narvi...
4	tt8093700	The Woman King	2022	PG-13	135 min	Action, Drama, History	6.7	A historical epic inspired by true events that...
...	...	...	...	...	...	...	...	...
368295	tt26744289	Vermin	II	NaN	NaN	Horror	NaN	Kaleb, who's about to turn 30 and has never be...
368296	tt26744483	Epicuri	NaN	NaN	NaN	Horror	NaN	Add a Plot
368297	tt26745131	1910 Evil Awakens	2023	NaN	NaN	Horror	NaN	Ishaan and Kathy live in Germany and visit Ind...
368298	tt26745411	Death Walk	2003	NaN	9 min	Horror	NaN	Add a Plot
368299	tt26745752	Iravu	NaN	NaN	NaN	Horror, Thriller	NaN	Add a Plot

director	director_id	star	star_id	votes
Damien Chazelle	/name/nm3227090/	Brad Pitt, \nMargot Robbie, \nJean Smart, \nOL...	/name/nm0000093/,/name/nm3053338/,/name/nm0005...	66051.0
David O. Russell	/name/nm0751102/	Christian Bale, \nMargot Robbie, \nJohn David ...	/name/nm0000288/,/name/nm3053338/,/name/nm0913...	65277.0
Robert Eggers	/name/nm3211470/	Alexander Skarsgård, \nNicole Kidman, \nClaes ...	/name/nm0002907/,/name/nm0000173/,/name/nm0051...	217036.0
Erik Skjoldbjærg	/name/nm0804408/	Billy Campbell, \nKristine Hartgen, \nStig Hen...	/name/nm0001004/,/name/nm12242794/,/name/nm038...	9051.0
Gina Prince-Bythewood	/name/nm0697656/	Viola Davis, \nThuso Mbedu, \nLashana Lynch, \n...	/name/nm0205626/,/name/nm9096847/,/name/nm2736...	48483.0
...	...	...	...	...
Sébastien Vanicek	/name/nm7146773/	Finnegan Oldfield, \nSofia Lesaffre, \nJérôme ...	/name/nm2028613/,/name/nm5400953/,/name/nm5684...	NaN
Hugo Cobo	/name/nm7952838/	NaN	NaN	NaN
Iqbal B	/name/nm14569065/	Vani Vashishth, \nZeeshan Khan, \nAman S	/name/nm9200488/,/name/nm14569066/,/name/nm145...	NaN
Murat Erkok	/name/nm14569208/	Murat Erkok	/name/nm14569208/	NaN
Jagadeesan Subu	/name/nm9134023/	Vetri	/name/nm8904192/	NaN

Σχ2. Στις παραπάνω εικόνες παρουσιάζεται η μορφή των ακατέργαστων δεδομένων του συνόλου δεδομένων των κινηματογραφικών ταινιών που χρησιμοποιούνται στην συγκεκριμένη εργασία.

Η στήλη **movie_id** περιλαμβάνει το μοναδικό αναγνωριστικό κάθε ταινίας, το οποίο αποτελείται από το πρόθεμα "tt" ακολουθούμενο από ακέραιους αριθμούς. Τα δεδομένα είναι τύπου συμβολοσειράς.

Η στήλη **movie_name** περιλαμβάνει τον τίτλο κάθε ταινίας και είναι επίσης τύπου συμβολοσειράς. Το ίδιο ισχύει και για τη στήλη **description**, η οποία περιέχει μια σύντομη περιγραφή του σεναρίου της ταινίας.

Η στήλη **year** αναφέρεται στο έτος κυκλοφορίας της κάθε ταινίας και είναι τύπου συμβολοσειράς, ενώ η στήλη **run_time** περιλαμβάνει τη συνολική διάρκεια της ταινίας σε λεπτά και είναι τύπου ακέραιος.

Η στήλη **certificate** περιέχει το πιστοποιητικό καταλληλότητας της ταινίας και είναι επίσης τύπου συμβολοσειράς, ενώ η στήλη **genre** περιέχει το είδος της ταινίας και είναι επίσης τύπου συμβολοσειράς, με δεδομένα όπως Action, Drama, History.

Η στήλη **rating** περιέχει τη μέση βαθμολογία της ταινίας και είναι τύπου δεκαδικός αριθμός.

Η στήλη **director** περιλαμβάνει το όνομα του σκηνοθέτη της ταινίας και είναι τύπου συμβολοσειράς. Η στήλη **director_id** είναι το μοναδικό αναγνωριστικό του σκηνοθέτη στο IMDB και έχει μορφή /name/nmXXXXXXXXX/, όπου τα X αντιστοιχούν σε ακέραιους αριθμούς.

Η στήλη **star** περιέχει τα ονόματα των κύριων ηθοποιών της ταινίας και είναι τύπου συμβολοσειράς. Το **star_id** αναφέρεται στα μοναδικά αναγνωριστικά των ηθοποιών στο IMDB και έχει μορφή /name/nmXXXXXXXXX/, όπου τα X αντιστοιχούν σε ακέραιους αριθμούς.

Η στήλη **votes** περιέχει τον συνολικό αριθμό ψήφων που έχει λάβει η ταινία στην ιστοσελίδα του IMDB και είναι τύπου ακέραιος αριθμός.

Προεπεξεργασία και καθαρισμός του συνόλου δεδομένων

Το πρώτο βήμα για την υλοποίηση της εργασίας, έχει ως στόχο την προεπεξεργασία του συνόλου δεδομένων των ταινιών, ώστε να τα προετοιμάσει για περαιτέρω ανάλυση και χρήση στο σύστημα πρότασης ταινιών. Πιο συγκεκριμένα θα πρέπει να πραγματοποιηθεί τροποποίηση των δεδομένων με την αλλαγή του τύπου των επιμέρους στηλών, την αφαίρεση στηλών που δεν προσφέρουν χρήσιμη πληροφορία στο σύστημα όπως επίσης και διαχείριση κενών τιμών του συνόλου δεδομένων

Αρχικά, ως πρώτο βήμα για τον καθαρισμό των δεδομένων ορίζονται οι στήλες που πρέπει να αφαιρεθούν από το σύνολο δεδομένων, διότι δεν προσφέρουν κάποια χρήσιμη πληροφορία. Οι

στήλες αυτές είναι οι: **director_id**, **star_id** και **certificate**. Επιπλέον στήλες που δεν περιέχουν ουσιαστική πληροφορία για το σύστημα πρότασης ταινιών είναι οι στήλες: **rating**, **votes** και **run_time** ωστόσο δεν αφαιρούνται σε περίπτωση που ενδεχομένως χρησιμοποιηθούν μελλοντικά για λόγους ταξινόμησης των ταινιών στις δοθείσες προτάσεις του συστήματος.

Στο σύνολο δεδομένων παρατηρείται ότι στη στήλη της περιγραφής (description) που περιλαμβάνει την περιγραφή των ταινιών υπάρχουν κελιά που περιέχουν την τιμή "Add a Plot" σε περίπτωση που η περιγραφή δεν είναι διαθέσιμη. Επίσης στις στήλες director, star, και description παρατηρείται ότι υπάρχουν ειδικοί χαρακτήρες αλλαγής γραμμής (\n) καθώς και κενές τιμές (<NA>). Για το λόγο αυτό οι ειδικοί χαρακτήρες αφαιρούνται από τα συγκεκριμένα κελιά ενώ οι κενές τιμές αντικαθίστανται με την ειδική τιμή NaN.

Στη συνέχεια οι στήλες runtime και year μετατρέπονται από τιμές τύπου συμβολοσειράς σε αριθμητική μορφή για ευκολότερη ανάλυση και επεξεργασία των δεδομένων. Τέλος, από το σύνολο δεδομένων αφαιρούνται οι διπλότυπες εγγραφές με βάση την στήλη movie_id, εξασφαλίζοντας ότι κάθε μοναδική ταινία υπάρχει μόνο μία φορά στο σύνολο δεδομένων.

Με αυτόν τον τρόπο, τα δεδομένα του συνόλου μετατρέπονται σε κατάλληλη μορφή και προετοιμάζονται ώστε να χρησιμοποιηθούν μετέπειτα από το σύστημα πρότασης ταινιών, διασφαλίζοντας με τον τρόπο αυτό ότι τα δεδομένα είναι ομοιογενή και δεν περιέχουν μη χρήσιμες ή λανθασμένες πληροφορίες.

4.4 Επεξεργασία Φυσικής Γλώσσας (NLP)

Όπως αναφέρεται ήδη παραπάνω, αρκετά από τα δεδομένα που χρησιμοποιούμε για την ανάπτυξη του συστήματος είναι σε φυσική γλώσσα. Για το λόγο αυτό δεν είναι εύκολο να χρησιμοποιηθούν αυτούσια και άμεσα από το σύστημα πρότασης ταινιών. Τα δεδομένα σε φυσική γλώσσα περιέχουν αρκετές ασάφειες, παραλλαγές και περιττές πληροφορίες που καθιστούν δύσκολη την επεξεργασία τους και τη χρήση τους από το σύστημα που αναπτύσσεται. Για να είναι εφικτή η καλύτερη και πιο αποδοτική η ανάλυση και η αξιοποίησή των δεδομένων αυτών από το σύστημα πρότασης ταινιών, πρέπει πρώτα να εφαρμοστούν μέθοδοι επεξεργασίας φυσικής γλώσσας ή αλλιώς Natural Language Processing (NLP) στο συγκεκριμένο σύνολο δεδομένων.

Οι μέθοδοι Επεξεργασίας Φυσικής Γλώσσας (NLP) που περιγράφονται παραπάνω καθιστούν εφικτή την επεξεργασία των δεδομένων, μετατρέποντάς τα σε μια πιο συνεκτική και διαχειρίσιμη μορφή. Πιο συγκεκριμένα, για την επεξεργασία των δεδομένων χρησιμοποιούνται τεχνικές όπως την αφαίρεση τερματικών όρων (Stopwords), η λημματοποίηση (Lemmatization) καθώς και η τεχνική της αποκοπής καταλήξεων στις επιμέρους λέξεις (Stemming) με στόχο να πραγματοποιηθεί μετατροπή των λέξεων στη βασικότερη τους μορφή ή αλλιώς στη ρίζα τους.

Η εφαρμογή αυτών των μεθόδων της επεξεργασίας φυσικής γλώσσας (NLP) είναι απαραίτητη ώστε να μετατραπούν τα ακατέργαστα δεδομένα εισόδου σε μορφή φυσικής γλώσσας (Natural Language) του συνόλου δεδομένων σε χρήσιμη και αξιοποιήσιμη πληροφορία στο σύστημα πρότασης ταινιών που αναπτύσσεται στην εργασία.

4.5 Μετατροπή των χαρακτήρων σε πεζούς (Lowercase)

Η διαδικασία μετατροπής όλων των χαρακτήρων ενός κειμένου σε πεζά είναι ένα σημαντικό βήμα στην προ επεξεργασία φυσικής γλώσσας. Αυτή η μετατροπή βοηθά στη διατήρηση της ομοιομορφίας και στην αποφυγή της διάκρισης μεταξύ λέξεων που διαφέρουν μόνο ως προς το αν είναι γραμμένες με κεφαλαία ή πεζά γράμματα. Για παράδειγμα, οι λέξεις "Ταινία" και "ταινία" θεωρούνται ίδιες μετά τη μετατροπή τους σε πεζά γράμματα. Με αυτόν τον τρόπο, βελτιώνεται η ακρίβεια των αλγορίθμων ανάλυσης κειμένου, καθώς οι λέξεις αντιμετωπίζονται ομοιόμορφα ανεξάρτητα από τη μορφή των γραμμάτων τους.

4.6 Αφαίρεση Συμβόλων και Αριθμών

Η αφαίρεση των συμβόλων και των αριθμών από τα κείμενα είναι ένα ακόμη κρίσιμο βήμα στην προεπεξεργασία δεδομένων κειμένου. Τα σύμβολα και οι αριθμοί, όπως τα σημεία στίξης και οι αριθμητικές τιμές, συχνά δεν προσθέτουν ουσιαστική πληροφορία στο περιεχόμενο και μπορεί να προκαλέσουν θόρυβο κατά την ανάλυση. Η διαδικασία περιλαμβάνει την απομάκρυνση αυτών των χαρακτήρων για να διασφαλιστεί ότι το κείμενο περιέχει μόνο λέξεις που συμβάλλουν στη σημασία του. Αυτό βοηθά στη μείωση της πολυπλοκότητας του κειμένου και διευκολύνει τους αλγόριθμους μηχανικής μάθησης να εστιάσουν στις πραγματικά σημαντικές λέξεις.

4.7 Αφαίρεση των τερματικών όρων (Stopwords)

Η αφαίρεση των τερματικών όρων (Stopwords) σε δεδομένα φυσικής γλώσσας είναι μια σημαντική διαδικασία στην Επεξεργασία Φυσικής Γλώσσας (NLP), που χρησιμοποιείται για την προεπεξεργασία κειμένων. Οι τερματικοί όροι (Stopwords) είναι λέξεις που εμφανίζονται συχνά σε ένα κείμενο, αλλά δεν προσφέρουν σημαντική ή αξιοποιήσιμη πληροφορία για την κατανόηση του περιεχομένου του. Παραδείγματα τέτοιων λέξεων είναι τα άρθρα, οι προθέσεις και οι σύνδεσμοι, όπως "και", "είναι", "το", "σε", "από". Αυτές οι λέξεις είναι απαραίτητες για τη σύνταξη και τη γραμματική της γλώσσας, αλλά δεν συμβάλλουν ουσιαστικά πουθενά στην ανάλυση ή τη σημασιολογική κατανόηση του κειμένου.

Η αφαίρεση των τερματικών όρων (Stopwords) από ένα κείμενο βοηθά στη μείωση του θορύβου και της πολυπλοκότητας των δεδομένων, καθιστώντας τα δεδομένα πιο εύκολα διαχειρίσιμα και κατανοητά για το σύστημα πρότασης ταινιών. Χωρίς αυτήν τη διαδικασία, οι τερματικοί όροι (Stopwords) θα καταλάμβαναν πολύτιμο χώρο στη μνήμη και θα επιβάρυναν τους αλγόριθμους

ανάλυσης και επεξεργασίας των κειμένων, χωρίς να προσθέτουν κάποια ουσιαστική αξία στο τελικό αποτέλεσμα. Με την αφαίρεσή τους, το σύστημα μπορεί να επικεντρωθεί σε πιο σημαντικές λέξεις και φράσεις που καθορίζουν το νόημα και το περιεχόμενο του κειμένου όπως για παράδειγμα τον τίτλο και τη περιγραφή κάθε ταινίας στο σύνολο δεδομένων.

Η μέθοδος της αφαίρεσης των τερματικών όρων (Stopwords) πραγματοποιείται συνήθως με τη χρήση έτοιμων συνόλων που περιέχουν κοινά Stopwords που παρέχονται από εργαλεία NLP, όπως το Natural Language Toolkit (NLTK). Αυτές οι λίστες περιέχουν κοινούς τερματικούς όρους (Stopwords) για διάφορες γλώσσες, όπως για παράδειγμα Αγγλικά, Ελληνικά ή Γερμανικά. Στη συγκεκριμένη εργασία, η αφαίρεση των τερματικών όρων (Stopwords) είναι ένα από τα πρώτα βήματα στην προεπεξεργασία των δεδομένων, προκειμένου να διασφαλιστεί ότι το σύστημα προτάσεων ταινιών μπορεί να επεξεργαστεί και να αναλύσει τους τίτλους καθώς και τις περιγραφές με μεγαλύτερη ακρίβεια και αποτελεσματικότητα.

Αφαιρώντας τα stopwords, το σύστημα μπορεί να εξάγει πιο καθαρές και σαφείς πληροφορίες από το κείμενο, επιτρέποντας καλύτερη κατανόηση των προτιμήσεων και των αναγκών των χρηστών. Αυτό είναι κρίσιμο για την παροχή εξατομικευμένων προτάσεων ταινιών που ανταποκρίνονται στις συγκεκριμένες απαιτήσεις και προτιμήσεις του κάθε χρήστη. Με αυτόν τον τρόπο, η αφαίρεση των stopwords συμβάλλει σημαντικά στη βελτίωση της απόδοσης και της αποτελεσματικότητας του συστήματος προτάσεων ταινιών που αναπτύσσουμε στη συγκεκριμένη εργασία.

Αρχικό κείμενο

A computer is an electronic device that manipulates information, or data. It has the ability to store, retrieve, and process data. You may already know that you can use a computer to type documents, send email, play games, and browse the Web.

Αφαίρεση Stopwords



A, is, an, that, or, It, has, the, and, You, may, already, that, you, can, a, to, and, the

computer electronic device manipulates information, data. ability store, retrieve, process data. already know use computer type documents, send email, play games, browse Web.

Τελικό κείμενο

Σχ3. Παρουσιάζεται ένα ενδεικτικό παράδειγμα της διαδικασίας αφαίρεσης τερματικών όρων (Stopwords), από ένα κείμενο φυσικής γλώσσας. Παρατηρείται ότι πολύ κοινές λέξεις που δεν δίνουν ουσιαστικό νόημα στα δεδομένα, αφαιρούνται από το κείμενο μειώνοντας έτσι τον όγκο δεδομένων, χωρίς ωστόσο να χαθεί πολύτιμη πληροφορία.

4.8 Εφαρμογή αποκοπής καταλήξεων (Stemming)

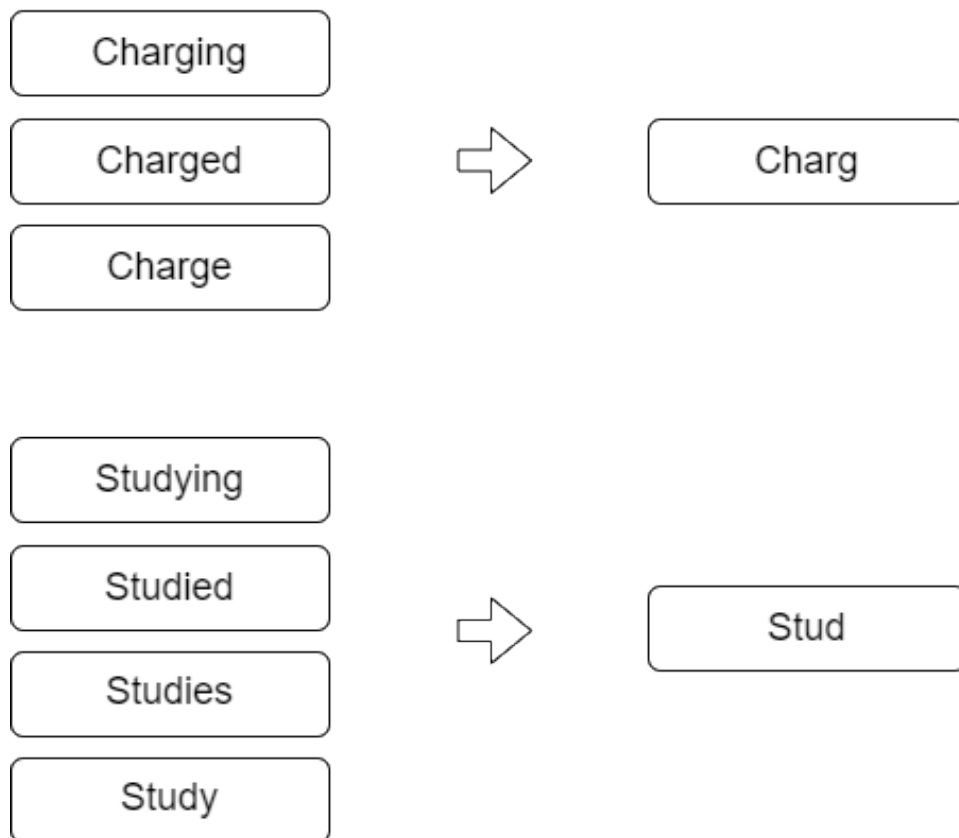
Η αποκοπή καταλήξεων ή αλλιώς ριζοποίηση (stemming) είναι μια τεχνική στον κλάδο της επεξεργασίας φυσικής γλώσσας (NLP) που χρησιμοποιείται κατά κόρον για την προεπεξεργασία κειμένων με στόχο την απλοποίηση και την κανονικοποίηση των λέξεων. Η διαδικασία της αποκοπής καταλήξεων περιλαμβάνει την αφαίρεση των επιθημάτων από τις λέξεις, προκειμένου να μειωθούν στις βασικές τους μορφές ή "ρίζες". Αυτή η διαδικασία δεν λαμβάνει υπόψη τη γραμματική ή τα συμφραζόμενα του κειμένου στο οποίο εφαρμόζεται η μέθοδος, και συνεπώς η παραγόμενη "ρίζα" μπορεί να μην είναι μια πραγματική λέξη, αλλά μια απλουστευμένη μορφή της.

Για παράδειγμα, οι λέξεις "running", "runner" και "runs" μπορεί να μετατραπούν όλες στη ρίζα "run" μέσω της διαδικασίας αποκοπής καταλήξεων. Αυτή η μέθοδος είναι ιδιαίτερα χρήσιμη όταν η ακριβής μορφή της λέξης δεν είναι τόσο σημαντική όσο το γενικό νόημα ή την έννοια η λέξη που εκπροσωπεί. Με την αποκοπή καταλήξεων, οι αλγόριθμοι επεξεργασίας μπορούν να μειώσουν τον αριθμό των μοναδικών λέξεων που χρειάζεται να αναλυθούν, βελτιώνοντας έτσι την απόδοση και την αποδοτικότητα των συστημάτων που χρησιμοποιούν φυσική γλώσσα για την λειτουργία τους.

Η μέθοδος της αποκοπής καταλήξεων είναι χρήσιμη σε συστήματα προτάσεων ταινιών, καθώς επιτρέπει στο σύστημα να αναγνωρίζει και να ομαδοποιεί περιγραφές και κριτικές που χρησιμοποιούν διαφορετικές μορφές της ίδιας λέξης. Αυτό διευκολύνει την ανάλυση των κειμένων και την εξαγωγή σχετικών πληροφοριών που μπορούν να χρησιμοποιηθούν για την παροχή εξατομικευμένων προτάσεων. Για παράδειγμα, αν ένας χρήστης αναζητήσει ταινίες που περιέχουν το πρόθεμα "αγαπ", το σύστημα μπορεί να αναγνωρίσει και να προτείνει ταινίες που περιέχουν λέξεις όπως "αγαπημένος", "αγάπη", "αγαπώντας" ή "αγάπησα".

Στην εργασία μας, η μέθοδος της αποκοπής καταλήξεων εφαρμόζεται ως μέρος της προ επεξεργασίας δεδομένων. Χρησιμοποιούμε εργαλεία όπως το Natural Language Toolkit (NLTK) για να πραγματοποιήσουμε τη συγκεκριμένη μέθοδο, εξασφαλίζοντας ότι τα δεδομένα είναι πιο ομοιογενή και εύκολα διαχειρίσιμα. Αυτή η διαδικασία συμβάλλει στη μείωση της πολυπλοκότητας των δεδομένων και επιτρέπει στο σύστημα προτάσεων να λειτουργεί πιο αποτελεσματικά. Αν και η ριζοποίηση μπορεί να μην είναι τόσο ακριβής όσο η μέθοδος της λημματοποίησης που θα περιγράψουμε παρακάτω, παραμένει ένα πολύτιμο εργαλείο για την προεπεξεργασία των δεδομένων.

Εφαρμογή αποκοπής καταλήξεων (Stemming)



Σχ4. Παρουσιάζεται ένα ενδεικτικό παράδειγμα της διαδικασίας αποκοπής ή αφαίρεσης καταλήξεων των λέξεων (Stemming), σε ένα κείμενο φυσικής γλώσσας. Παρατηρείται ότι οι καταλήξεις των λέξεων αφαιρούνται, διατηρώντας ένα κοινό πρόθεμα, δηλαδή τη βασική μορφή τους.

4.9 Εφαρμογή λημματοποίησης (Lemmatization)

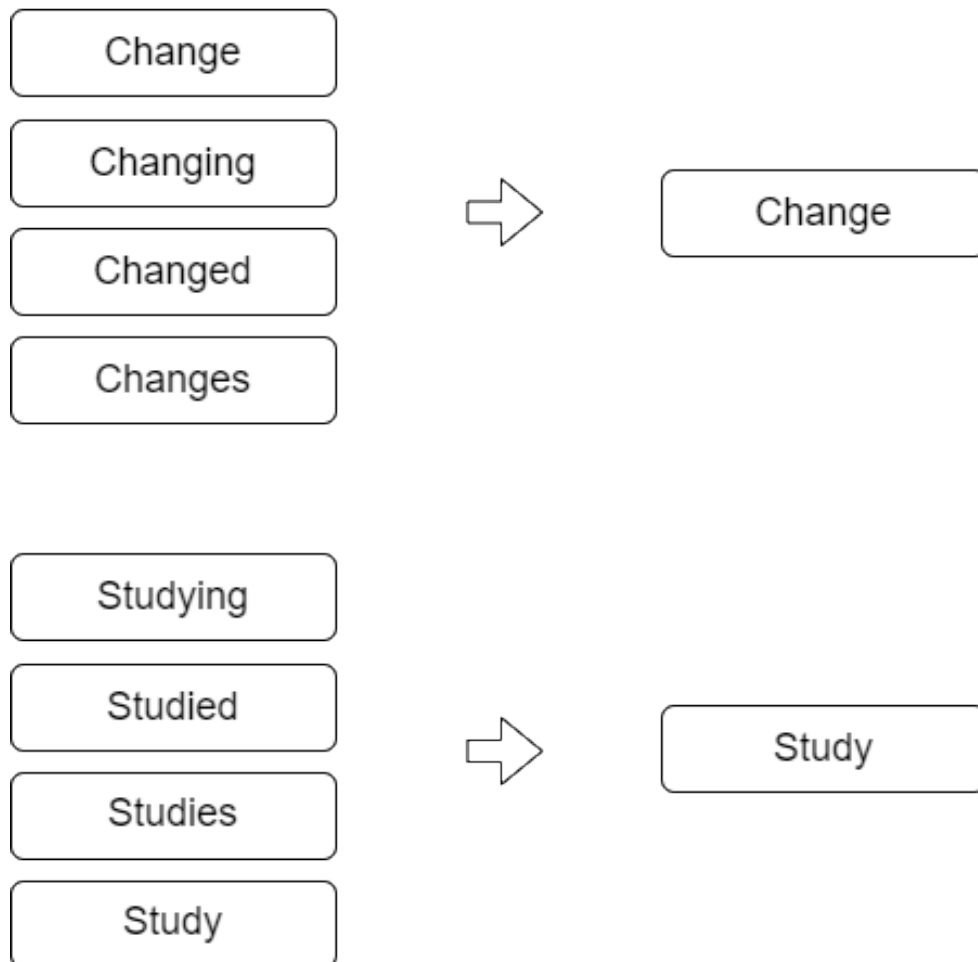
Η λημματοποίηση (lemmatization) είναι μια σημαντική μέθοδος στην επεξεργασία της φυσικής γλώσσας (NLP) που χρησιμοποιείται για την προεπεξεργασία και τον καθαρισμό κειμένων. Η διαδικασία της λημματοποίησης περιλαμβάνει τη μετατροπή των λέξεων στη βασική ή λεξική τους μορφή, γνωστή ως λήμμα. Σε αντίθεση με την πιο απλή διαδικασία της ριζοποίησης (stemming) που περιγράφουμε παραπάνω, η οποία απλώς αφαιρεί τα επιθήματα από τις λέξεις, η λημματοποίηση λαμβάνει υπόψη τη γραμματική και τα συμφραζόμενα για να εξασφαλίσει ότι η κάθε λέξη μετατρέπεται στη σωστή της μορφή.

Για παράδειγμα, οι λέξεις "running", "ran" και "runs" θα μετατραπούν στη βασική μορφή "run" μέσω της εφαρμογής της λημματοποίησης. Αυτή η διαδικασία βοηθά στην ενοποίηση των λέξεων που έχουν διαφορετικές μορφές ή εκδοχές, αλλά την ίδια σημασία, διευκολύνοντας έτσι την ανάλυση και την κατανόηση του κειμένου από το σύστημα. Με αυτόν τον τρόπο, οι αλγόριθμοι επεξεργασίας μπορούν να αναγνωρίσουν και να επεξεργαστούν τις λέξεις με μεγαλύτερη ακρίβεια, συμβάλλοντας στη μείωση της πολυπλοκότητας των δεδομένων και αυξάνοντας την αποτελεσματικότητα.

Η λημματοποίηση είναι ιδιαίτερα χρήσιμη σε συστήματα προτάσεων ταινιών, καθώς επιτρέπει στο σύστημα να αναγνωρίζει και να συγκρίνει περιγραφές και κριτικές ταινιών με μεγαλύτερη ακρίβεια. Με τη μετατροπή των λέξεων στη βασική τους μορφή, το σύστημα μπορεί να κατανοήσει καλύτερα τις προτιμήσεις των χρηστών και να παρέχει πιο σχετικές και ακριβείς προτάσεις. Για παράδειγμα, αν ένας χρήστης αναζητήσει ταινίες με θέμα το "αγαπώντας", το σύστημα θα μπορέσει να αναγνωρίσει και να προτείνει ταινίες που περιέχουν τις λέξεις "αγαπημένος", "αγάπη" ή "αγάπησα" στη περιγραφή τους.

Στην συγκεκριμένη εργασία, η λημματοποίηση εφαρμόζεται ως μέρος της διαδικασίας προεπεξεργασίας των δεδομένων εισόδου. Χρησιμοποιούνται εργαλεία όπως το Natural Language Toolkit (NLTK) για να πραγματοποιήσουμε τη λημματοποίηση, όπως ακριβώς και για την διαδικασία αποκοπής καταλήξεων, εξασφαλίζοντας ότι τα δεδομένα είναι καθαρά και ομοιογενή. Αυτή η διαδικασία συμβάλλει σημαντικά στη βελτίωση της ακρίβειας και της απόδοσης του συστήματος προτάσεων, καθιστώντας το πιο ικανό να κατανοεί και να ανταποκρίνεται στις ανάγκες των χρηστών. Με αυτόν τον τρόπο, η λημματοποίηση παίζει έναν κρίσιμο ρόλο στην αποτελεσματική λειτουργία του συστήματος προτάσεων ταινιών.

Εφαρμογή λημματοποίησης (Lemmatization)



Σχ5. Παρουσιάζεται ένα ενδεικτικό παράδειγμα της διαδικασίας της λημματοποίησης η οποία μετατρέπει τις επιμέρους λέξεις στην βασική τους λεκτική μορφή ή αλλιώς ρίζα, λαμβάνοντας υπόψη την γραμματική και τα συμφραζόμενα των λέξεων.

4.10 Term Frequency - Inverse Document Frequency (TF-IDF)

Η έννοια TF-IDF αναπαριστά τη συντομογραφία των όρων "Term Frequency - Inverse Document Frequency". Πρόκειται για ένα στατιστικό αλγόριθμο που χρησιμοποιείται στον τομέα της επεξεργασίας φυσικής γλώσσας για την αξιολόγηση της σημαντικότητας μιας λέξης σε ένα έγγραφο σε σχέση με ένα σύνολο εγγράφων.

Ο στατιστικός αλγόριθμος TF-IDF αξιολογεί τη σημαντικότητα μιας λέξης εντός ενός εγγράφου με βάση δύο κύριους παράγοντες.

- Πρώτος παράγοντας είναι η συχνότητα εμφάνισης της λέξης στο εν λόγω έγγραφο (TF - Term Frequency). Αυτός ο παράγοντας μετράει πόσο συχνά εμφανίζεται μια λέξη σε ένα συγκεκριμένο έγγραφο. Όσο πιο συχνά εμφανίζεται μια λέξη σε ένα έγγραφο, τόσο μεγαλύτερη είναι η τιμή του TF για αυτήν τη λέξη στο συγκεκριμένο έγγραφο.
- Δεύτερος παράγοντας αποτελεί η αντίστροφη συχνότητα εμφάνισης της λέξης σε όλα τα έγγραφα (IDF - Inverse Document Frequency). Ο συγκεκριμένος παράγοντας μετράει τη σπανιότητα της λέξης σε όλα τα έγγραφα που είναι διαθέσιμα. Λέξεις που εμφανίζονται σε λίγα έγγραφα έχουν υψηλότερη τιμή IDF, ενώ λέξεις που εμφανίζονται σε πολλά έγγραφα έχουν χαμηλότερη τιμή IDF.

Η συνολική τιμή TF-IDF για μια λέξη υπολογίζεται ως το γινόμενο της τιμής TF και της τιμής IDF. Έτσι, λέξεις που εμφανίζονται συχνά σε ένα συγκεκριμένο έγγραφο αλλά εμφανίζονται σπάνια σε άλλα έγγραφα έχουν υψηλή τιμή TF-IDF, δείχνοντας ότι είναι σημαντικές για το εν λόγω έγγραφο. Αυτή η τεχνική χρησιμοποιείται ευρέως σε πολλές εφαρμογές, όπως η ανάλυση κειμένου, η αναζήτηση πληροφοριών και η αναγνώριση θεμάτων.

Ο τύπος του TF-IDF εκφράζεται μαθηματικά ως εξής:

TF-IDF = Term Frequency (TF) * Inverse Document Frequency (IDF)

$$TF = \frac{\text{αριθμός εμφανίσεων της λέξης στο έγγραφο}}{\text{συνολικός αριθμός λέξεων στο έγγραφο}}$$

$$IDF = \log \left(\frac{\text{συνολικός αριθμός εγγράφων}}{\text{αριθμός εγγράφων που περιέχουν τη λέξη}} \right)$$

Για τη συγκεκριμένη εργασία, αφού έχει ήδη γίνει η επεξεργασία φυσικής γλώσσας (Natural Language Processing), εφαρμόζεται ο αλγόριθμος TF-IDF (Term Frequency-Inverse Document Frequency) για να εντοπιστούν οι πιο σημαντικές λέξεις στα κείμενα των περιγραφών των κινηματογραφικών ταινιών. Ο στόχος μας είναι να μειώσουμε το μέγεθος των περιγραφών χωρίς όμως να χάσουμε την ουσία των πληροφοριών που περιέχουν. Η ανάλυση TF-IDF είναι ένας ισχυρός στατιστικός τρόπος μέτρησης που μας βοηθά να αξιολογήσουμε τη σημασία μιας λέξης σε ένα έγγραφο σε σχέση με μια συλλογή εγγράφων

Με την εφαρμογή του αλγορίθμου TF-IDF στην εργασία, κάθε λέξη στο κείμενο της περιγραφής λαμβάνει μια τιμή βαθμολογίας (Score) που καθορίζει τη σημαντικότητά της στο κείμενο. Η τιμή αυτή υπολογίζεται με βάση τη συχνότητα εμφάνισης της λέξης στο συγκεκριμένο έγγραφο και την αντίστροφη συχνότητα εμφάνισης της λέξης σε όλα τα έγγραφα της συλλογής. Έτσι, λέξεις που εμφανίζονται συχνά σε ένα έγγραφο αλλά σπάνια σε άλλα, λαμβάνουν υψηλή τιμή TF-IDF, επισημαίνοντας τη μοναδική συμβολή τους στο κείμενο. Αντίθετα, κοινές λέξεις που εμφανίζονται παντού λαμβάνουν χαμηλή τιμή TF-IDF, καθώς δεν προσφέρουν πολύτιμη πληροφορία.

Αυτή η διαδικασία μας επιτρέπει να αναγνωρίσουμε ποιες λέξεις έχουν τη μεγαλύτερη σημασία και συμβάλλουν περισσότερο στην κατανόηση του περιεχομένου της περιγραφής. Στο στάδιο αυτό, οι περιγραφές των κινηματογραφικών ταινιών, οι οποίες μπορεί να περιέχουν πολλές περιττές ή κοινές λέξεις, γίνονται πιο συμπυκνωμένες και περιεκτικές. Με τον τρόπο αυτό, καταφέρνουμε να επικεντρωθούμε στις λέξεις-κλειδιά που καθορίζουν την ουσία της περιγραφής κάθε ταινίας, διευκολύνοντας την ανάλυση και την επεξεργασία των δεδομένων από το σύστημα προτάσεων ταινιών.

Στο τελικό στάδιο αυτού του βήματος, αφαιρούμε από τις περιγραφές τις λέξεις που δεν είναι αρκετά σημαντικές και των οποίων η τιμή της βαθμολογίας (Score) δεν υπερβαίνει ένα καθορισμένο κατώφλι (Threshold) που έχουμε ορίσει. Με αυτόν τον τρόπο, καταφέρνουμε να διατηρήσουμε μόνο τις πιο κρίσιμες λέξεις στις περιγραφές, μειώνοντας έτσι το μέγεθος των δεδομένων ενώ διατηρούμε τις πιο σημαντικές πληροφορίες. Αυτή η επιλογή των λέξεων

βασίζεται στην τιμή TF-IDF που λαμβάνει κάθε λέξη, επιτρέποντας μας να φιλτράρουμε τις λιγότερο σημαντικές λέξεις και να επικεντρωθούμε στις πιο σημαντικές.

Εν κατακλείδι, η χρήση της ανάλυσης TF-IDF για την επεξεργασία των περιγραφών των κινηματογραφικών ταινιών μας επιτρέπει την εξαγωγή των πιο κρίσιμων στοιχείων από κάθε περιγραφή. Αυτό το βήμα είναι κρίσιμο για την αποτελεσματική λειτουργία του συστήματος προτάσεων ταινιών, καθώς επιτρέπει στο σύστημα να επικεντρώνεται στις λέξεις που πραγματικά καθορίζουν το νόημα της περιγραφής κάθε ταινίας. Με την απομάκρυνση των λιγότερο σημαντικών λέξεων, το σύστημα μπορεί να παρέχει πιο ακριβείς και σχετικές προτάσεις ταινιών στους χρήστες.

4.11 Διανυσματική αναπαράσταση λέξεων (Word Vectorization)

Η διανυσματική αναπαράσταση λέξεων (Word Vectorization) αποτελεί μια θεμελιώδη τεχνική στην επεξεργασία φυσικής γλώσσας, επιτρέποντας τη μετατροπή των λέξεων σε διανύσματα στον πολυδιάστατο χώρο, τα οποία μπορούν να επεξεργαστούν από υπολογιστικά συστήματα. Υπάρχουν διάφοροι τρόποι για να επιτευχθεί αυτή η αναπαράσταση, ο καθένας με τα δικά του πλεονεκτήματα και μειονεκτήματα.

Μία από τις πρώτες και πιο απλές μεθόδους αναπαράστασης είναι η Bag of Words (BoW), η οποία μετατρέπει ένα κείμενο σε έναν διάνυσμα συχνοτήτων των λέξεων χωρίς να λαμβάνει υπόψη τη σειρά των λέξεων. Παρόλο που είναι εύκολη στην εφαρμογή, η μέθοδος αυτή δεν διατηρεί τη σημασιολογική πληροφορία των λέξεων που περιέχονται στο κείμενο.

Η μέθοδος TF-IDF (Term Frequency-Inverse Document Frequency) βελτιώνει την προσέγγιση της μεθόδου Bag of Words (BoW), λαμβάνοντας υπόψη τη συχνότητα εμφάνισης μιας λέξης σε σχέση με το πόσο συχνά εμφανίζεται σε όλα τα έγγραφα. Αυτή η τεχνική βοηθάει να αναδειχθούν οι σημαντικές λέξεις ενός κειμένου και να μειωθεί ο θόρυβος από τις πολύ κοινές λέξεις.

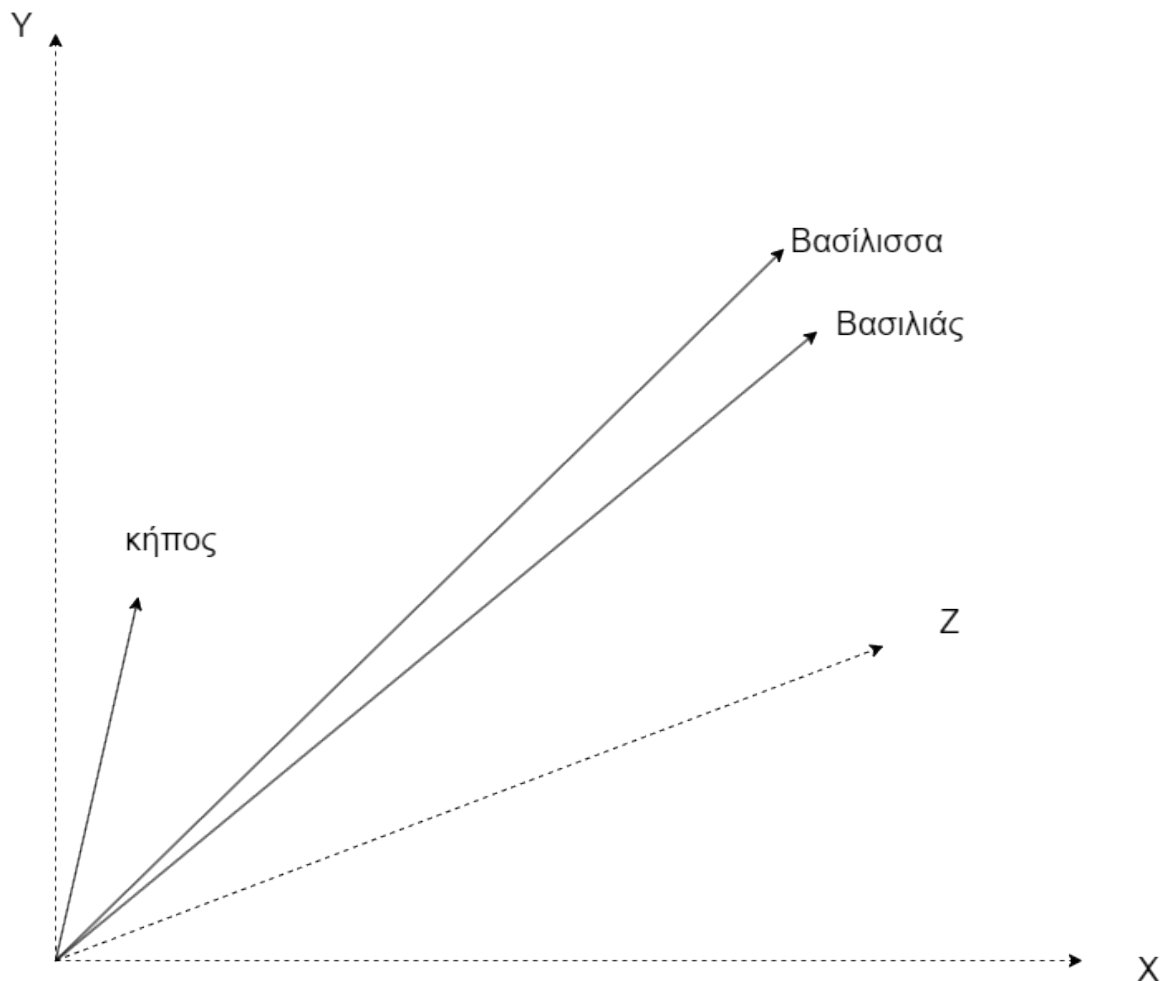
Μία πιο βελτιωμένη και σύγχρονη μέθοδος είναι η διανυσματική αναπαράσταση των λέξεων με την χρήση του Word2Vec, το οποίο χρησιμοποιεί νευρωνικά δίκτυα για να δημιουργήσει διανύσματα λέξεων. Υπάρχουν δύο κύριες προσεγγίσεις στο Word2Vec: το Continuous Bag of Words (CBOW) και το Skip-gram. Το CBOW προβλέπει μια λέξη από τις γειτονικές λέξεις, ενώ το Skip-gram προβλέπει τις γειτονικές λέξεις από μία δεδομένη λέξη. Το Word2Vec καταφέρνει να διατηρεί τη σημασιολογική πληροφορία των λέξεων, κάνοντας τις λέξεις με παρόμοια σημασία να βρίσκονται κοντά στο διανυσματικό χώρο.

Μια άλλη βελτιωμένη μέθοδος είναι η διανυσματική αναπαράσταση των λέξεων με την χρήση του GloVe (Global Vectors for Word Representation). Το GloVe κατασκευάζει έναν πίνακα συνύπαρξης λέξεων και χρησιμοποιεί στατιστικά δεδομένα από έναν μεγάλο όγκο

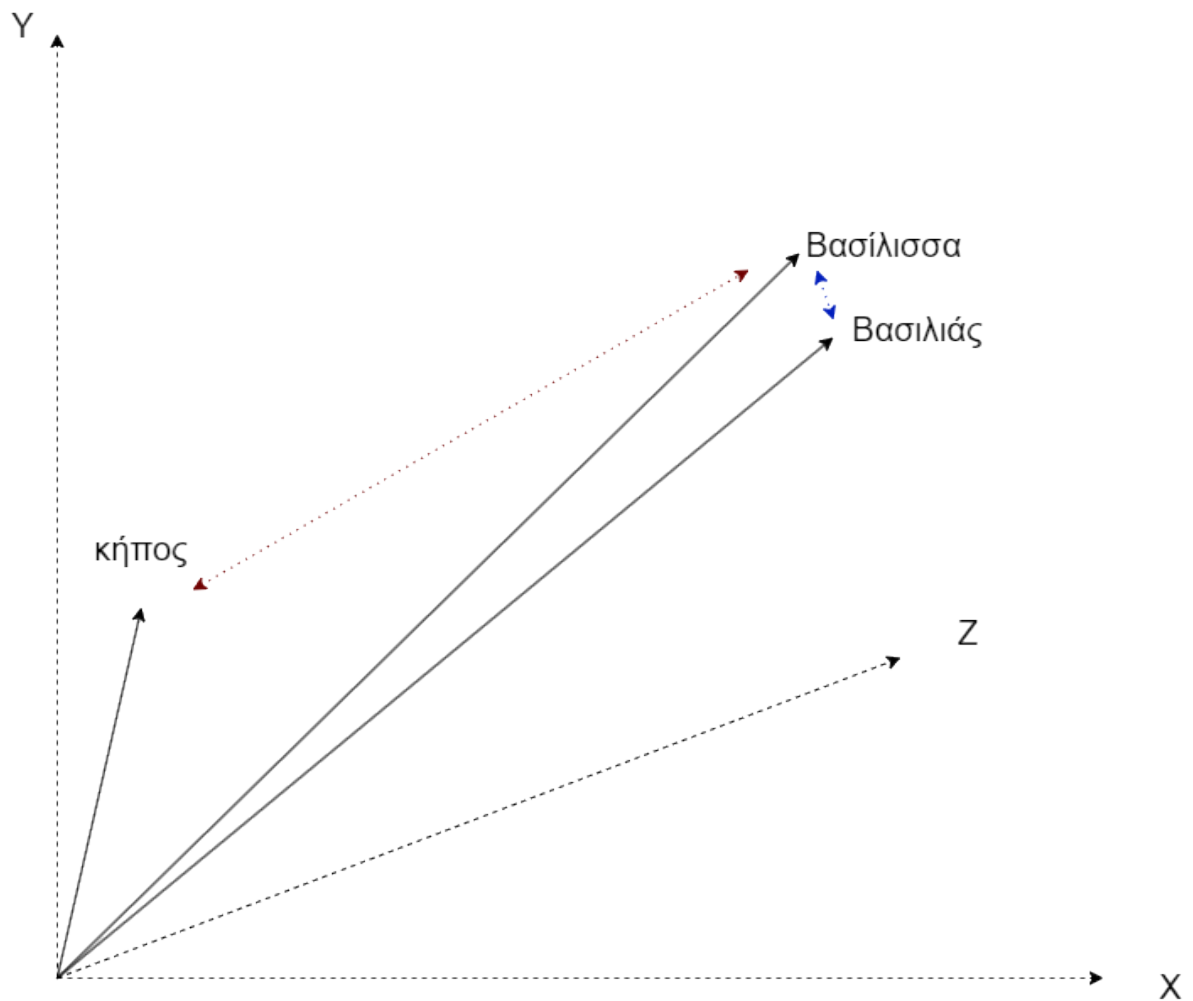
κειμένων για να εκπαιδεύσει ένα μηχανισμό μετατροπής λέξεων σε διανύσματα. Αυτό το μοντέλο αναδεικνύει τις σημασιολογικές σχέσεις μεταξύ λέξεων με μεγαλύτερη ακρίβεια.

Για τους σκοπούς στην συγκεκριμένη εργασία, επιλέξαμε τη διανυσματική αναπαράσταση των λέξεων με τη χρήση του Global Vectors for Word Representation (GloVe) επειδή μας επιτρέπει να αντιπροσωπεύουμε τις σημασιολογικές σχέσεις μεταξύ τους σε έναν πολυδιάστατο χώρο. Το GloVe δημιουργεί διανύσματα βάσει της συχνότητας εμφάνισης των λέξεων σε διάφορα κείμενα, οι οποίες συνήθως αντιστοιχούν σε σημασιολογικές συνδέσεις. Αυτό μας βοηθάει να κατανοήσουμε και να αντλήσουμε τη σημασιολογία των λέξεων, κάτι που είναι χρήσιμο για αναζητήσεις σημασιολογικών σχέσεων και άλλες εργασίες επεξεργασίας κειμένου

Διανυσματική αναπαράσταση λέξεων



Διανυσματική αναπαράσταση λέξεων



Σχ6. Στα παραπάνω σχήματα παρουσιάζεται σχηματικά η διανυσματική αναπαράσταση διαφόρων λέξεων με την μέθοδο Word Vectorization. Στο συγκεκριμένο παράδειγμα γίνεται ενδεικτική αναπαράσταση τριών λέξεων, οι οποίες είναι Βασιλιάς, Βασίλισσα και κήπος. Διαπιστώνεται ότι λέξεις που εμφανίζουν κοινή σημασιολογική έννοια έχουν κοντινά διανύσματα. Στο σχήμα κατανοούμε ότι οι λέξεις που έχουν κοινή σημασιολογία, έχουν κοντινά διανύσματα. Σε αντίθεση σε λέξεις που δεν υπάρχει σημασιολογική συσχέτιση, τα διανύσματα εμφανώς διαφοροποιούνται.

4.12 Global Vectors for Word Representation (GloVe)

Ως ένα από τα τελευταία βήματα για την προεπεξεργασία του συνόλου δεδομένων των κινηματογραφικών ταινιών που θα χρησιμοποιηθεί στην εργασία, εφαρμόζεται μια τεχνική για την αναπαράσταση των επιμέρους λέξεων των κειμένων ως διανύσματα, γνωστή και ως Global Vectors for Word Representation (GloVe). Η συγκεκριμένη μέθοδος επιτρέπει την αναπαράσταση των ξεχωριστών λέξεων ως διανυσμάτων σε έναν πολυδιάστατο χώρο, όπου κάθε λέξη αποδίδεται ως ένα σημείο στον χώρο αυτό. Η διαδικασία αυτή διευκολύνει την κατανόηση και την ανάλυση των λέξεων των κειμένων με βάση τις σημασιολογικές τους σχέσεις.

Τα διανύσματα GloVe έχουν την ιδιότητα να περιέχουν και να αποτυπώνουν τις σημασιολογικές σχέσεις μεταξύ ξεχωριστών λέξεων. Οι λέξεις που έχουν παρόμοιες ή παραπλήσιες σημασίες βρίσκονται διανυσματικά κοντά η μία στην άλλη στον πολυδιάστατο χώρο, ενώ λέξεις με διαφορετικές σημασίες που δεν συσχετίζονται βρίσκονται πιο μακριά στον πολυδιάστατο αυτό χώρο. Για παράδειγμα, οι λέξεις "βασιλιάς" και "βασιλίτσα" θα βρίσκονται κοντά στον χώρο GloVe, όπως επίσης και οι λέξεις "νερό" και "ποτάμι", αναπαριστώντας την εννοιολογική τους συγγένεια. Αντιθέτως λέξεις όπως "λουλούδι" και "πόρτα" δεν βρίσκονται κοντά στο διανυσματικό χώρο λόγω της χαμηλής τους συσχέτισης. Αυτή η ιδιότητα των διανυσμάτων GloVe καθιστά δυνατή την αναζήτηση και κατανόηση σημασιολογικών σχέσεων μεταξύ λέξεων με έναν πιο αποδοτικό και ακριβή τρόπο.

Η χρήση των διανυσμάτων GloVe έχει πολλές εφαρμογές στην ανάλυση κειμένου. Μπορούν να χρησιμοποιηθούν για την αναζήτηση σημασιολογικών σχέσεων μεταξύ λέξεων, επιτρέποντας την αναγνώριση λέξεων που έχουν παρόμοια σημασία ή ρόλο σε διαφορετικά κείμενα. Επιπλέον, τα διανύσματα GloVe μπορούν να αξιοποιηθούν για συστάσεις κειμένων, κατηγοριοποίηση κειμένων, και άλλες εργασίες που απαιτούν κατανόηση του νοήματος των λέξεων και της σημασίας τους.

Τα διανύσματα GloVe (Global Vectors for Word Representation) επιτυγχάνουν την αναπαράσταση των λέξεων με τη χρήση στατιστικών δεδομένων από έναν μεγάλο όγκο κειμένων. Το μοντέλο GloVe μαθαίνει τις αναπαραστάσεις των λέξεων λαμβάνοντας υπόψη τη συχνότητα εμφάνισης των λέξεων σε διάφορα κείμενα. Με αυτόν τον τρόπο, οι λέξεις που συνυπάρχουν συχνά αποκτούν διανύσματα που είναι κοντά το ένα στο άλλο στον πολυδιάστατο χώρο, αναδεικνύοντας τη σημασιολογική τους σύνδεση και σχέση.

Η τεχνική της αναπαράστασης λέξεων ως διανύσματα τύπου GloVe σε ένα πολυδιάστατο χώρο, αποτελεί ένα ισχυρό εργαλείο για την ανάλυση και την επεξεργασία της φυσικής γλώσσας, επιτρέποντας την κατανόηση και τη χρήση της σημασιολογίας των λέξεων με μεγαλύτερη ακρίβεια και αποτελεσματικότητα σε σχέση με άλλες μεθόδους, όπως για παράδειγμα τις πιο απλές μεθόδους One-Hot Encoding για την αναπαράσταση λέξεων και Count Vectors (ή αλλιώς Bag of Words) που αναπαριστά ένα κείμενο ως ένα διάνυσμα συχνοτήτων των λέξεων που περιέχει.

Διανυσματική αναπαράσταση λέξεων

An undercover cop and a mole in the police attempt to identify each other while infiltrating an Irish gang in South Boston.



"an" -> [0.134, 0.56, -0.25, 0.87, ...]
"undercover" -> [0.12, 0.34, -0.67, 0.56, ...]
"cop" -> [0.45, -0.24, 0.19, 0.72, ...]
"and" -> [0.89, 0.12, -0.33, 0.44, ...]
"a" -> [0.34, 0.67, -0.45, 0.88, ...]
"mole" -> [0.11, -0.34, 0.44, 0.67, ...]
"in" -> [0.24, 0.56, -0.45, 0.34, ...]
"the" -> [0.56, -0.12, 0.34, 0.45, ...]
"police" -> [0.78, 0.11, -0.23, 0.67, ...]
"attempt" -> [0.23, -0.56, 0.45, 0.89, ...]
"to" -> [0.67, 0.12, -0.34, 0.45, ...]
"identify" -> [0.45, -0.23, 0.56, 0.78, ...]
"each" -> [0.12, 0.89, -0.45, 0.67, ...]
"other" -> [0.56, -0.34, 0.12, 0.45, ...]
"while" -> [0.34, 0.45, -0.12, 0.89, ...]
"infiltrating" -> [0.67, -0.23, 0.78, 0.45, ...]
"an" -> [0.134, 0.56, -0.25, 0.87, ...]
"Irish" -> [0.11, 0.45, -0.34, 0.67, ...]
"gang" -> [0.78, -0.45, 0.23, 0.56, ...]
"in" -> [0.24, 0.56, -0.45, 0.34, ...]
"South" -> [0.45, 0.78, -0.34, 0.12, ...]
"Boston" -> [0.56, -0.34, 0.45, 0.67, ...]

Σχ7. Στο σχήμα περιγράφεται ο τρόπος με τον οποίο πραγματοποιείται η διανυσματική αναπαράσταση επιμέρους λέξεων σε ένα κείμενο. Παρατηρείται ότι κάθε λέξη μετατρέπεται σε ένα αριθμητικό διάνυσμα στο πολυδιάστατο χώρο. Όλες οι λέξεις παράγουν διανύσματα ίδιων διαστάσεων μεταξύ τους. Τα διανύσματα αντιπροσωπεύουν τη σημασιολογική τους συσχέτιση στο χώρο.

Διανυσματική αναπαράσταση λέξεων

An undercover cop and a mole in the police attempt to identify each other while infiltrating an Irish gang in South Boston.



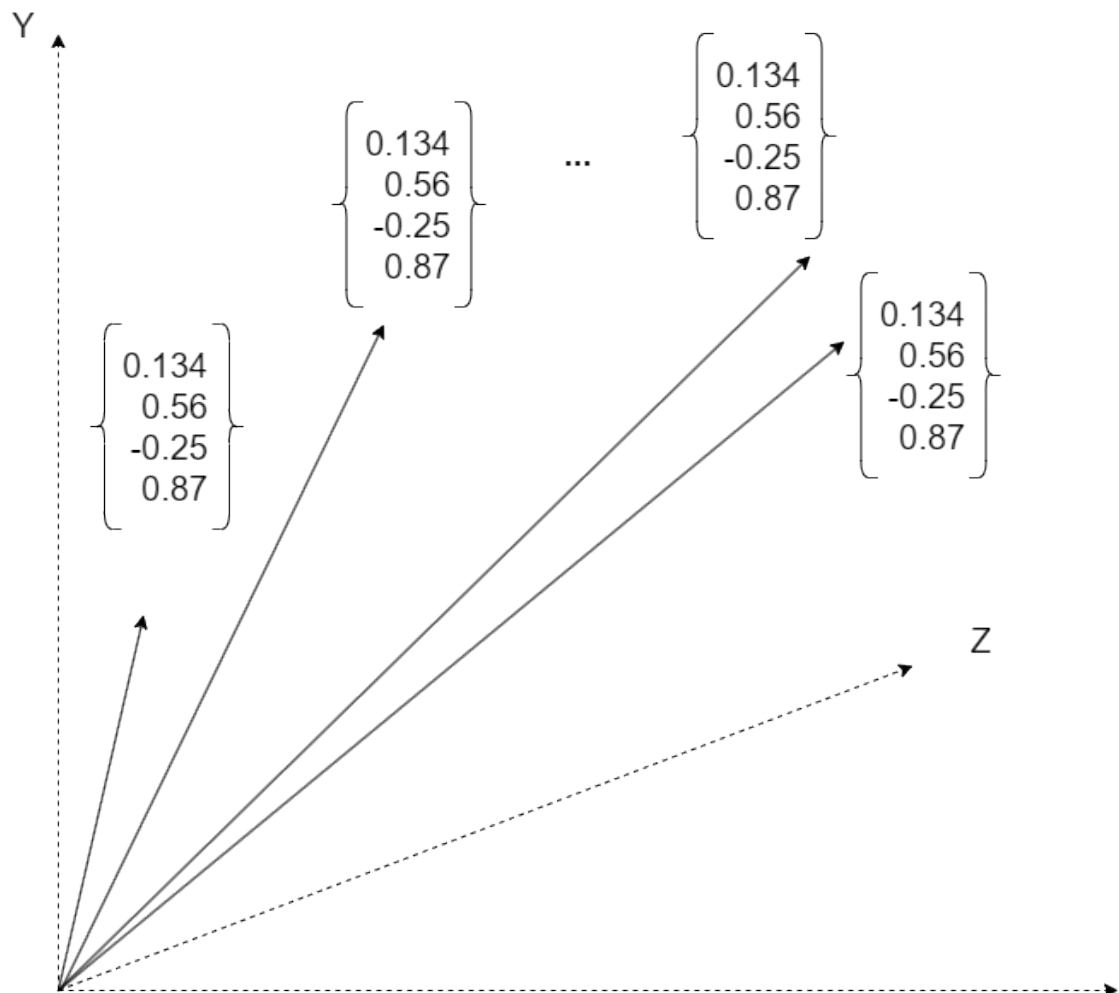
$$\begin{bmatrix} 0.134 \\ 0.56 \\ -0.25 \\ 0.87 \end{bmatrix} \begin{bmatrix} 0.134 \\ 0.56 \\ -0.25 \\ 0.87 \end{bmatrix} \begin{bmatrix} 0.134 \\ 0.56 \\ -0.25 \\ 0.87 \end{bmatrix} \begin{bmatrix} 0.134 \\ 0.56 \\ -0.25 \\ 0.87 \end{bmatrix} \dots \begin{bmatrix} 0.134 \\ 0.56 \\ -0.25 \\ 0.87 \end{bmatrix} \begin{bmatrix} 0.134 \\ 0.56 \\ -0.25 \\ 0.87 \end{bmatrix}$$

Σχ8. Στο σχήμα παρουσιάζεται η διαδικασία μετατροπής ενός κειμένου σε φυσική γλώσσα σε ένα πίνακα διανυσμάτων. Παρατηρείται ότι κάθε λέξη μετατρέπεται σε ένα διάνυσμα και με τον τρόπο αυτό, δεδομένα τύπου συμβολοσειράς παράγουν ένα πίνακα διανυσμάτων.

An undercover cop and a mole in the police attempt to identify each other while infiltrating an Irish gang in South Boston.



$$\begin{Bmatrix} 0.134 \\ 0.56 \\ -0.25 \\ 0.87 \end{Bmatrix} \begin{Bmatrix} 0.134 \\ 0.56 \\ -0.25 \\ 0.87 \end{Bmatrix} \begin{Bmatrix} 0.134 \\ 0.56 \\ -0.25 \\ 0.87 \end{Bmatrix} \begin{Bmatrix} 0.134 \\ 0.56 \\ -0.25 \\ 0.87 \end{Bmatrix} \dots \begin{Bmatrix} 0.134 \\ 0.56 \\ -0.25 \\ 0.87 \end{Bmatrix} \begin{Bmatrix} 0.134 \\ 0.56 \\ -0.25 \\ 0.87 \end{Bmatrix}$$



Σχ9. Στο σχήμα παρατηρείται, ότι κάθε λέξη μετασχηματίζεται σε ένα διάνυσμα δημιουργώντας την διανυσματική του αναπαράσταση στο χώρο.

Το GloVe (Global Vectors for Word Representation) πρόκειται για ένα αλγόριθμο μηχανικής εκμάθησης χωρίς επίβλεψη (unsupervised machine learning algorithm), ο οποίος χρησιμοποιείται για την αναπαράσταση λέξεων σε έναν πολυδιάστατο διανυσματικό χώρο. Αναπτύχθηκε από τον Jeffrey Pennington, Richard Socher και Christopher D. Manning στο Πανεπιστήμιο του Στάνφορντ. Ο αλγόριθμος GloVe συνδυάζει τα πλεονεκτήματα των παραδοσιακών μεθόδων στατιστικής αναπαράστασης λέξεων με τη χρήση μηχανικής μάθησης (Machine Learning ή αλλιώς ML).

Το μοντέλο GloVe δημιουργεί διανυσματικές αναπαραστάσεις λέξεων με βάση τις συνυπάρξεις τους σε ένα μεγάλο σύνολο κειμένων. Αντί να επικεντρωθεί αποκλειστικά στη συχνότητα εμφάνισης των λέξεων (όπως η μέθοδος TF-IDF), ο αλγόριθμος GloVe λαμβάνει υπόψη την πιθανότητα συνύπαρξης δύο λέξεων σε ένα κείμενο. Με αυτόν τον τρόπο, πραγματοποιείται κωδικοποίηση των σημασιολογικών σχέσεων μεταξύ των λέξεων στη διανυσματική τους αναπαράσταση.

Το GloVe (Global Vectors for Word Representation) παρέχει επίσης, ήδη προ εκπαιδευμένα μοντέλα για την αναπαράσταση λέξεων σε ένα πολυδιάστατο διανυσματικό χώρο. Το GloVe εκπαιδεύεται με σκοπό να υπολογίζει τις διανυσματικές αναπαραστάσεις λέξεων βάσει των συνδυασμών τους σε ένα μεγάλο σύνολο δεδομένων κειμένων, όπως για παράδειγμα τα κείμενα της Wikipedia. Αυτό το μοντέλο έχει προ εκπαιδευτεί ώστε να αντιπροσωπεύει το νόημα των λέξεων σε διανύσματα.

Γίνεται χρήση των ήδη προ εκπαιδευμένων μοντέλων GloVe επειδή έχουν εκπαιδευτεί σε τεράστια σύνολα δεδομένων όπως για παράδειγμα ακαδημαϊκά κείμενα, ιστοσελίδες, εφημερίδες και κείμενα βιβλίων, προκειμένου να αποκτήσουν πλήρη και πολύπλευρη κατανόηση της γλώσσας.

Η εκπαίδευση ενός νέου μοντέλου αναπαράστασης λέξεων εκ νέου μπορεί να είναι πολύ δαπανηρή και χρονοβόρα διαδικασία. Καταρχάς, απαιτεί μεγάλο όγκο δεδομένων για την εκπαίδευση του μοντέλου, ο οποίος μπορεί να περιλαμβάνει εκατομμύρια ή ακόμα και δισεκατομμύρια λέξεις και προτάσεις

Επιπλέον, η διαδικασία της εκπαίδευσης απαιτεί σημαντικούς υπολογιστικούς πόρους και χρόνο. Οι αλγόριθμοι μάθησης που χρησιμοποιούνται για τη δημιουργία των αναπαραστάσεων λέξεων εκτελούν πολλαπλούς υπολογισμούς και ενδεχομένως πρέπει να εφαρμοστούν επαναληπτικά σε όλο το σύνολο των δεδομένων.

4.13 Διανυσματική αναπαράσταση κειμένων (Document Vectorization)

Η τεχνική της διανυσματικής αναπαράστασης κειμένων είναι χρήσιμη σε πολλές εφαρμογές επεξεργασίας της φυσικής γλώσσας. Μπορεί να χρησιμοποιηθεί για την κατηγοριοποίηση κειμένων, την ανάλυση συναισθημάτων, την αναζήτηση πληροφοριών και την σύσταση περιεχομένου. Η ικανότητα να συλλαμβάνει τη σημασιολογική πληροφορία του κειμένου καθιστά αυτή τη μέθοδο εξαιρετικά ισχυρή και ευέλικτη για διάφορες χρήσεις στην επεξεργασία της φυσικής γλώσσας.

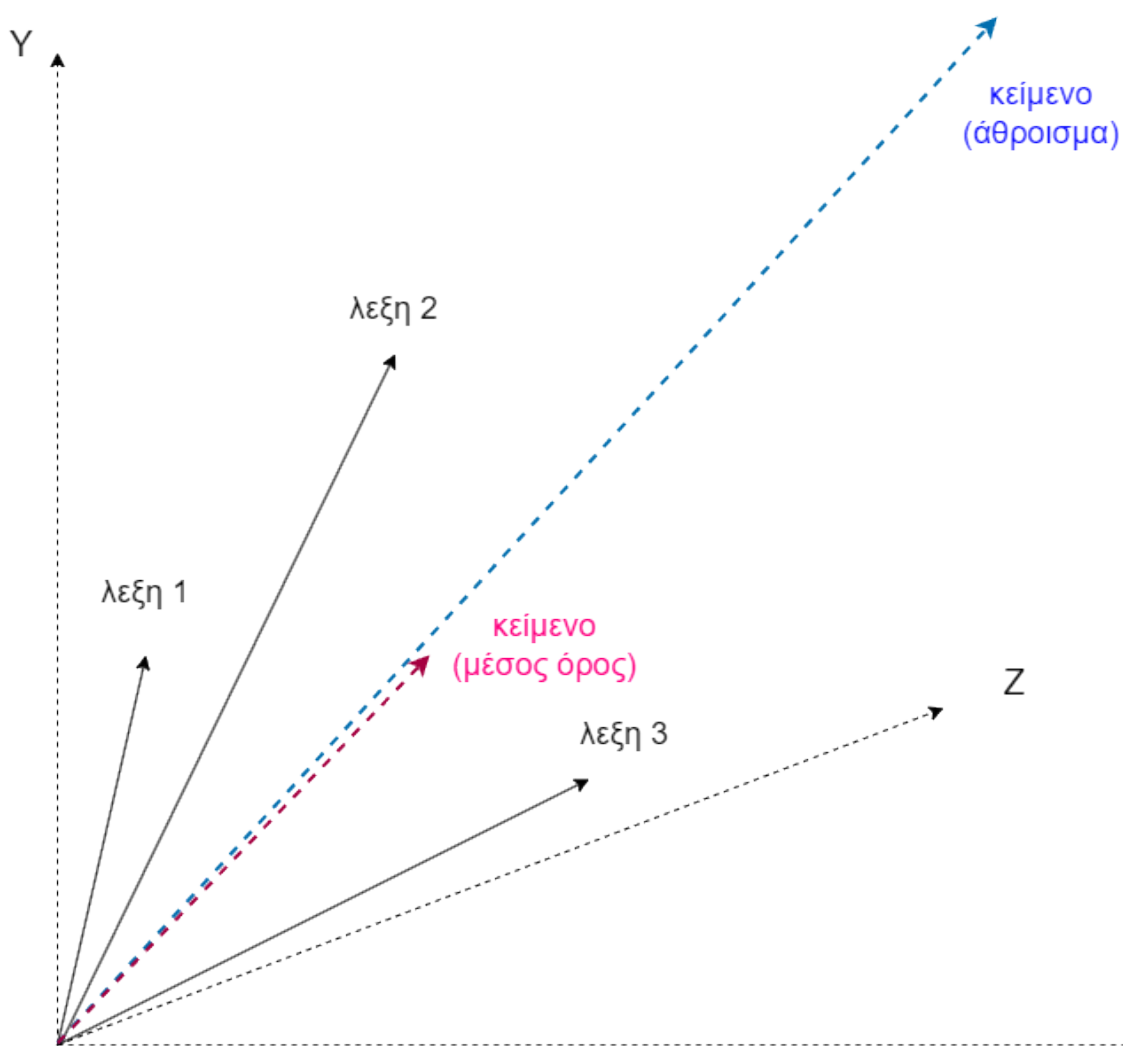
Όπως παρουσιάζεται παραπάνω, οι επιμέρους λέξεις των κειμένων στο σύνολο δεδομένων μετατρέπονται σε διανύσματα χρησιμοποιώντας το μοντέλο GloVe. Τελικός στόχος ωστόσο δεν είναι η διανυσματική αναπαράσταση των επιμέρους λέξεων του κειμένου αλλά η διανυσματική αναπαράσταση του ίδιου του κειμένου. Για να παραχθεί η διανυσματική αναπαράσταση του συνολικού κειμένου, τα διανύσματα των ξεχωριστών λέξεων συνδυάζονται σε ένα ενιαίο διάνυσμα. Ο συνδυασμός αυτός μπορεί να επιτευχθεί μέσω διαφόρων μεθόδων, όπως το άθροισμα, ο μέσος όρος ή άλλες στατιστικές τεχνικές. Η βασική ιδέα πίσω από τον συνδυασμό είναι ότι κάθε λέξη συνεισφέρει στη συνολική σημασία του κειμένου, και το τελικό διάνυσμα αντικατοπτρίζει αυτή τη συλλογική σημασιολογική πληροφορία του κειμένου.

Ο συνδυασμός των διανυσμάτων λέξεων επιτρέπει τη διατήρηση της σημασιολογικής πληροφορίας του κειμένου. Καθώς κάθε λέξη έχει το δικό της μοναδικό διάνυσμα που περιέχει πληροφορίες για τη σημασία και τις σχέσεις της με άλλες λέξεις, η ενιαία διανυσματική αναπαράσταση που προκύπτει από τον συνδυασμό αυτών των διανυσμάτων περιλαμβάνει πληροφορίες για τη συνολική σημασιολογική δομή του κειμένου. Έτσι, το τελικό διάνυσμα μπορεί να χρησιμοποιηθεί για την κατανόηση του κειμένου σε υψηλότερο επίπεδο καθώς επιτρέπει την διατήρηση της σημασιολογικής πληροφορίας των επιμέρους λέξεων που περιλαμβάνονται στο κείμενο.

Στο επόμενο βήμα πραγματοποιείται συνδυασμός των επιμέρους διανυσμάτων των λέξεων για κάθε κείμενο με στόχο να δημιουργηθεί μια διανυσματική αναπαράσταση για ολόκληρο το κείμενο της περιγραφής των κινηματογραφικών ταινιών. Η διαδικασία αυτή περιλαμβάνει τον συνδυασμό των επιμέρους διανυσματικών αναπαραστάσεων όλων των λέξεων που περιλαμβάνονται στο κείμενο της περιγραφής.

Υπάρχουν διάφοροι τρόποι για να συνδυαστούν τα διανύσματα των λέξεων ώστε να παραχθεί μία ενιαία διανυσματική αναπαράσταση για ένα κείμενο. Κάθε μέθοδος έχει τα δικά της πλεονεκτήματα και μειονεκτήματα, και η επιλογή της κατάλληλης μεθόδου εξαρτάται από τη συγκεκριμένη εφαρμογή και τους στόχους της ανάλυσης. Οι πιο κοινές μέθοδοι είναι το άθροισμα, ο μέσος όρος και η χρήση πιο σύνθετων τεχνικών.

Διανυσματική αναπαράσταση κειμένου (με άθροισμα και με μέσο όρο)



Σχ10. Στο σχήμα περιγράφονται οι δύο μέθοδοι, με τις οποίες μπορούμε να δημιουργήσουμε την διανυσματική αναπαράσταση ενός κειμένου χρησιμοποιώντας τα ξεχωριστά διανύσματα των λέξεων που περιέχονται σε αυτό. Η μια μέθοδος υπολογίζει το άθροισμα των επιμέρους διανυσμάτων, ώστε να προκύψει η διανυσματική αναπαράσταση ενός κειμένου, ενώ η άλλη μέθοδος υπολογίζει τον μέσο όρο τους. Όπως φαίνεται στο σχήμα, το διάνυσμα που προκύπτει από τις δύο μεθόδους έχει την ίδια κατεύθυνση ή αλλιώς γωνία και το μόνο που διαφοροποιείται είναι το μέγεθός του.

4.13.1 Άθροισμα των επιμέρους διανυσμάτων

Το άθροισμα των επιμέρους διανυσμάτων είναι μια απλή και άμεση μέθοδος. Σε αυτήν την προσέγγιση, τα διανύσματα όλων των λέξεων σε ένα κείμενο προστίθενται για να παραχθεί ένα τελικό διάνυσμα. Βασικό πλεονέκτημα της συγκεκριμένης μεθόδου είναι η απλότητα της, καθώς η υλοποίηση αυτής της μεθόδου είναι εύκολη και γρήγορη.

4.13.2 Μέσος Όρος των επιμέρους διανυσμάτων

Η εύρεση του μέσου όρου των ξεχωριστών διανυσμάτων είναι μια άλλη σχετικά απλή μέθοδος. Σε αυτή την προσέγγιση, τα διανύσματα όλων των λέξεων σε ένα κείμενο προστίθενται και στη συνέχεια διαιρούνται με τον αριθμό των λέξεων. Βασικό πλεονέκτημα της συγκεκριμένης μεθόδου είναι πως ο μέσος όρος εξισορροπεί τη συνεισφορά κάθε λέξης, αποφεύγοντας την υπερεκτίμηση των συνηθισμένων λέξεων.

Και οι δύο μέθοδοι έχουν τα πλεονεκτήματα και τα μειονεκτήματά τους, αλλά προσφέρουν μια καλή βάση για περαιτέρω ανάλυση και επεξεργασία. Η τελική απόφαση σχετικά με ποια μέθοδος θα χρησιμοποιηθεί εξαρτάται συχνά από την επόμενη φάση της διαδικασίας, όπως τον υπολογισμό αποστάσεων ή τη χρήση τους σε αλγορίθμους μηχανικής μάθησης. Ο τρόπος με τον οποίο θα υπολογιστεί η απόσταση μεταξύ των διανυσμάτων μπορεί να επηρεάσει την απόφαση για το αν θα επιλεγεί το άθροισμα ή ο μέσος όρος διανυσμάτων.

Το άθροισμα και ο μέσος όρος διανυσμάτων είναι και οι δύο αξιόπιστες και αποδεκτές μέθοδοι για τη συνδυαστική αναπαράσταση των λέξεων ενός κειμένου. Η επιλογή μεταξύ τους μπορεί να βασίζεται στις προτιμήσεις και στις ανάγκες του εκάστοτε συστήματος. Έτσι, η επιλογή της μεθόδου συνδυασμού των διανυσμάτων είναι ευέλικτη και μπορεί να προσαρμοστεί ανάλογα με τις απαιτήσεις και τους στόχους της ανάλυσης.

index	movie_name	description	clean_description
0	Babylon	A tale of outsized ambition and outrageous excess, It traces the rise and fall of multiple c...	tale outsiz ambit outrag excess trace rise fall multipl charact era unbridl decad deprav earli hollywo...
1	Amsterdam	In the 1930s, three friends witness a murder, are framed for it, and uncover one of the mo...	three friend wit murder frame uncov one outrag plot american histori
2	The Northman	A young Viking prince is on a quest to avenge his father's murder.	young vike princ quest aveng father s murder
3	Narvik: Hitler's First De...	April 1940. The eyes of the world are on Narvik, a small town in northern Norway, source ...	april eye world narvik small town northern norway sourc iron ore need hitler s war machineri two ...
4	The Woman King	A historical epic inspired by true events that took place in The Kingdom of Dahomey, one...	histor epic inspir true event took place kingdom dahomey one power state africa th th centuri
5	Oppenheimer	The story of American scientist J. Robert Oppenheimer and his role in the development of...	stori american scientis t robert oppenheim role develop atom bomb
6	Killers of the Flower M...	Members of the Osage tribe in the United States are murdered under mysterious circum...	member osag tribe unit state murder mysteri circumst spark major fbi investig involv j edgar hoover
7	Schindler's List	In German-occupied Poland during World War II, industrialist Oskar Schindler gradually b...	germanoccupi poland world war ii industrialist oskar schindler gradual becom concern jewish work...
8	She Said	New York Times reporters Megan Twohey and Jodi Kantor break one of the most importa...	new york time report megan twohey jodi kantor break one import stori gener stori help ignit move...
9	Dunkirk	Allied soldiers from Belgium, the British Commonwealth and Empire, and France are surro...	alli soldier belgium british commonwealth empir franc surround german armi evacu fierc battl worl...
10	Hidden Figures	The story of a team of female African-American mathematicians who served a vital role in...	stori team femal africanamerican mathematician serv vital role nasa earli year us space program
11	Till	In 1955, after Emmett Till is murdered in a brutal lynching, his mother vows to expose the...	emmett till murder brutal lynch mother vow expos racism behind attack work involv brought justic
12	Mission Majnu	In the 1970s, an undercover Indian spy takes on a deadly mission to expose a covert nucl...	undercov indian spi take deadli mission expos covert nuclear weapon program heart pakistan
13	Hacksaw Ridge	World War II American Army Medic Desmond T. Doss, who served during the Battle of Ok...	world war ii american armi medic desmond t doss serv battl okinawa refus kill peopl becom first m...
14	Medieval	The story of fifteenth century Czech icon and warlord Jan Zizka, who defeated armies of t...	stori fifteenth centuri czech icon warlord jan zizka defeat armi teuton order holi roman empir
15	The Big Short	In 2006-2007 a group of investors bet against the United States mortgage market. In thei...	group investor bet unit state mortgag market research discov flaw corrupt market
16	Braveheart	Scottish warrior William Wallace leads his countrymen in a rebellion to free his homeland ...	scottish warrior william wallac lead countryman rebellion free homeland tyranni king edward engla...
17	Argentina, 1985	A team of lawyers takes on the heads of Argentina's bloody military dictatorship during t...	team lawyer take head argentina s bloodi militari dictatorship battl odd race time
18	Tombstone	A successful lawman's plans to retire anonymously in Tombstone, Arizona are disrupted b...	success lawman s plan retir anonym tombston arizona disrupt kind outlaw famou elimin
19	The Last Duel	King Charles VI declares that Knight Jean de Carrouges settle his dispute with his squire b...	king charl vi declar knight jean de carroug settl disput squir challeng duel
20	Hamilton	The real life of one of America's foremost founding fathers and first Secretary of the Treas...	real life one america s foremost found father first secretari treasuri alexand hamilton captur live bro...
21	BlackBerry	The story of the meteoric rise and catastrophic demise of the world's first smartphone.	stori meteor rise catastroph demis world s first smartphon
22	The King	Hal, wayward prince and heir to the English throne, is crowned King Henry V after his tyra...	hal wayward princ heir english throne crown king henri v tyrann father dy young king must navig p...
23	Kingdom of Heaven	Balian of ibelin travels to Jerusalem during the Crusades of the 12th century, and there he...	balian ibelin travel jerusalem crusad th centuri find defend citi peopl
24	Caligula	A dramatization of the ascent to Caesar and subsequent reign of Caligula, one of the mos...	dramat ascent caesar subsequ reign caligula one notori leader ancient rome see ambit scheme per...
25	JFK	New Orleans District Attorney Jim Garrison discovers there's more to the Kennedy assassi...	new orlean district attorney jim garrison discov s kennedi assassin offici stori
26	History of the World: P...	Mel Brooks brings his one-of-a-kind comic touch to the history of mankind covering even...	mel brook bring oneofakind comic touch histori mankind cover event old testament french revolut ...
27	Black Hawk Down	The story of 160 elite U.S. soldiers who dropped into Mogadishu in October 1993 to capt...	stori elit us soldier drop mogadishu octob captur two top lieuten renegad warlord found desper ba...
28	Malcolm X	Biographical epic of the controversial and influential Black Nationalist leader, from his earl...	biograph epic controversi influenti black nationalist leader earli life career smaltim gangster minist...
29	12 Years a Slave	In the antebellum United States, Solomon Northup, a free black man from upstate New Y...	antebellum unit state solomon northup free black man upstat new york abduct sold slaveri
30	Corsage	A fictional account of one year in the life of Empress Elisabeth of Austria. On Christma...	fiction account one year life empress elisabeth austria christma eve elisabeth idol beauti turn offici ...
31	The Impossible	The story of a tourist family in Thailand caught in the destruction and chaotic aftermath o...	stori tourist famili thailand caught destruct chaotic aftermath indian ocean tsunami
32	Inherit the Wind	Based on a real-life case in 1925, two great lawyers argue the case for and against a Tenn...	base reallif case two great lawyer argu case tennesse scienc teacher accus crime teach evolut
33	Pearl Harbor	A tale of war and romance mixed in with history. The story follows two lifelong friends an...	tale war romanc mix histori stori follow two lifelong friend beauti nurs caught horror infam sunday ...

Σχ11. Στο παραπάνω σχήμα παρουσιάζεται το αποτέλεσμα του μετασχηματισμού των δεδομένων, με την εφαρμογή μεθόδων επεξεργασίας φυσικής γλώσσας (Natural Language Processing η αλλιώς NLP). Οι μέθοδοι αυτοί περιλαμβάνουν την εφαρμογή της αφαίρεσης τερματικών όρων (Stopwords), της αποκοπής καταλήξεων των λέξεων (Stemming), καθώς και της λημματοποίησης (Lemmatization). Το σχήμα αναφέρεται σε πραγματικά δεδομένα που χρησιμοποιούνται στην εργασία και παρουσιάζονται για λόγους κατανόησης της μορφής των δεδομένων της εργασίας.

index	movie_name	description	c	tfidf_descriptions
0	Babylon	A tale of outsize ambition and outrageous excess, it traces the rise and fall of multipl...	t...	trace tale rise outrag multipl hollywood fall excess era earli deprav decad charact ambit
1	Amsterdam	In the 1930s, three friends witness a murder, are framed for it, and uncover one of the...	t...	wit uncov three plot outrag one murder histor friend frame american
2	The Northman	A young Viking prince is on a quest to avenge his father's murder.	y...	young vike quest princ murder father aveng
3	Narvik: Hitler's First De...	April 1940. The eyes of the world are on Narvik, a small town in northern Norway, sou...	a...	world winter warfar war two town sourc small ore norway northern need month iron hitler first fierc eye defeat dealt april
4	The Woman King	A historical epic inspired by true events that took place in The Kingdom of Dahomey, ...	h...	true took th state power place one kingdom inspir histor event epic centuri africa
5	Oppenheimer	The story of American scientist J. Robert Oppenheimer and his role in the developme...	s...	stori scientist role robert develop bomb atom american
6	Killers of the Flower M...	Members of the Osage tribe in the United States are murdered under mysterious circ...	m...	unit tribe state spark mysteri murder member major involv investig fbi edgar circumst
7	Schindler's List	In German-occupied Poland during World War II, industrialist Oskar Schindler graduall...	g...	world wit war poland persecut oskar nazi jewish industrialist i gradual germanoccupi concern becom
8	She Said	New York Times reporters Megan Twohey and Jodi Kantor break one of the most imp...	n...	nyork time subject stori silenc shatter sexual report one new movement megan jodi import ignit hollywood help gener decad b...
9	Dunkirk	Allied soldiers from Belgium, the British Commonwealth and Empire, and France are s...	a...	world war surround soldier i german franc fierc evacu empir british belgium batti armi alli
10	Hidden Figures	The story of a team of female African-American mathematicians who served a vital rol...	s...	year vital us team stori space serv role program nasa mathematician femal earli africanamerican
11	Till	In 1955, after Emmett Till is murdered in a brutal lynching, his mother vows to expose ...	e...	work vow till racism murder mother lynch justic involv expos brutal brought behind attack
12	Mission: Impossible	In the 1970s, an undercover Indian spy takes on a deadly mission to expose a covert n...	u...	weapon undercov take spi program pakistan nuclear mission indian heart expos deadli covert
13	Hacksaw Ridge	World War II American Army Medic Desmond T. Doss, who served during the Battle of...	w...	world without war shot serv refus receiv peopl okinawa medic medal man kill ii honor histori first fire desmond becom batti ar...
14	Medieval	The story of fifteenth century Czech icon and warlord Jan Zizka, who defeated armies ...	s...	warlord stori roman order jan ikon holi empir defeat czech centuri armi
15	The Big Short	In 2006-2007 a group of investors bet against the United States mortgage market. In t...	g...	unit state research mortgag market investor group flaw discov corrupt bet
16	Braveheart	Scottish warrior William Wallace leads his countrymen in a rebellion to free his homel...	s...	william warrior wallac tyranni scottish rebellion lead king homeland free england edward countryman
17	Argentina, 1985	A team of lawyers takes on the heads of Argentina's bloody military dictatorship durin...	t...	time team take race odd militari lawyer head dictatorship bloodi batti argentina
18	Tombstone	A successful lawman's plans to retire anonymously in Tombstone, Arizona are disrupte...	s...	success retir plan outlaw lawman kind famiou elimin disrupt arizona anonym
19	The Last Duel	King Charles VI declares that Knight Jean de Carrouges settle his dispute with his squ...	k...	vi squir settl knight king jean duel disput declar de charl challeng
20	Hamilton	The real life of one of America's foremost founding fathers and first Secretary of the T...	r...	treasuri theater secretari richard real origin one live hamilton found foremost first father cast captur broadway america alexand
21	BlackBerry	The story of the meteoric rise and catastrophic demise of the world's first smartphone...	s...	world stori rise meteor first demis catastroph
22	The King	Hal, wayward prince and heir to the English throne, is crowned King Henry V after his ...	h...	wayward war tyrann throne string princ polit past palac navig must left king henri heir hal father english emot dy crown behind
23	Kingdom of Heaven	Balian of Ibelin travels to Jerusalem during the Crusades of the 12th century, and ther...	b...	travel th peopl jerusalem find defend crusad citi centuri
24	Caligula	A dramatization of the ascent to Caesar and subsequent reign of Caligula, one of the ...	d...	subsequ scheme rome reign pervers one notori leader dramatt decad caesar brutal ancient ambit
25	JFK	New Orleans District Attorney Jim Garrison discovers there's more to the Kennedy ass...	n...	stori orlean offic new kennedi jim garrison district discov attorney assassin
26	History of the World: P...	Mel Brooks brings his one-of-a-kind comic touch to the history of mankind covering ...	m...	vignett touch testament seri revolut old mel mankind histori french event episod cover comic comedi brook bring
27	Black Hawk Down	The story of 160 elite U.S. soldiers who dropped into Mogadishu in October 1993 to c...	s...	warlord us two top stori somali soldier renegad octob lieutenant larg heavili found forc elt drop desper captur battli arm
28	Malcolm X	Biographical epic of the controversial and influential Black Nationalist leader, from his ...	b...	smalltim nationalist nation ministri member life leader islam influenti gangster epic earli controversi career black biograph
29	12 Years a Slave	In the antebellum United States, Solomon Northup, a free black man from upstate Ne...	a...	nyork upstat unit state solomon sold slaveri new man free black abduct
30	Corsage	A fictional account of one year in the life of Empress Elisabeth of Austria. On Christm...	f...	year woman tam tri start public one old offic maintain imag idol fiction eve empress elisabeth deem christma beauti austria a...
31	The Impossible	The story of a tourist family in Thailand caught in the destruction and chaotic afterma...	s...	tsunami tourist thailand stori ocean indian famili destruct chaotic caught aftermath
32	Inherit the Wind	Based on a real-life case in 1925, two great lawyers argue the case for and against a T...	b...	two tennesse teacher teach scienc evallif lawyer great evoluit crime case base argu accus
33	Pearl Harbor	A tale of war and romance mixed in with history. The story follows two lifelong friends...	t...	war two tale sunday stori romanc nuns morn mix lifelong infam horror histori friend follow caught beauti

Σχ12. Στο παραπάνω σχήμα παρουσιάζεται το αποτέλεσμα του μετασχηματισμού των δεδομένων, με την εφαρμογή της μεθόδου TF-IDF. Στο συγκεκριμένο βήμα, υπολογίζεται η σημαντικότητα κάθε λέξης στα επιμέρους κείμενα της περιγραφής των κινηματογραφικών ταινιών, όπου οι λέξεις χαμηλής σημασίας αφαιρούνται. Σκοπός του συγκεκριμένου βήματος, είναι η μείωση του όγκου δεδομένων, χωρίς να χαθεί σημαντική πληροφορία από τα δεδομένα. Το σχήμα αναφέρεται σε πραγματικά δεδομένα που χρησιμοποιούνται στην εργασία και παρουσιάζονται για λόγους κατανόησης της μορφής των δεδομένων της εργασίας.

index	movie_name	description	e	t	vector_description
0	Babylon	A tale of outsized ambition and outrageous excess, it traces the rise and fall of multiple ch...	t...	t...	0.0019306312,-0.1054424943,-0.0031505004,0.0218352736,-0.0023948774,0.0131836291,-0.0463951267,0.1177897528,0.203211993...
1	Amsterdam	In the 1930s, three friends witness a murder, are framed for it, and uncover one of the mo...	t...	w...	-0.0858278722,0.0945162624,-0.1716418925,-0.2242411361,0.0954759081,0.2090927511,-0.2074833629,-0.1025132537,0.0059482494...
2	The Northman	A young Viking prince is on a quest to avenge his father's murder.	y...	y...	0.1494083286,0.1403121799,0.0239616632,-0.1179613397,-0.0477326661,0.2042426884,-0.0505155027,-0.0335721783,-0.0625559835...
3	Narvik: Hitler's First De...	April 1940. The eyes of the world are on Narvik, a small town in northern Norway, source ...	a...	w...	-0.1466106502,-0.1540019407,-0.1380464272,-0.1919043213,0.0507835102,0.1829893291,-0.0754509568,0.0339509062,-0.035551480...
4	The Woman King	A historical epic inspired by true events that took place in The Kingdom of Dahomey, one ...	t...	t...	0.1220257187,0.0572973862,-0.0445921868,-0.0819420815,-0.1366476056,0.1832107455,-0.0575258695,0.1948110867,0.130673771...
5	Oppenheimer	The story of American scientist J. Robert Oppenheimer and his role in the development of...	s...	s...	0.1241064168,0.1027112305,0.0126299929,-0.3251366615,-0.1623138785,0.0302460874,-0.2287817597,0.0678137839,-0.105212227...
6	Killers of the Flower M...	Members of the Osage tribe in the United States are murdered under mysterious circumst...	m...	u...	0.0674770996,0.1592016071,-0.1497180909,-0.0653445423,0.1178029999,0.0699410513,-0.0423859124,0.1865318269,0.139644363...
7	Schindler's List	In German-occupied Poland during World War II, industrialist Oskar Schindler gradually b...	g...	w...	-0.0603516586,0.1052814126,0.0059203259,-0.1266559213,0.1281054169,0.0299801622,-0.2105790294,-0.2587678134,0.046505988...
8	She Said	New York Times reporters Megan Twohey and Jodi Kantor break one of the most importa...	n...	y...	-0.1605184078,-0.0501674972,-0.1888795346,0.0325812024,-0.029042933,0.1232213452,0.0257472172,-0.0956457928,0.0756846294...
9	Dunkirk	Allied soldiers from Belgium, the British Commonwealth and Empire, and France are surr...	a...	w...	-0.0847830772,0.0214237515,0.0373261725,-0.0535325892,-0.0110085951,0.1351884156,-0.2326154858,0.0838885009,-0.075707249...
10	Hidden Figures	The story of a team of female African-American mathematicians who served a vital role in ...	t...	y...	0.1658722162,-0.0143361017,0.0813079029,-0.1590699852,-0.0863048062,-0.1103995007,-0.2094114125,-0.0638643056,-0.10966578...
11	Till	In 1955, after Emmett Till is murdered in a brutal lynching, his mother vows to expose the ...	e...	w...	0.0617638864,0.0582426365,-0.142478226,-0.1215125397,0.0074382816,-0.0197326075,-0.1912765354,-0.1365403235,0.0521109179...
12	Mission: Impossible	In the 1970s, an undercover Indian spy takes on a deadly mission to expose a covert nucle...	u...	w...	0.0154158184,0.0177046377,0.0717662647,-0.2417772859,-0.0388084538,0.0228100959,-0.157893252,0.0559328793,-0.0705504417...
13	Hacksaw Ridge	World War II American Army Medic Desmond T. Doss, who served during the Battle of Ok...	w...	w...	-0.0276997052,0.1197091565,0.0059038536,-0.0449153557,-0.1221318468,0.1007854193,-0.3135459721,0.1083155647,-0.030881604...
14	Medieval	The story of fifteenth century Czech icon and warlord Jan Zizka, who defeated armies of L...	t...	w...	0.2593633831,-0.0968879908,-0.0170751832,-0.0620089093,0.0906979069,0.0154873244,-0.1440763623,-0.0305653545,-0.030511618...
15	The Big Short	In 2006-2007 a group of investors bet against the United States mortgage market. In their...	g...	u...	-0.0378679037,0.0684805959,-0.153660146,0.0594985119,-0.034592595,-0.0959380046,0.0144103002,0.3320938945,0.1180272022...
16	Braveheart	Scottish warrior William Wallace leads his countrymen in a rebellion to free his homeland ...	s...	w...	-0.0537865385,-0.1307963896,0.1765336394,-0.0309774559,0.2310354561,-0.0086188912,0.004805191,0.2880190902,0.0581096323...
17	Argentina, 1985	A team of lawyers takes on the heads of Argentina's bloody military dictatorship during th...	t...	t...	0.1351627256,0.3052660227,-0.1280456483,-0.1745616049,0.0651092604,-0.1985331476,-0.0833400041,0.0630369261,0.2739559144...
18	Tombstone	A successful lawman's plans to retire anonymously in Tombstone, Arizona are disrupted b...	t...	s...	0.1011827439,0.1469472349,-0.0876446676,0.0489673921,-0.1452410231,0.1019705012,-0.0878724977,0.2704167664,-0.1052304807...
19	The Last Duel	King Charles VI declares that Knight Jean de Carrouges settle his dispute with his squire b...	t...	w...	0.2488328478,-0.1175028607,0.0820842812,-0.2183542056,0.1541314125,0.3799079485,-0.0997114331,-0.1085834354,-0.014442718...
20	Hamilton	The real life of one of America's foremost founding fathers and first Secretary of the Treas...	t...	t...	-0.0363988051,0.0020669072,-0.0563921779,-0.0983277077,0.0011879418,0.1156970114,-0.1341595965,-0.0520340631,-0.040002558...
21	BlackBerry	The story of the meteoric rise and catastrophic demise of the world's first smartphone.	t...	w...	0.056221664,0.231287818,-0.1057005003,-0.0432983339,-0.0886266604,0.2569681406,-0.3935849965,0.135439828,-0.170946002,-0...
22	The King	Hal, wayward prince and heir to the English throne, is crowned King Henry V after his tyr...	h...	w...	0.0258894434,-0.0268955678,0.1172908992,-0.044048477,0.076509878,0.076454334,-0.1137905195,0.0017298046,0.0821535364,0.3...
23	Kingdom of Heaven	Balian of Ibelin travels to Jerusalem during the Crusades of the 12th century, and there he...	b...	t...	0.0259811282,-0.0619175062,-0.1043287441,0.0064217523,-0.1264007539,-0.0522719994,0.04483327512,0.0923798829,-0.007527872...
24	Caligula	A dramatization of the ascent to Caesar and subsequent reign of Caligula, one of the mos...	d...	s...	0.1088350788,-0.1223848388,0.170905903,-0.1179114208,0.1483059913,0.0703380778,-0.0988483354,0.0585316718,0.085588336,-0...
25	JFK	New Orleans District Attorney Jim Garrison discovers there's more to the Kennedy assassi...	n...	s...	-0.2892852699,-0.2009760141,-0.2383276893,0.0173399986,0.0128257442,-0.0727124958,0.2408134043,0.02047386225,-0.043306879...
26	History of the World: P...	Mel Brooks brings his one-of-a-kind comic touch to the history of mankind covering even...	m...	v...	-0.1083065813,-0.106316834,0.021453321,-0.1355568235,-0.1234036088,0.1726746559,-0.1272256672,-0.0988305666,-0.02090590738...
27	Black Hawk Down	The story of 160 elite U.S. soldiers who dropped into Mogadishu in October 1993 to capti...	s...	w...	0.0853265956,0.0838335049,-0.0141086699,0.0425047539,-0.1064374372,0.053289663,-0.0439086445,0.131625154,0.157342419,-0...
28	Malcolm X	Biographical epic of the controversial and influential Black Nationalist leader, from his earl...	s...	s...	0.0379777714,-0.1411115974,0.1387766451,0.0154650025,0.0938496366,-0.28625553296,-0.3135735691,0.1115629002,-0.002478827...
29	12 Years a Slave	In the antebellum United States, Solomon Northup, a free black man from upstate New Y...	a...	y...	-0.2195353359,-0.0743138194,-0.2421762496,0.172295837,-0.000500061,0.0775314122,0.1474706084,0.0495899469,0.0856416747...
30	Centage	A fictional account of one year in the life of Empress Elisabeth of Austria. On Christmas E...	f...	y...	-0.1741993874,0.0310682069,-0.1391353095,-0.1486340386,0.2268045472,0.0531397872,-0.0970977396,-0.0444380432,-0.02541284...
31	The Impossible	The story of a tourist family in Thailand caught in the destruction and chaotic aftermath of...	t...	t...	0.1592877036,-0.3258400261,0.0383876488,0.1425787308,0.1400404748,0.029198803,0.0948531106,0.1964926869,0.0410734564,-0...
32	Inherit the Wind	Based on a real-life case in 1925, two great lawyers argue the case for and against a Tenn...	b...	t...	-0.1586536298,0.0745370016,-0.1078148931,-0.1831324695,0.1731707454,-0.0296369977,0.1064566001,-0.1679780036,0.024614401...
33	Pearl Harbor	A tale of war and romance mixed in with history. The story follows two Melong friends an...	t...	w...	-0.1316684633,-0.0484332778,0.0666576475,0.0083881775,0.1145197675,0.114043355,-0.0832287222,-0.0453292727,-0.087951184...

Σχ13. Στο παραπάνω σχήμα παρουσιάζεται το αποτέλεσμα του μετασχηματισμού των δεδομένων, μετά την μετατροπή των επιμέρους λέξεων στην διανυσματική τους αναπαράσταση σε συνδυασμό με τον υπολογισμό της διανυσματικής αναπαράστασης των κειμένων περιγραφής των κινηματογραφικών ταινιών. Παρατηρείται ότι, για κάθε κείμενο του συνόλου δεδομένων προκύπτει ένα διάνυσμα στο πολυδιάστατο χώρο, το οποίο αναπαριστά σημασιολογικά το κείμενο της περιγραφής των δεδομένων.

4.14 Διανυσματική αναπαράσταση του αιτήματος του χρήστη (Query Vectorization)

Όταν ο χρήστης υποβάλλει ένα αίτημα ή ερώτημα, προκύπτει το πρόβλημα της σύγκρισης του αιτήματος με τα δεδομένα των ταινιών προκειμένου να βρεθούν αποτελέσματα που να ταιριάζουν με το ζητούμενο που υποβάλλει ο χρήστης. Το πρόβλημα έγκειται στο πώς μπορούμε να αντιστοιχίσουμε το φυσικό κείμενο του αιτήματος του χρήστη με τα κείμενα των περιγραφών των ταινιών με έναν αποδοτικό και ακριβή τρόπο.

Η παραδοσιακή προσέγγιση της σύγκρισης λέξεων με απλή αντιστοίχιση συμβολοσειρών δεν είναι αποτελεσματική, διότι δεν λαμβάνει υπόψη τη σημασιολογική σύνδεση μεταξύ των λέξεων. Για παράδειγμα, λέξεις που έχουν παρόμοια σημασία ή χρησιμοποιούνται σε παρόμοια συμφραζόμενα μπορεί να μην εντοπιστούν σωστά αν χρησιμοποιούμε απλές αντιστοιχίσεις. Εδώ έρχεται η ανάγκη για προηγμένες μεθόδους διανυσματικής αναπαράστασης του αιτήματος.

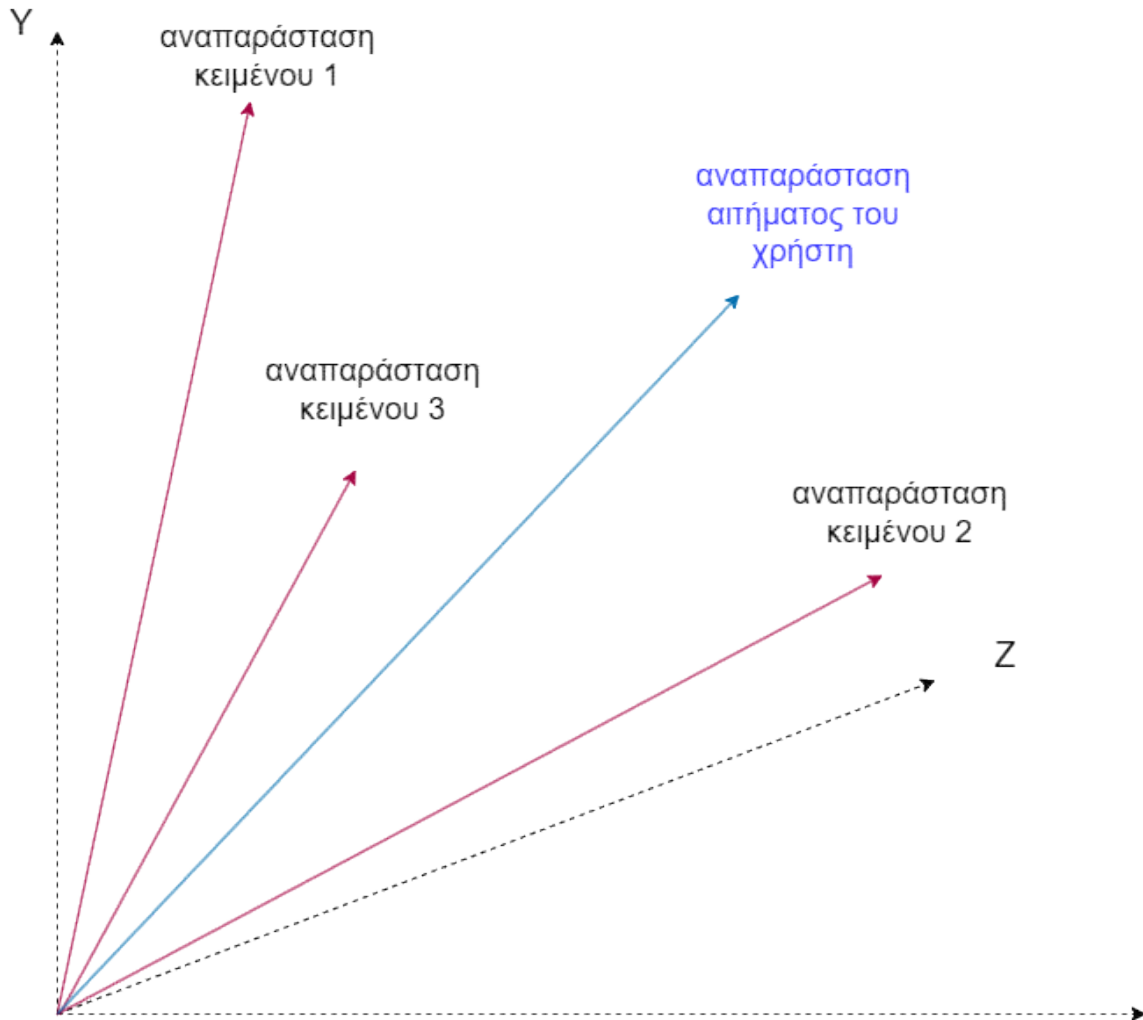
Όταν ο χρήστης υποβάλλει ένα αίτημα ή ερώτημα, αυτό το αίτημα επίσης μετατρέπεται σε διανυσματική αναπαράσταση, όπως ακριβώς και τα κείμενα του συνόλου δεδομένων που περιγράφονται παραπάνω. Η συγκεκριμένη διαδικασία επιτρέπει τη σύγκριση του αιτήματος του χρήστη με τα υπάρχοντα κείμενα για την εύρεση των πιο σχετικών και κοντινών κειμένων που είναι διαθέσιμα στο σύνολο.

Η μετατροπή του αιτήματος σε διάνυσμα (query vectorization) περιλαμβάνει τα ίδια βήματα προεπεξεργασίας που εφαρμόστηκαν στα κείμενα του συνόλου δεδομένων. Αυτό σημαίνει ότι το αίτημα περνάει από διαδικασίες όπως η αφαίρεση stopwords, η λημματοποίηση ή η αποκοπή καταλήξεων των λέξεων, και η μετατροπή των χαρακτήρων σε πεζούς. Στη συνέχεια, οι λέξεις του αιτήματος μετατρέπονται σε διανύσματα μέσω της ίδιας τεχνικής που χρησιμοποιήθηκε για τα κείμενα του συνόλου δεδομένων.

Αυτή η διανυσματική αναπαράσταση του αιτήματος επιτρέπει τον υπολογισμό της ομοιότητας μεταξύ του αιτήματος και των κειμένων του συνόλου δεδομένων. Μέσω αυτής της διαδικασίας, μπορούμε να βρούμε τα κείμενα που είναι πιο κοντά και σχετίζονται περισσότερο με το αίτημα του χρήστη. Αυτό γίνεται συνήθως με τη χρήση μεθόδων υπολογισμού απόστασης, όπως η ομοιότητα συνημιτόνου (Cosine Similarity), που μετράει την γωνιακή απόσταση μεταξύ των διανυσμάτων ή την Ευκλείδεια απόσταση (Euclidean distance) που υπολογίζει την Ευκλείδεια απόσταση των διανυσμάτων. Με αυτόν τον τρόπο, επιτυγχάνεται η αποδοτική και ακριβής εύρεση των σχετικών πληροφοριών, βελτιώνοντας την εμπειρία του χρήστη.

Η διαδικασία αυτή είναι εξαιρετικά χρήσιμη διότι, αντί να εξετάζουμε κάθε λέξη των κειμένων των κινηματογραφικών ταινιών ξεχωριστά όπως επίσης και του αιτήματος του χρήστη, μπορούμε να χρησιμοποιούμε αυτά τα διανύσματα για να κατανοήσουμε τη συνολική σημασία των περιγραφών και του αιτήματος. Αυτό βοηθά στη μείωση της διάστασης των δεδομένων (dimensionality reduction, ή dimension reduction) και στην εξαγωγή σημαντικών πληροφοριών (Feature extraction) με έναν αποδοτικότερο και αποτελεσματικό τρόπο.

Διανυσματική αναπαράσταση του αιτήματος χρήστη



Σχ14. Στο σχήμα παρουσιάζεται ενδεικτικά η διανυσματική αναπαράσταση μερικών κειμένων από το σύνολο δεδομένων, καθώς επίσης και η διανυσματική αναπαράσταση από το αίτημα ή αλλιώς ερώτημα του χρήστη, που δίνεται σε φυσική γλώσσα (Natural Language). Όλα τα κείμενα, καθώς και το ερώτημα μετατρέπονται σε ξεχωριστά διανύσματα.

4.15 Μέθοδοι σύγκρισης απόστασης διανυσμάτων

Καθώς έχει προηγηθεί η μετατροπή των δεδομένων του συνόλου σε διανυσματική μορφή, είναι κρίσιμο να γίνει κατανοητό πώς πραγματοποιείται η σύγκριση των διανυσμάτων που αντιπροσωπεύουν τα επιμέρους κείμενα της περιγραφής των κινηματογραφικών ταινιών και των αιτημάτων των χρηστών της εφαρμογής. Αφού μετατρέψουμε τα κείμενα σε διανύσματα, η επόμενη σημαντική διαδικασία είναι η σύγκρισή των διανυσμάτων για την εύρεση ομοιότητας ή απόστασης μεταξύ τους.

Σε αυτό το κεφάλαιο, θα παρουσιαστούν και θα εξεταστούν διάφορες μέθοδοι σύγκρισης απόστασης διανυσμάτων και θα γίνει ανάλυση σχετικά με τα πλεονεκτήματα και τα μειονεκτήματά για κάθε μία από τις μεθόδους που αναφέρονται. Αυτό θα βοηθήσει να γίνει κατανοητό πώς μπορούν αυτές οι μέθοδοι να εφαρμοστούν στη διανυσματική αναπαράσταση κειμένων, όπως στην περίπτωση της αναζήτησης κοντινών κειμένων ή στην ανάλυση των αιτημάτων των χρηστών.

Η σύγκριση απόστασης διανυσμάτων είναι κρίσιμη για την εύρεση της ομοιότητας μεταξύ του αιτήματος του χρήστη και των δεδομένων των ταινιών. Υπάρχουν αρκετές μέθοδοι που χρησιμοποιούνται για τον υπολογισμό αυτής της απόστασης, κάθε μία με τα δικά της πλεονεκτήματα και μειονεκτήματα.

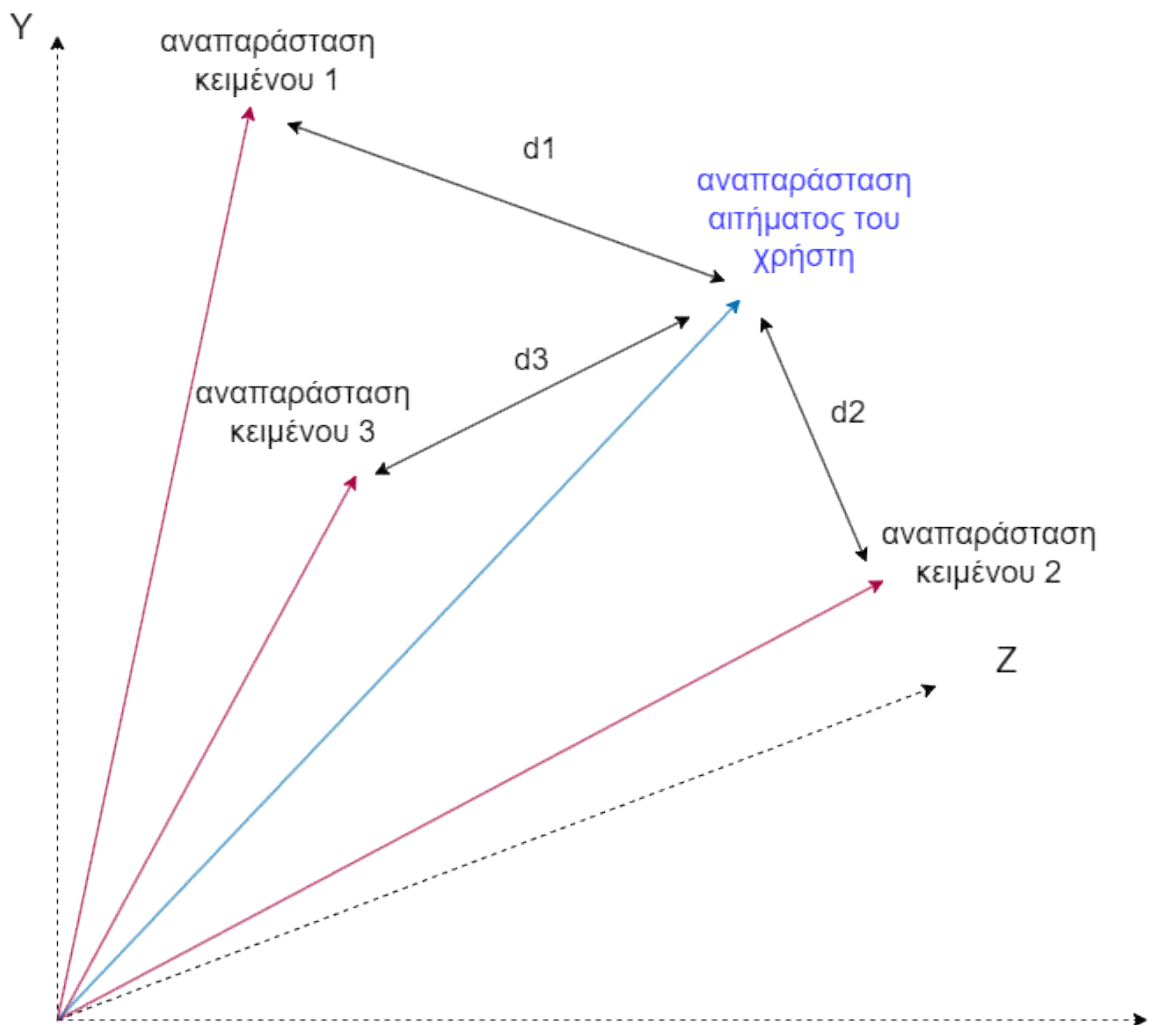
4.15.1 Ευκλείδεια Απόσταση (Euclidean Distance)

Η ευκλείδεια απόσταση είναι η πιο διαδεδομένη μέθοδος μέτρησης της απόστασης μεταξύ δύο σημείων σε έναν πολυδιάστατο χώρο. Υπολογίζεται ως η τετραγωνική ρίζα του αθροίσματος των τετραγώνων των διαφορών των συντεταγμένων τους. Αυτή η μέθοδος είναι απλή και εύκολα κατανοητή, καθώς και χρήσιμη για διανύσματα με συνεχείς τιμές. Ωστόσο, μπορεί να επηρεαστεί από την κλίμακα των δεδομένων και δεν λαμβάνει υπόψη την κατεύθυνση των διανυσμάτων.

Η ευκλείδεια απόσταση μεταξύ δύο σημείων (x_1, y_1) και (x_2, y_2) στο δισδιάστατο χώρο υπολογίζεται με τον ακόλουθο μαθηματικό τύπο:

$$\text{Ευκλείδεια Απόσταση} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Υπολογισμός Ευκλείδειας απόστασης



Σχ15. Στο παραπάνω σχήμα γίνεται αναπαράσταση του υπολογισμού της Ευκλείδειας απόστασης (Euclidean Distance), μεταξύ διαφορετικών διανυσμάτων στο χώρο. Παρατηρείται ότι, η απόσταση υπολογίζεται και ισούται με την απόλυτη απόσταση μεταξύ των σημείων των διανυσμάτων.

4.15.2 Ομοιότητα Συνημιτόνων (Cosine Similarity)

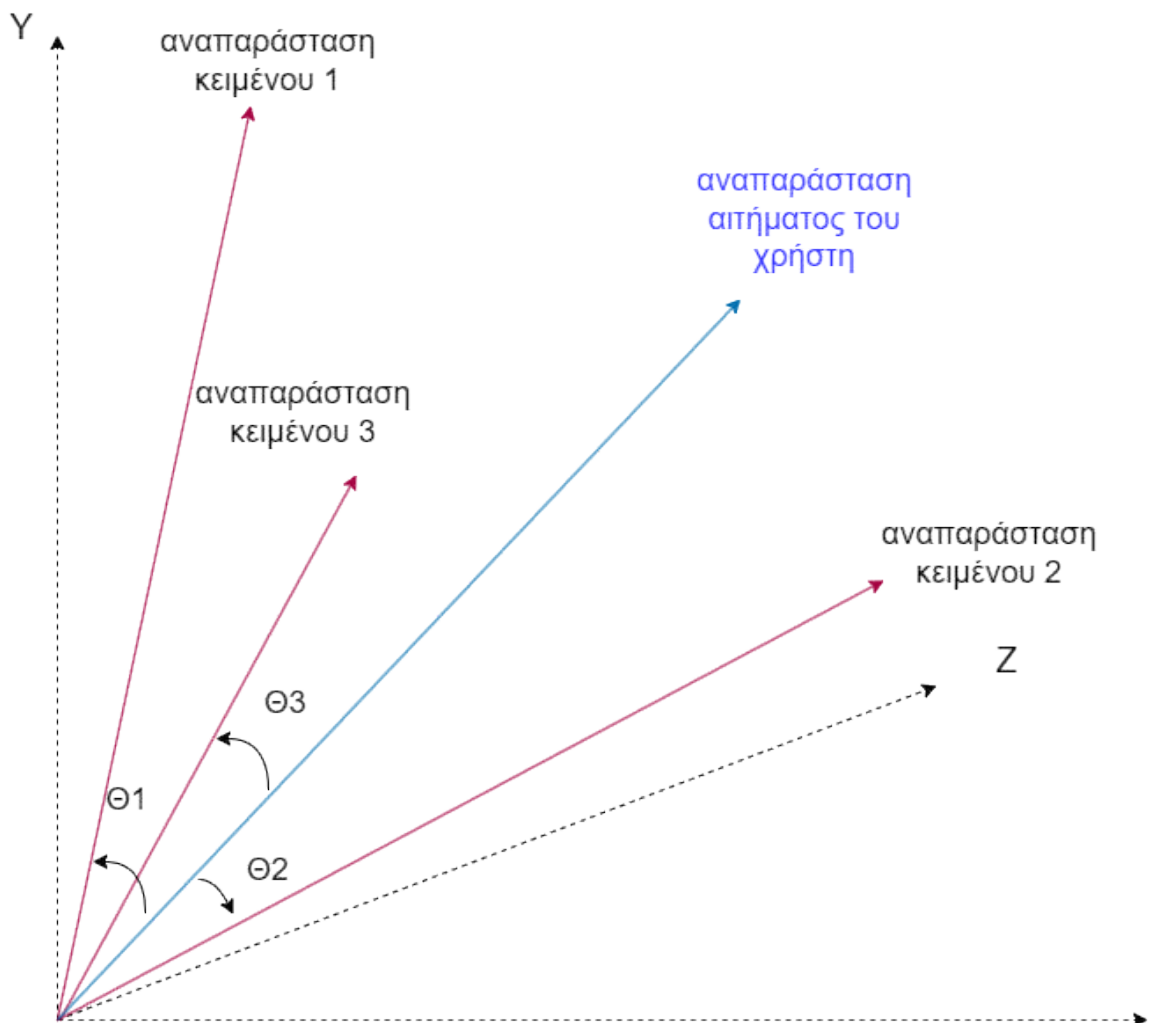
Η συνημιτονοειδής ομοιότητα μετράει τη γωνία μεταξύ δύο διανυσμάτων και όχι την απόστασή τους. Υπολογίζεται ως το εσωτερικό γινόμενο των διανυσμάτων διαιρεμένο με το γινόμενο των μεγεθών τους. Αυτή η μέθοδος είναι κατάλληλη για διανύσματα με διαφορετικά μεγέθη και δεν επηρεάζεται από την κλίμακα των δεδομένων. Ωστόσο, δεν λαμβάνει υπόψη τη διαφορά στο μέγεθος των διανυσμάτων.

Για την ομοιότητα συνημιτόνων, μπορούμε να χρησιμοποιήσουμε τη συνημιτονική ομοιότητα, η οποία ορίζεται ως το κανονικοποιημένο εσωτερικό γινόμενο δύο διανυσμάτων. Για δύο διανύσματα a , b η ομοιότητα συνημιτόνων υπολογίζεται ως εξής:

$$\text{Ομοιότητα Συνημιτόνων} = \frac{a \cdot b}{\|a\| \cdot \|b\|}$$

όπου $a \cdot b$ είναι το εσωτερικό γινόμενο των δύο διανυσμάτων και $\|a\|$ και $\|b\|$ είναι οι νόρμες των διανυσμάτων a και b αντίστοιχα.

Υπολογισμός Ομοιότητας Συνημιτόνου



Σχ16. Στο παραπάνω σχήμα γίνεται αναπαράσταση του υπολογισμού της Ομοιότητας Συνημιτόνου (Cosine Similarity), μεταξύ διαφορετικών διανυσμάτων στο χώρο. Παρατηρείται ότι, η απόσταση υπολογίζεται και ισούται με τη γωνία που σχηματίζεται μεταξύ των διαφορετικών διανυσμάτων.

4.15.3 Απόσταση Manhattan (Manhattan Distance ή L1 Distance)

Η απόσταση Manhattan υπολογίζει το άθροισμα των απόλυτων διαφορών των συντεταγμένων των διανυσμάτων. Αυτή η μέθοδος είναι απλή και ευνοεί τις ευθείες διαδρομές, και είναι χρήσιμη για διανύσματα με σπάνιες τιμές. Ωστόσο, μπορεί να μην αποτυπώνει σωστά την πραγματική απόσταση σε πολυδιάστατους χώρους.

4.15.4 Ομοιότητα Jaccard (Jaccard Similarity)

Η ομοιότητα Jaccard μετράει την ομοιότητα μεταξύ δύο συνόλων διαιρώντας το μέγεθος της τομής με το μέγεθος της ένωσης των συνόλων. Αυτή η μέθοδος είναι χρήσιμη για δυαδικά και σπάνια διανύσματα, αλλά δεν λαμβάνει υπόψη τη συχνότητα εμφάνισης των στοιχείων.

Για τη σύγκριση επιμέρους των διανυσματικών αναπαραστάσεων που έχουν προκύψει από τα κείμενα της περιγραφής των κινηματογραφικών ταινιών, προτιμούμε να επικεντρωθούμε σε δύο συγκεκριμένες μεθόδους απόστασης διανυσμάτων, αποφεύγοντας τη λεπτομερή ανάλυση των άλλων δύο μεθόδων.

Η απόσταση Manhattan, γνωστή και ως L1 απόσταση, υπολογίζει το άθροισμα των απόλυτων διαφορών των συντεταγμένων των διαφορετικών διανυσμάτων. Παρόλο που αυτή η μέθοδος είναι απλή, ενδέχεται να μην αποτυπώνει σωστά την πραγματική απόσταση σε πολυδιάστατους χώρους. Συνεπώς, δεν θα γίνει λεπτομερής αναφορά της συγκεκριμένης μεθόδου

Η ομοιότητα Jaccard υπολογίζει πόσο παρόμοια είναι δύο σύνολα, διαιρώντας τον αριθμό των κοινών στοιχείων με τον αριθμό των στοιχείων που υπάρχουν συνολικά στα δύο σύνολα. Αν και είναι χρήσιμη για δυαδικά διανύσματα, δεν λαμβάνει υπόψη τη συχνότητα εμφάνισης των στοιχείων. Επομένως, η χρήση αυτής της μεθόδου είναι περιορισμένη για τα ζητούμενα της συγκεκριμένης εργασίας.

Αντ' αυτού, θα εστιάσουμε στην ευκλείδεια απόσταση και την ομοιότητα συνημιτόνων, οι οποίες έχουν καλύτερη εφαρμογή στη συγκεκριμένη εργασία καθώς καθιστούν εφικτό να εξεταστεί η σχέση μεταξύ των διανυσμάτων, λαμβάνοντας υπόψη τόσο την απόστασή των διανυσμάτων όσο και την κατεύθυνσή τους ή τη γωνία που σχηματίζεται μεταξύ τους.

4.16 Επιλογή κατάλληλης μεθόδου υπολογισμού απόστασης διανυσμάτων

Η επιλογή της κατάλληλης μεθόδου απόστασης διανυσμάτων αποτελεί κρίσιμο βήμα στην σύγκριση διαφορετικών κειμένων μέσω της διανυσματικής τους αναπαράστασης. Κάθε μέθοδος συγκρίνει την ομοιότητα διανυσμάτων χρησιμοποιώντας διαφορετική μετρική απόστασης και ως αποτέλεσμα κάθε μέθοδος λύνει διαφορετικά προβλήματα ανάλογα με το ποια πτυχή της κοντινότητας θέλουμε να τονίσουμε ανάμεσα σε διαφορετικά διανύσματα.

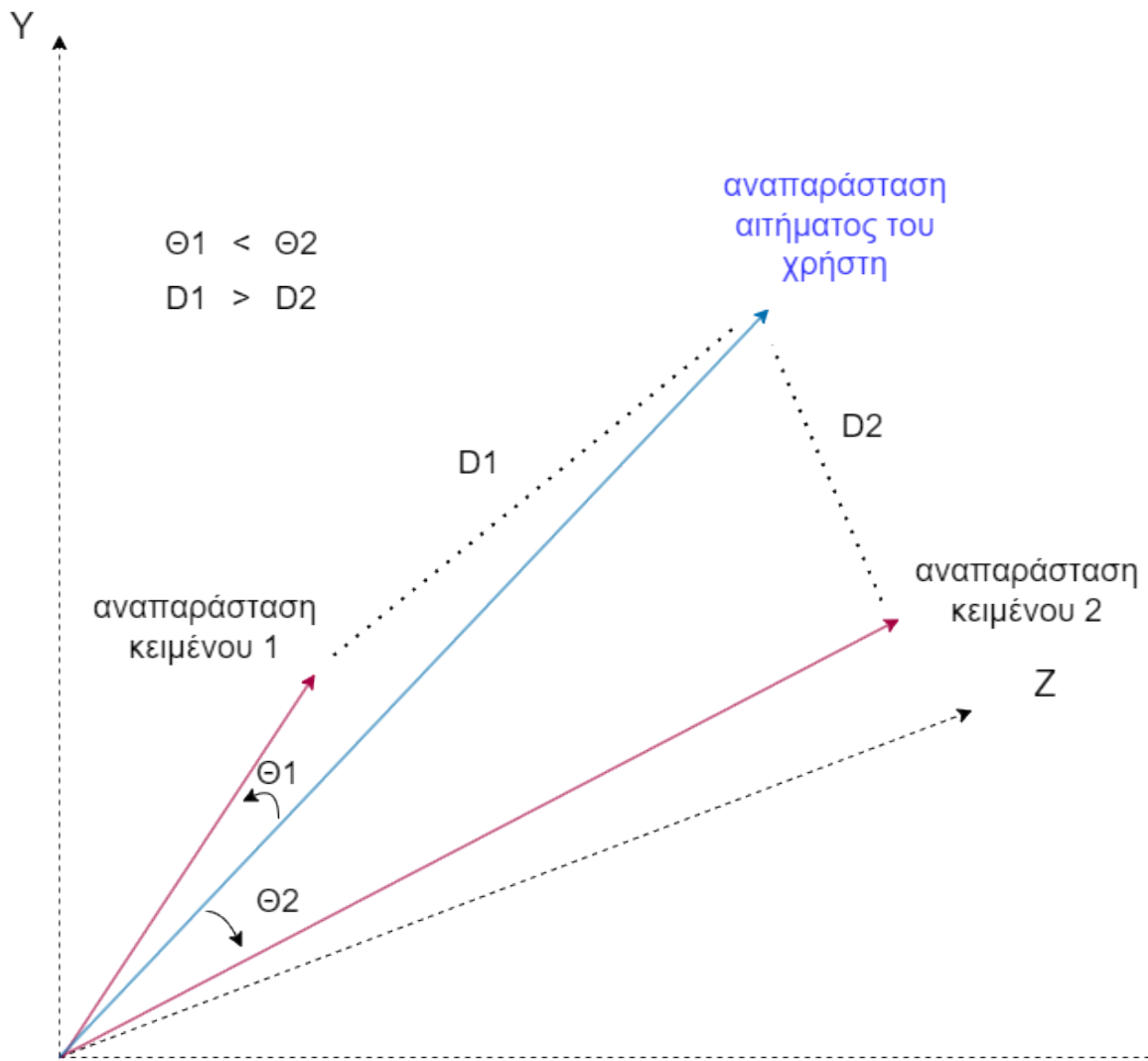
Η Ευκλείδεια απόσταση και η ομοιότητα συνημιτόνου είναι δύο διαφορετικές μετρικές που χρησιμοποιούνται για να αξιολογήσουν την κοντινότητα μεταξύ διανυσμάτων σε έναν πολυδιάστατο χώρο. Αν και οι δύο μετρικές μπορούν να χρησιμοποιηθούν για να μετρήσουν την απόσταση μεταξύ διανυσμάτων, διαφέρουν στον τρόπο που αξιολογούν την κοντινότητα ανάμεσα σε διαφορετικά διανύσματα.

Η Ευκλείδεια απόσταση υπολογίζει την απόλυτη απόσταση μεταξύ δύο σημείων στον χώρο, λαμβάνοντας υπόψη την απόσταση των συντεταγμένων τους. Αντίθετα, η ομοιότητα συνημιτόνου εστιάζει περισσότερο στη γωνία που σχηματίζεται μεταξύ των διαφορετικών διανυσμάτων στο χώρο, και δεν εστιάζει τόσο στην απόσταση των σημείων στον πολυδιάστατο χώρο. Αυτή η διαφορά επιτρέπει στην ομοιότητα συνημιτόνου να αντιληφθεί τη σημασιολογική σχέση μεταξύ των διανυσμάτων, ενώ η Ευκλείδεια απόσταση εστιάζει αποκλειστικά στην απόσταση μεταξύ των σημείων.

Η επιλογή αυτή είναι σημαντική κυρίως για εφαρμογές όπου η σημασιολογική ομοιότητα μεταξύ των κειμένων είναι πιο σημαντική από την απόλυτη απόστασή τους. Για παράδειγμα, στην ανάλυση των αιτημάτων των χρηστών ή στην ανάλυση κοντινών κειμένων, η συνημιτονοειδής ομοιότητα μπορεί να παρέχει πιο συνεκτικά αποτελέσματα, καθώς λαμβάνει υπόψη τη σημασιολογική συνάφεια των κειμένων ανεξαρτήτως της απόστασής τους στον πολυδιάστατο χώρο.

Η καλύτερη επιλογή ανάμεσα στις μεθόδους απόστασης διανυσμάτων που έχουν ήδη παρουσιαστεί για τη συγκεκριμένη εργασία είναι η συνημιτονοειδής ομοιότητα, καθώς δεν μετρά απλά την απόσταση μεταξύ των διανυσμάτων, αλλά εστιάζει στη γωνία που σχηματίζεται μεταξύ τους. Αυτό είναι σημαντικό επειδή η γωνία παρέχει μια πιο αντιπροσωπευτική ένδειξη της σημασιολογικής σχέσης μεταξύ των διανυσματικών αναπαραστάσεων του κειμένου. Αντί λοιπόν να εστιάζεται απλώς στην απόσταση, η συνημιτονοειδής ομοιότητα λαμβάνει υπόψη το πώς σχετίζονται τα κείμενα μεταξύ τους σημασιολογικά, προσφέροντας ένα πιο ολοκληρωμένο και αντιπροσωπευτικό μέτρο σύγκρισης.

Σύγκριση Συνημιτονοειδής Ομοιότητας και Ευκλείδειας Απόστασης



Σχ17. Στο σχήμα παρουσιάζεται η σύγκριση της απόστασης, με την χρήση των δύο μεθόδων που περιγράφονται παραπάνω. Η πρώτη μέθοδος είναι η Ευκλείδεια Απόσταση (Euclidean Distance), ενώ η δεύτερη είναι η Ομοιότητα Συνημιτόνων (Cosine Similarity). Παρατηρείται ότι για τα ίδια διανύσματα, οι αποστάσεις που υπολογίζονται είναι αρκετά διαφορετικές. Η Ευκλείδεια Απόσταση (Euclidean Distance), υπολογίζει την απόλυτη απόσταση μεταξύ των διανυσμάτων, χωρίς να λαμβάνει υπόψιν την κατεύθυνσή τους. Όπως διαπιστώνεται, διανύσματα πλησιέστερα μεταξύ τους ελαχιστοποιούν την ομοιότητα συνημιτόνων, ενώ η ευκλείδεια απόσταση παραμένει μεγάλη. Σε εφαρμογές που ζητούμενο είναι η σημασιολογία, η ομοιότητα συνημιτόνων προτιμάτε καθώς είναι πιο αποδοτική και παράγει καλύτερα αποτελέσματα.

4.17 Αποθήκευση των προ επεξεργασμένων δεδομένων

Όταν ολοκληρωθεί η προεπεξεργασία του συνόλου δεδομένων, είναι σημαντικό να αποθηκευτούν τα αποτελέσματα για μελλοντική χρήση, χωρίς την ανάγκη για επαναεπεξεργασία. Αυτό γίνεται με σκοπό να εξασφαλιστεί ότι τα δεδομένα θα είναι προσβάσιμα και θα μπορούν να χρησιμοποιηθούν από την τελική εφαρμογή διαδικτύου που αναπτύσσεται. Αυτή η εφαρμογή αξιοποιεί τα προ επεξεργασμένα δεδομένα για να προτείνει ταινίες στους χρήστες βάσει των πληροφοριών από το σύνολο δεδομένων και τα αιτήματά τους.

Η αποθήκευση των αποτελεσμάτων σε ενδιάμεσο αρχείο εξασφαλίζει την ευκολία χρήσης και διαχείρισης των δεδομένων από την τελική εφαρμογή. Έτσι, η εφαρμογή δεν χρειάζεται να επαναλάβει την προεπεξεργασία κάθε φορά που χρειάζεται νέες πληροφορίες, κάτι που βοηθά στην εξοικονόμηση χρόνου και πόρων από το σύστημα.

4.18 Ανασκόπηση Προεπεξεργασίας Δεδομένων

Στην εργασία για την ανάλυση και την προεπεξεργασία δεδομένων χρησιμοποιείται η γλώσσα προγραμματισμού Python. Η γλώσσα προγραμματισμού Python είναι ευρέως διαδεδομένη για τη διαχείριση και ανάλυση δεδομένων, καθώς προσφέρει μια πληθώρα βιβλιοθηκών που διευκολύνουν την υλοποίηση αυτών των διαδικασιών. Αρχικά, τα δεδομένα φορτώνονται και διαχειρίζονται χρησιμοποιώντας τη βιβλιοθήκη Pandas, η οποία επιτρέπει την αποτελεσματική ανάγνωση, καθαρισμό και χειρισμό μεγάλων συνόλων δεδομένων.

Για την αρχική προεπεξεργασία των δεδομένων, εφαρμόζουμε τεχνικές όπως η αφαίρεση στηλών που δεν θα χρειαστούν στην ανάλυση, η αντικατάσταση κενών τιμών και η μετατροπή στηλών σε αριθμητική μορφή για δεδομένα ημερομηνιών και αξιολογήσεων. Η βιβλιοθήκη Pandas συνδυάζεται με τη βιβλιοθήκη Numpy παρέχει αποδοτικές και γρήγορες μαθηματικές λειτουργίες σε σύνολα δεδομένων.

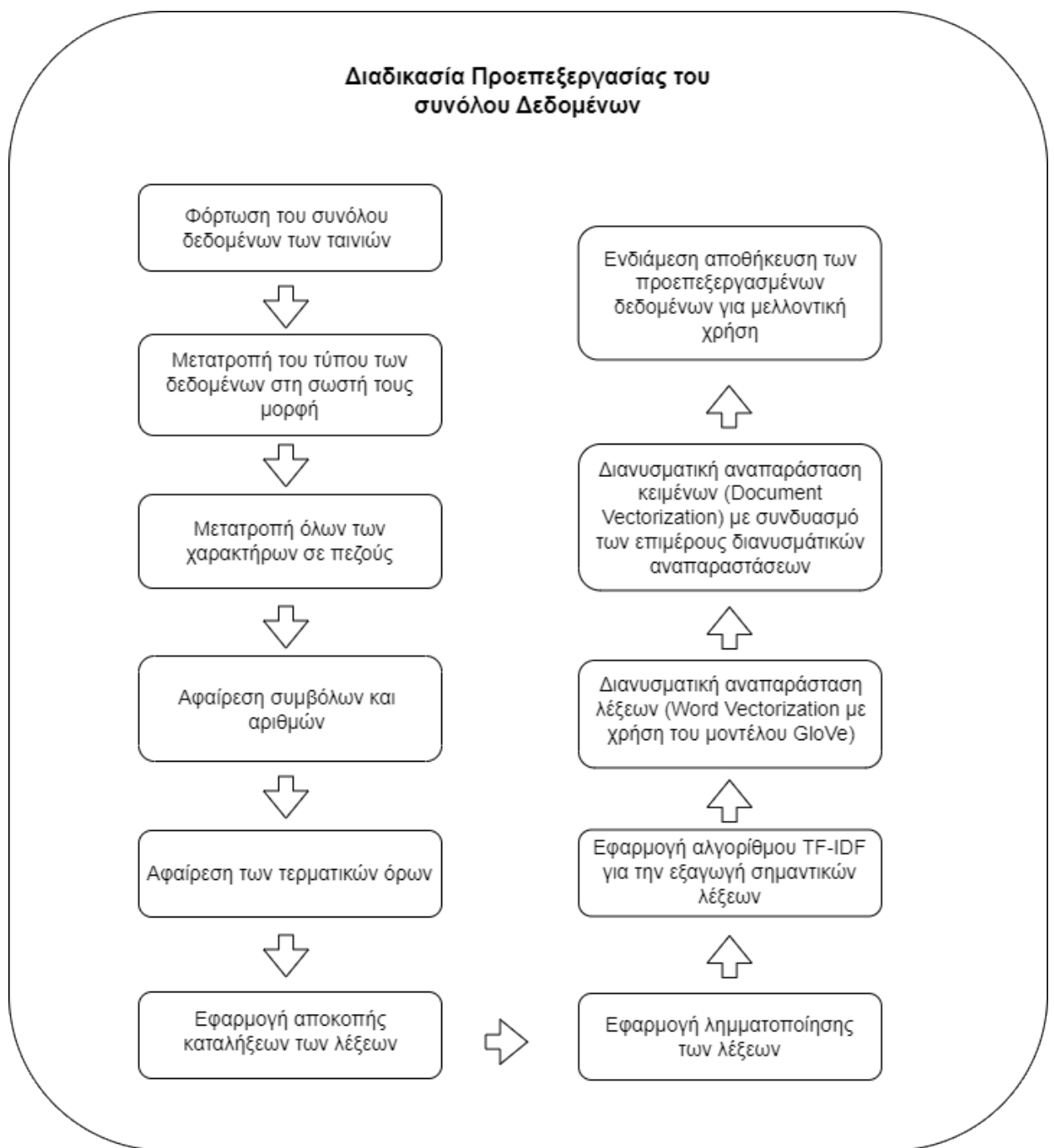
Στη συνέχεια, χρησιμοποιούμε τη βιβλιοθήκη NLTK (Natural Language Toolkit) για την επεξεργασία των περιγραφών των ταινιών. Η διαδικασία αυτή περιλαμβάνει τον διαχωρισμό του κειμένου σε λέξεις, την αφαίρεση των τερματικών όρων (Stopwords) την εφαρμογή λημματοποίησης (Lemmatization) και της εφαρμογής αποκοπής καταλήξεων των λέξεων (Stemming) στα δεδομένα φυσικής γλώσσας (Natural Language), την αφαίρεση μη αλφαριθμητικών χαρακτήρων και τη μετατροπή όλων των χαρακτήρων σε πεζούς.

Για την εξαγωγή των σημαντικότερων λέξεων από τις περιγραφές των ταινιών, χρησιμοποιείται ο αλγόριθμος TF-IDF (Term Frequency-Inverse Document Frequency). Η βιβλιοθήκη Scikit-learn παρέχει τα εργαλεία για τον υπολογισμό του TF-IDF, βοηθώντας μας να εντοπίσουμε τις λέξεις που είναι πιο σημαντικές για την κατανόηση του κειμένου. Με αυτόν τον τρόπο, μειώνεται ο όγκος των δεδομένων, ενώ διατηρείται η ουσιαστική πληροφορία.

Ο αλγόριθμος GloVe (Global Vectors for Word Representation) χρησιμοποιείται για την αναπαράσταση των λέξεων ως διανύσματα σε πολυδιάστατο χώρο. Η βιβλιοθήκη Gensim παρέχει προεκπαιδευμένα μοντέλα για την διανυσματική αναπαράσταση λέξεων με τη χρήση του GloVe. Αυτή η αναπαράσταση βοηθά στη αξιοποίηση των σημασιολογικών σχέσεων μεταξύ των λέξεων, καθιστώντας τη διαδικασία ανάλυσης κειμένου πιο αποδοτική και εύκολη.

Τέλος, η βιβλιοθήκη Scikit-learn παρέχει επίσης τα εργαλεία για τον υπολογισμό της ομοιότητας μεταξύ των διανυσμάτων χρησιμοποιώντας μεθόδους όπως η ομοιότητα συνημιτόνου και η ευκλείδεια απόσταση. Αυτές οι μετρικές μας επιτρέπουν να συγκρίνουμε τις περιγραφές των κινηματογραφικών ταινιών βάση των διανυσματικών τους αναπαραστάσεων και την διανυσματική ομοιότητα των περιγραφών τους.

Με τη συνδυαστική χρήση αυτών των βιβλιοθηκών και τεχνικών, είναι εφικτή η ανάπτυξη ενός συστήματος για την προεπεξεργασία και την ανάλυση των δεδομένων των κινηματογραφικών ταινιών, διευκολύνοντας την ανάπτυξη της εφαρμογής που θα περιγράψουμε παρακάτω για την πρόταση των ταινιών στους χρήστες.



Σχ18. Στο σχήμα παρουσιάζεται πιο αναλυτικά η αρχιτεκτονική και τα τμήματα του συστήματος προ επεξεργασίας των δεδομένων. Αναγράφονται όλα τα βήματα που έχουν παρουσιαστεί παραπάνω για την προεπεξεργασία του συνόλου δεδομένων.

Κεφάλαιο 5 - Ανάπτυξη Διαδικτυακής Εφαρμογής για Προτάσεις Ταινιών

5.1 Εισαγωγή

Στο προηγούμενο κεφάλαιο έγινε περιγραφή της προεπεξεργασίας του συνόλου δεδομένων των κινηματογραφικών ταινιών που θα χρησιμοποιηθούν από την διαδικτυακή εφαρμογή προτάσεων ταινιών. Στο κεφάλαιο αυτό θα παρουσιαστούν αναλυτικά τα βήματα και η διαδικασία για την ανάπτυξη της διαδικτυακής εφαρμογής, η οποία θα αξιοποιεί τα δεδομένα αυτά και θα λειτουργεί ως διεπαφή με τους τελικούς χρήστες, προτείνοντας κινηματογραφικές ταινίες σύμφωνα με τα αιτήματά τους.

Ο στόχος σε αυτό το στάδιο της εργασίας είναι η ανάπτυξη μιας διαδραστικής εφαρμογής που επιτρέπει στους χρήστες να εισάγουν σε μορφή ελεύθερου κειμένου, το ερώτημα ή αλλιώς το αίτημα τους ώστε να περιγράψουν την ταινία που επιθυμούν να παρακολουθήσουν. Η εφαρμογή δεν χρησιμοποιεί συγκεκριμένες, προκαθορισμένες ερωτήσεις για να καθορίσει τις προτιμήσεις των χρηστών. Αντίθετα, βασίζεται στην ανάλυση του κειμένου που παρέχουν οι χρήστες, προτείνοντας ταινίες που ανταποκρίνονται καλύτερα στις προτιμήσεις και τις ανάγκες τους.

5.2 Περιγραφή λειτουργίας της εφαρμογής

Η εφαρμογή θα πρέπει να επεξεργάζεται το κείμενο που εισάγει ο χρήστης, εξάγοντας λέξεις κλειδιά και βασικές πληροφορίες από το αίτημα του χρήστη. Με βάση αυτές τις πληροφορίες, θα προτείνει ταινίες που ταιριάζουν στις προτιμήσεις και ανάγκες του χρήστη. Έτσι, η εφαρμογή επιτυγχάνει την παροχή προτάσεων ταινιών που είναι ουσιαστικά προσαρμοσμένες στην περιγραφή του κάθε χρήστη.

Η αλληλεπίδραση με τον χρήστη θα ξεκινάει με μια γενική ερώτηση, ζητώντας από τον χρήστη να περιγράψει ελεύθερα το είδος της ταινίας που επιθυμεί να δει. Η επικοινωνία θα πραγματοποιείται μέσω ζωντανής συνομιλίας, όπου ο χρήστης θα παρέχει το αίτημά του και η εφαρμογή θα προτείνει μια λίστα ταινιών που ανταποκρίνονται στις προτιμήσεις του, σύμφωνα με τις πληροφορίες που έχει δώσει.

Αφού ο χρήστης λάβει τις αρχικές προτάσεις, θα έχει τη δυνατότητα να προσθέσει επιπλέον πληροφορίες στην αρχική του αίτηση, επιτρέποντας την περαιτέρω βελτίωση και προσαρμογή των προτάσεων. Η συνομιλία θα γίνεται σε πραγματικό χρόνο, δίνοντας στον χρήστη τη δυνατότητα

να συνεχίσει να ρωτάει και να καθορίζει επιπλέον χαρακτηριστικά που επιθυμεί να έχουν οι προτεινόμενες ταινίες.

Αυτή η διαδραστική επικοινωνία σε πραγματικό χρόνο επιτρέπει στον χρήστη να λαμβάνει πιο ακριβείς και εξατομικευμένες προτάσεις, βελτιώνοντας τη συνολική εμπειρία αναζήτησης και επιλογής ταινιών. Με κάθε απάντηση, ο χρήστης μπορεί να προσαρμόζει τις παραμέτρους της αναζήτησής του, οδηγώντας σε συνεχώς βελτιωμένες προτάσεις που ανταποκρίνονται καλύτερα στις προτιμήσεις του.

5.3 Αρχιτεκτονική της εφαρμογής διαδικτύου

Η εφαρμογή ξεκινάει φορτώνοντας τα δεδομένα που έχουν προ επεξεργαστεί στο προηγούμενο στάδιο, όπως περιγράφηκε στην ενότητα της προεπεξεργασίας δεδομένων. Αυτό σημαίνει ότι η προεπεξεργασία των δεδομένων γίνεται μία φορά και δεν απαιτείται να επαναληφθεί κάθε φορά που ξεκινάει η εφαρμογή. Κατά την έναρξή της, διαβάζει τα προ επεξεργασμένα δεδομένα που έχουν ήδη ετοιμαστεί και αποθηκευτεί, ώστε να είναι έτοιμη να τα χρησιμοποιήσει όταν απαιτηθεί από τον χρήστη. Αυτή η διαδικασία διασφαλίζει ότι η εφαρμογή είναι άμεσα λειτουργική και έτοιμη να ανταποκριθεί στα αιτήματα του χρήστη και να προτείνει κινηματογραφικές ταινίες βάσει των κριτηρίων που ορίζει ο χρήστης.

Στη συνέχεια, η εφαρμογή περιλαμβάνει μια σελίδα συνομιλίας σε μορφή ζωντανής συνομιλίας, όπου ο χρήστης μπορεί να υποβάλει τα αιτήματά του μέσω μιας φόρμας. Αυτή η σελίδα λειτουργεί ως διεπαφή μεταξύ του χρήστη και του συστήματος της εφαρμογής, επιτρέποντας την ελεύθερη επικοινωνία και την υποβολή ερωτημάτων σε πραγματικό χρόνο. Στη σελίδα ζωντανής συνομιλίας, τα ερωτήματα του χρήστη εμφανίζονται στα δεξιά και οι απαντήσεις του συστήματος στα αριστερά, όπως ακριβώς συμβαίνει σε μια τυπική συνομιλία. Με αυτόν τον τρόπο, ο χρήστης μπορεί να περιγράψει το είδος της ταινίας που επιθυμεί να παρακολουθήσει, και η εφαρμογή να προτείνει ταινίες που ανταποκρίνονται στις προτιμήσεις του.

Η εφαρμογή περιλαμβάνει επίσης μια Διεπαφή Προγραμματισμού Εφαρμογών (Application Programming Interface ή REST API), η οποία διαχειρίζεται την επικοινωνία μεταξύ της σελίδας συνομιλίας που βλέπει ο χρήστης και του συστήματος προτάσεων ταινιών. Όταν ο χρήστης υποβάλλει ένα αίτημα μέσω της σελίδας ζωντανής συνομιλίας, η διεπαφή προγραμματισμού λαμβάνει το αίτημα, το επεξεργάζεται και απαντά στον χρήστη. Συγκεκριμένα, η διεπαφή

προγραμματισμού (REST API) δέχεται τα δεδομένα εισόδου που περιλαμβάνουν το ερώτημα του χρήστη και εκτελεί μια σειρά λειτουργιών για την εύρεση των κινηματογραφικών ταινιών που θα προτείνει στον χρήστη.

Μετά την επεξεργασία του αιτήματος, η διεπαφή προγραμματισμού (REST API) επιστρέφει τα αποτελέσματα στην εφαρμογή. Οι προτεινόμενες ταινίες αποστέλλονται πίσω στη σελίδα συνομιλίας, όπου εμφανίζονται στον χρήστη ως απαντήσεις. Αυτή η διαδικασία επιτρέπει στην εφαρμογή να ανταποκρίνεται σε πραγματικό χρόνο στα αιτήματα του χρήστη, εξασφαλίζοντας μια ομαλή και αποδοτική ροή δεδομένων μεταξύ της διεπαφής χρήστη και του συστήματος προτάσεων. Αυτό βελτιώνει τη συνολική εμπειρία του χρήστη και καθιστά τη διαδικασία αναζήτησης ταινιών γρήγορη και εύκολη.

Τέλος, η εφαρμογή περιλαμβάνει τις συναρτήσεις που εκτελούν τη διαδικασία εύρεσης των πιο πλησιέστερων κινηματογραφικών ταινιών βάσει των ερωτημάτων ή αιτήσεων του χρήστη. Αυτές οι συναρτήσεις χρησιμοποιούν διάφορα μοντέλα και μεθόδους που έχουν ήδη παρουσιαστεί, όπως η επεξεργασία φυσικής γλώσσας (Natural Language Processing) και η διανυσματική αναπαράσταση του αιτήματος του χρήστη με τη χρήση του GloVe (Global Vectors for Word Representation).

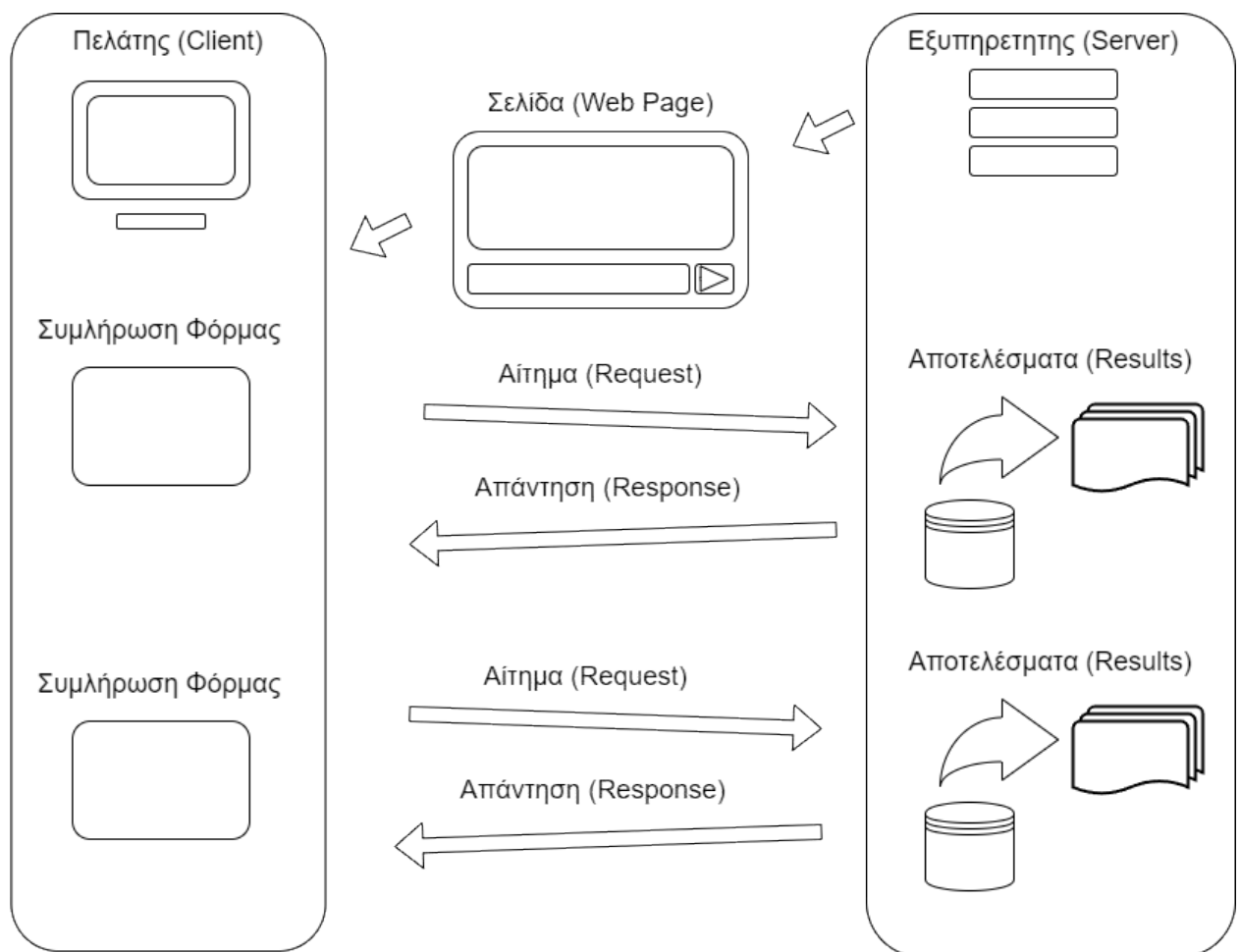
Το αίτημα του χρήστη υποβάλλεται επίσης σε επεξεργασία φυσικής γλώσσας (Natural Language Processing ή NLP), όπως έχουμε ήδη δει σε προηγούμενο κεφάλαιο, με τα εξής βήματα:

- Μετατροπή των χαρακτήρων σε πεζούς (Lowercase)
- Αφαίρεση Συμβόλων και Αριθμών
- Αφαίρεση των τερματικών όρων (Stopwords)
- Εφαρμογή αποκοπής καταλήξεων (Stemming)
- Εφαρμογή λημματοποίησης (Lemmatization)
- Διανυσματική αναπαράσταση λέξεων (Word Vectorization)
- Διανυσματική αναπαράσταση του αιτήματος (Query Vectorization)

Με τη χρήση αυτών των συναρτήσεων, η εφαρμογή μετατρέπει την αίτηση του χρήστη, που δίνεται σε φυσική γλώσσα, σε μια διανυσματική αναπαράσταση, όπως ακριβώς έχει πραγματοποιηθεί στις περιγραφές των κινηματογραφικών ταινιών. Αυτή η διαδικασία επιτρέπει στο σύστημα να εφαρμόσει μεθόδους και αλγορίθμους για να βρει τις πλησιέστερες κινηματογραφικές ταινίες βάσει του ερωτήματος του χρήστη.

Οι συναρτήσεις που περιγράφουμε, καλούνται από την εφαρμογή αφού λάβει αίτημα μέσω της Διεπαφής Προγραμματισμού Εφαρμογών (REST API). Όταν ο χρήστης υποβάλλει ένα αίτημα μέσω της σελίδας ζωντανής συνομιλίας, το REST API λαμβάνει το αίτημα, καλεί τις συναρτήσεις που προ επεξεργάζονται το αίτημα και εντοπίζουν τις σχετικές ταινίες. Στη συνέχεια, η διεπαφή προγραμματισμού απαντάει στον χρήστη με τα αποτελέσματα.

Παρακάτω παρουσιάζεται ένα διάγραμμα της ροής εκτέλεσης της εφαρμογής για την πρόταση κινηματογραφικών ταινιών:



Σχ19. Στο διάγραμμα παρουσιάζεται μια ενδεικτική ροή λειτουργίας του συστήματος πρότασης ταινιών. Αρχικά παρατηρούμε ότι ο χρήστης λαμβάνει την σελίδα της διαδικτυακής εφαρμογής.

Στη συνέχεια, μέσω αυτής υποβάλλει τα αιτήματά του, τα οποία το σύστημα δέχεται, επεξεργάζεται και του επιστρέφει αποτελέσματα σχετικά με το αίτημά του.

5.4 Ενοποίηση διαδοχικών αιτημάτων του χρήστη

Η εφαρμογή προτάσεων, όπως έχει ήδη παρουσιαστεί, λειτουργεί ως μια διαδραστική πλατφόρμα συνομιλίας, όπου ο χρήστης μπορεί να υποβάλει ένα αίτημα σχετικά με το είδος των ταινιών που επιθυμεί να δει. Αντίστοιχα, η εφαρμογή ανταποκρίνεται παρέχοντας ένα σύνολο κινηματογραφικών ταινιών που είναι σχετικές με το αίτημα που έχει υποβάλλει ο χρήστης. Κατόπιν, ο χρήστης έχει τη δυνατότητα να υποβάλλει ένα νέο αίτημα και να λάβει εκ νέου μια νέα πρόταση για ταινίες που ταιριάζουν στο νέο ερώτημά του. Ωστόσο, το νέο αίτημα του χρήστη μπορεί να χρησιμοποιηθεί σε συνδυασμό με το προηγούμενο για να το εμπλουτίσει και να παρέχει επιπλέον πληροφορία στο αρχικό του αίτημα ή την αρχική του ερώτηση. Αντί να χρησιμοποιηθεί μόνο το νέο αίτημα ως βάση για να προτείνει ταινίες, υπάρχει η δυνατότητα να χρησιμοποιηθεί για να επεκτείνει το προηγούμενο αίτημα, προσφέροντας έτσι μια πιο πλήρη και ολοκληρωμένη αναζήτηση.

Ένα παράδειγμα ξεκαθαρίζει τη διαδικασία αυτής της λειτουργίας. Έστω ότι ο χρήστης ξεκινά με το αίτημα "Θέλω να παρακολουθήσω μια ταινία για τον μεσαιωνικό κόσμο" και του προτείνονται μερικές ταινίες που ταιριάζουν στην περιγραφή αυτή. Στη συνέχεια, αν ο χρήστης δώσει ένα νέο αίτημα "Οι ταινίες πρέπει να αφορούν τον πόλεμο", τότε το αρχικό αίτημα για τον μεσαιωνικό κόσμο δεν θα ληφθεί υπόψη κατά την αναζήτηση ταινιών που αφορούν τον πόλεμο. Αντίθετα, θα γίνει μια νέα άμεση αναζήτηση βάσει του νέου ερωτήματος για ταινίες που αντιστοιχούν στη θεματική του πολέμου αγνοώντας το αίτημα που σχετίζεται με το μεσαιωνικό κόσμο.

Η μέθοδος ενοποίησης και επέκτασης των αιτημάτων του χρήστη είναι ένας σημαντικός μηχανισμός που επιτρέπει στην εφαρμογή να εξάγει περισσότερη πληροφορία από τη συνομιλία της με τον χρήστη. Όταν ο χρήστης δίνει ένα αρχικό αίτημα, το οποίο μπορεί να είναι, για παράδειγμα, "δείξε μου ταινίες επιστημονικής φαντασίας", η εφαρμογή αντιλαμβάνεται το αίτημα και επεξεργάζεται τη φυσική γλώσσα του χρήστη για να το μετατρέψει σε ένα συγκεκριμένο ερώτημα προς τη βάση δεδομένων ταινιών. Σε περίπτωση που ο χρήστης συνεχίσει τη συνομιλία δίνοντας περισσότερα αιτήματα, τότε αυτά τα νέα αιτήματα συνδυάζονται με το αρχικό για να δημιουργήσουν ένα ευρύτερο και πιο στοχευμένο ερώτημα. Για παράδειγμα, αν ο χρήστης συνεχίσει με "δείξε μου ταινίες που σχετίζονται με το διάστημα", τότε το νέο αίτημα συνδυάζεται

με το προηγούμενο, επεκτείνοντας έτσι το ερώτημα και προσθέτοντας περισσότερη πληροφορία στο αρχικό ερώτημα για την αναζήτηση κινηματογραφικών ταινιών.

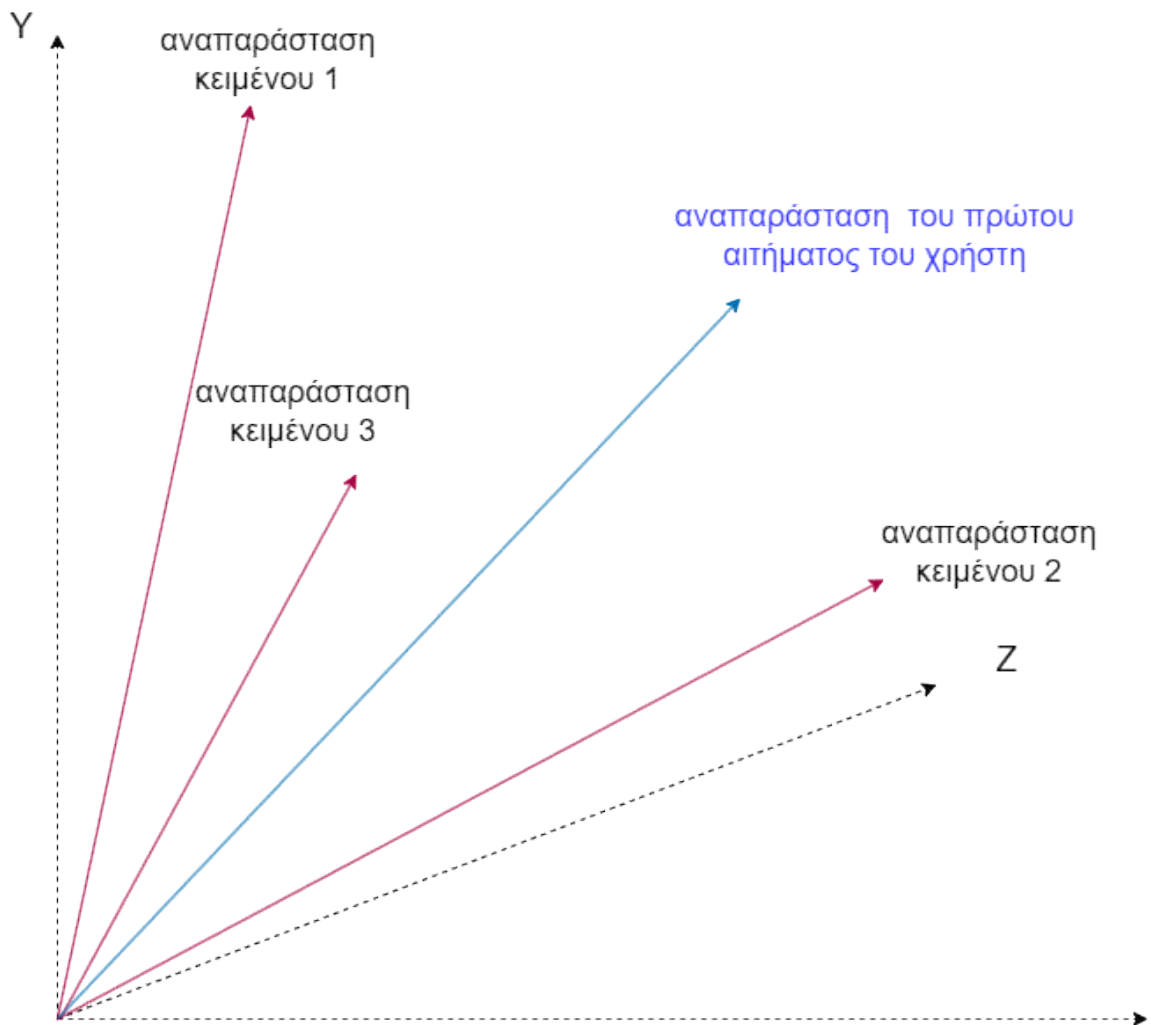
Μέσω αυτής της διαδικασίας, η εφαρμογή συλλέγει σταδιακά όλα τα αιτήματα του χρήστη και τα συνδυάζει για να δώσει μία πιο ολοκληρωμένη και ακριβή εικόνα των προτιμήσεών του. Καθώς οι συνεχείς αναφορές του χρήστη εμπλουτίζουν το αρχικό αίτημα με περισσότερες πληροφορίες, το σύστημα μπορεί να παράγει πιο στοχευμένες και εξατομικευμένες προτάσεις ταινιών που να ανταποκρίνονται ακριβώς στις ανάγκες και τις προτιμήσεις του χρήστη και των ερωτημάτων του.

Για την προσαρμογή στις απαιτήσεις της εργασίας, ορίσαμε το παράθυρο ή το πλαίσιο των ερωτημάτων να περιορίζεται στα τελευταία 4 ερωτήματα. Αυτό το περιορισμένο πλήθος ερωτημάτων βοηθά στη διατήρηση μιας συγκεκριμένης ποσότητας πληροφοριών που λαμβάνονται υπόψη για την αναζήτηση κινηματογραφικών ταινιών. Με αυτόν τον τρόπο, εξασφαλίζεται ότι οι προηγούμενες προτάσεις ταινιών δεν χάνονται πλήρως και μπορούν να εξακολουθήσουν να επηρεάζουν τη συνολική αναζήτηση. Επιπλέον, ο περιορισμός σε μια συγκεκριμένη ποσότητα ερωτημάτων εξασφαλίζει την καλύτερη απόκριση της εφαρμογής σε κάθε συνεχή αλληλεπίδραση με τον χρήστη, διατηρώντας τη συνολική διαδικασία αποτελεσματική και αποδοτική.

Κατά την προεπεξεργασία των αιτημάτων του χρήστη στο σύστημα, η εφαρμογή χρησιμοποιεί μεθόδους επεξεργασίας φυσικής γλώσσας για να κατανοήσει το περιεχόμενο του κειμένου. Στη συνέχεια, αυτό το κείμενο μετατρέπεται σε ένα διάνυμα αναπαράστασης του αιτήματος, όπως έχει ήδη περιγραφεί στην προεπεξεργασία των δεδομένων. Με αυτήν την μέθοδο, το αρχικό διάνυμα αναπαράστασης επεκτείνεται με την προσθήκη νέων ερωτημάτων, λαμβάνοντας υπόψη το περιεχόμενο αυτών των νέων ερωτημάτων σε συνδυασμό με τα προηγούμενα. Αυτή η διαδικασία ουσιαστικά δημιουργεί ένα επαυξημένο διάνυμα αναπαράστασης, το οποίο περιέχει πληροφορίες από τόσο το αρχικό αίτημα όσο και από τα νέα ερωτήματα, επιτρέποντας έτσι στην εφαρμογή να παράγει πιο στοχευμένες και εξατομικευμένες απαντήσεις ή προτάσεις βασισμένα σε περισσότερα ερωτήματα συνδυαστικά.

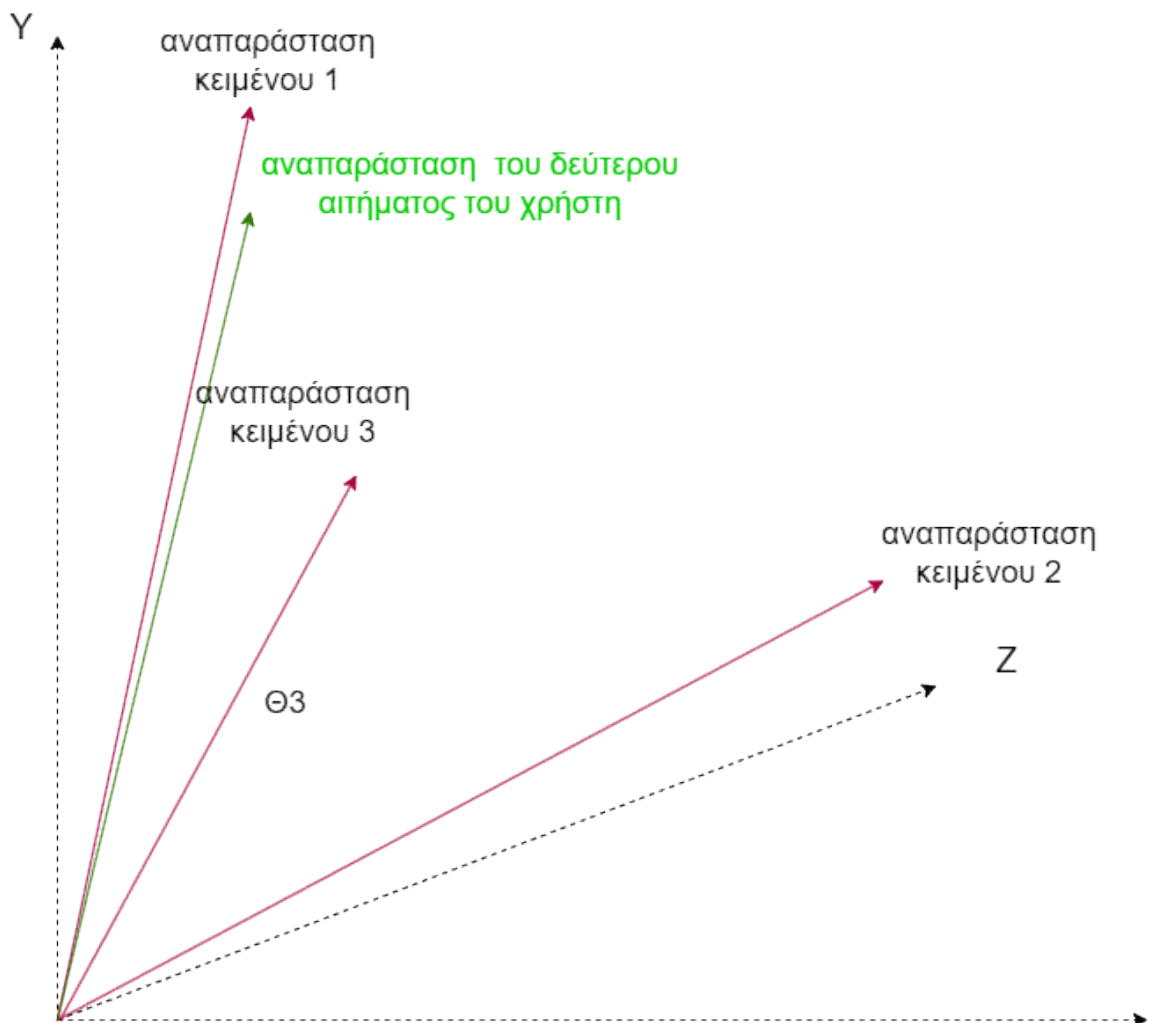
.Αυτή η προσέγγιση επιτρέπει στην εφαρμογή να προσεγγίσει καλύτερα τις κινηματογραφικές ταινίες που είναι πλησιέστερες στο ερώτημα του χρήστη, καθώς λαμβάνονται όλα τα ερωτήματα υπόψη. Αυτό διασφαλίζει ότι η σημασιολογική πληροφορία που περιέχεται σε κάθε αίτημα αναδεικνύεται πλήρως στο διάνυμα αναπαράστασης. Καθώς τα νέα ερωτήματα συνδυάζονται με το αρχικό, το διάνυμα εμπλουτίζεται με περισσότερες πληροφορίες, επιτρέποντας στην εφαρμογή να προτείνει ταινίες που πληρούν πιο ακριβώς τις ανάγκες και τις προτιμήσεις του χρήστη. Με αυτόν τον τρόπο, επιτυγχάνεται μια πιο εξειδικευμένη και προσαρμοσμένη πρόταση ταινιών, βασισμένη στο συνολικό περιεχόμενο της συνομιλίας με τον χρήστη.

Υπολογισμός Αναπαράστασης Αιτήματος Χρήστη



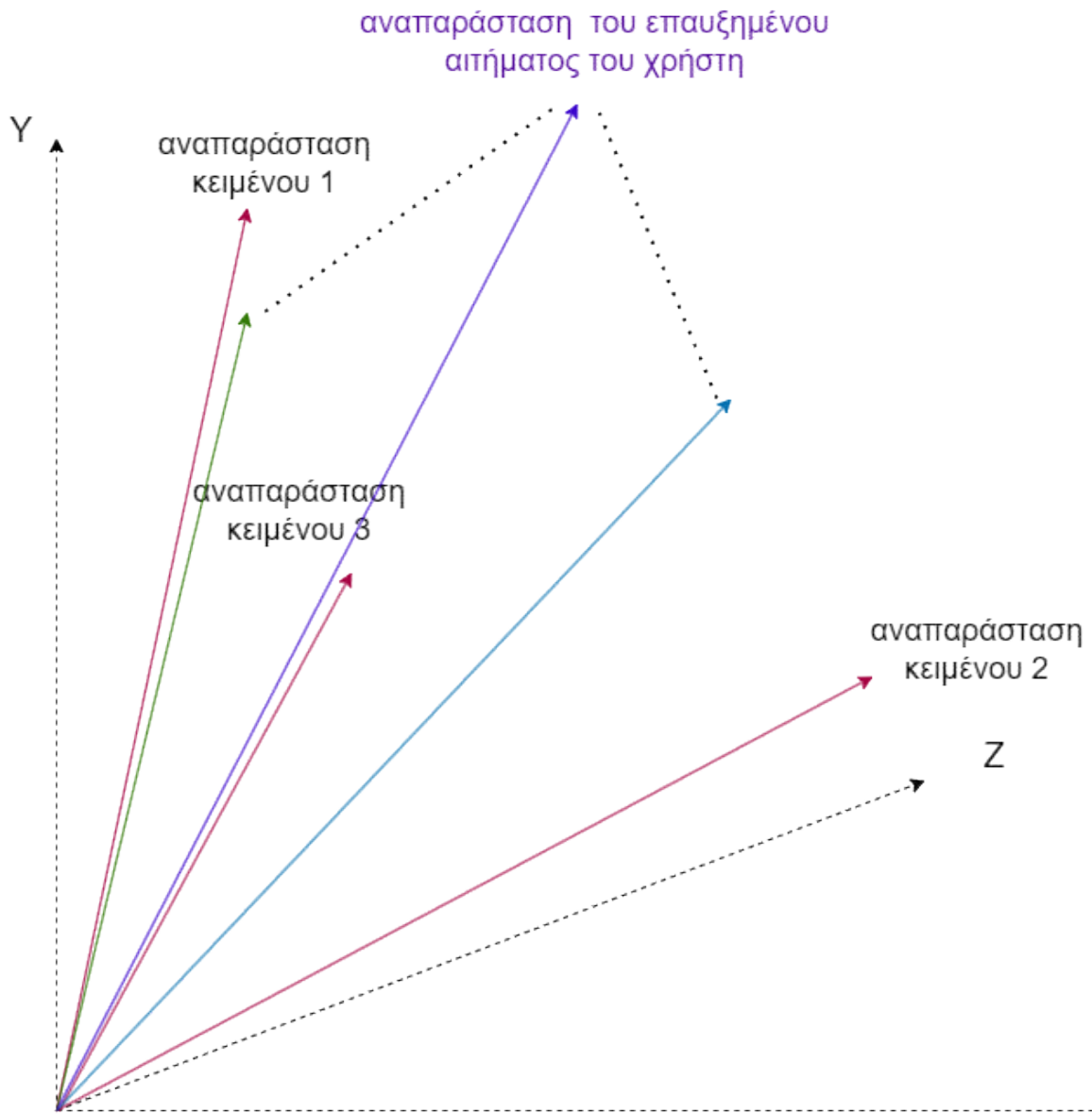
Σχ20. Στο παραπάνω σχήμα παρουσιάζεται ενδεικτικά η διανυσματική αναπαράσταση μερικών κειμένων από κινηματογραφικές ταινίες, καθώς η διανυσματική αναπαράσταση ενός αιτήματός του χρήστη. Στο συγκεκριμένο σχήμα αναφερόμαστε στο πρώτο αίτημα ή ερώτημα του χρήστη.

Υπολογισμός Αναπαράστασης Αιτήματος Χρήστη



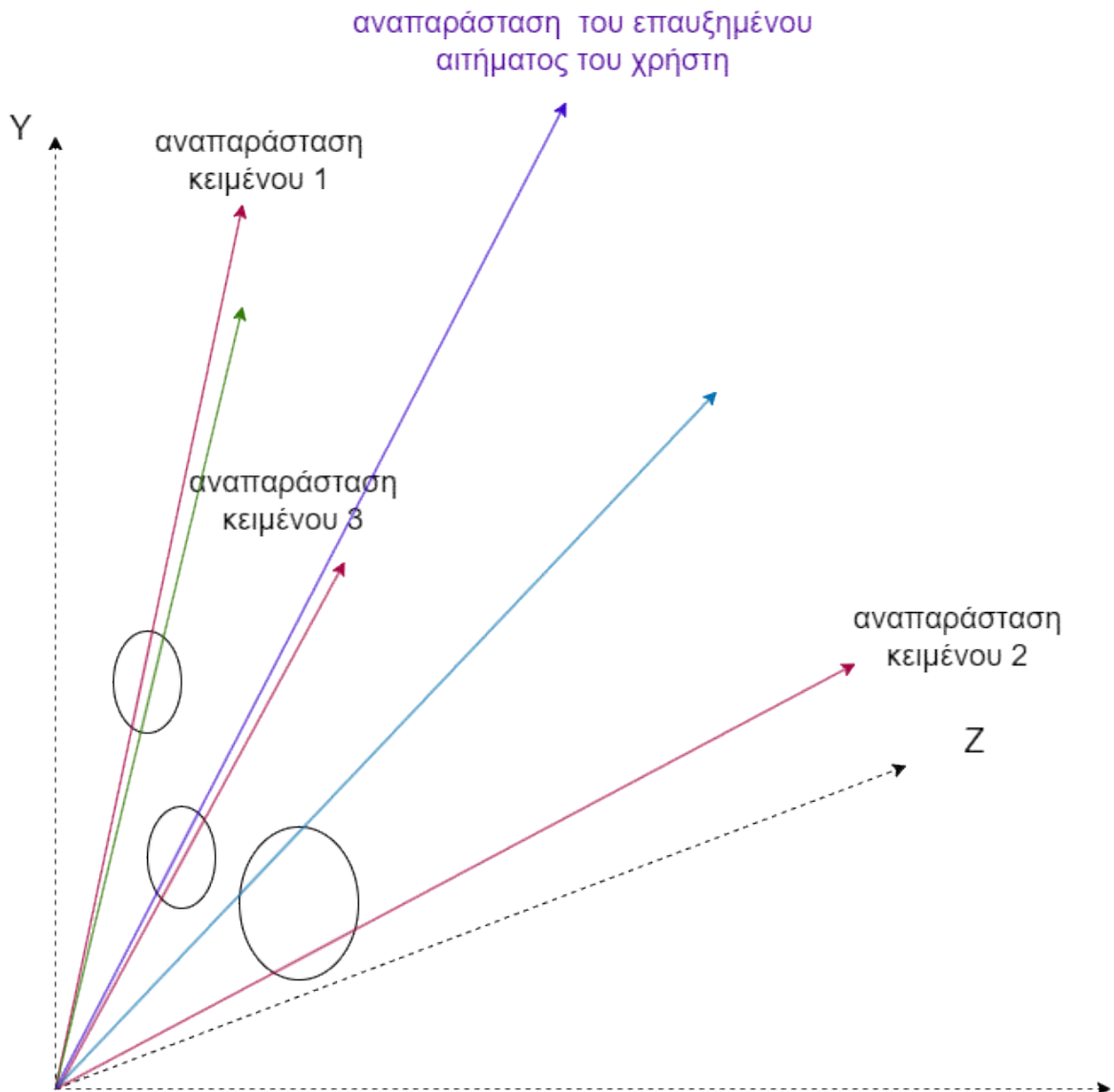
Σχ21. Στο παραπάνω σχήμα παρουσιάζεται ενδεικτικά η διανυσματική αναπαράσταση μερικών κειμένων από κινηματογραφικές ταινίες, καθώς και η διανυσματική αναπαράσταση ενός άλλου αιτήματός του χρήστη. Στο συγκεκριμένο σχήμα αναφερόμαστε σε επόμενα αιτήματα ή ερωτήματα του χρήστη. Παρατηρούμε ότι η διανυσματική αναπαράσταση των μετέπειτα ερωτημάτων δεν περιλαμβάνουν πληροφορίες από το πρώτο ερώτημά του. Αντιθέτως, δημιουργείται ένα καινούργιο ερώτημα.

Υπολογισμός Αναπαράστασης Αιτημάτων Χρήστη

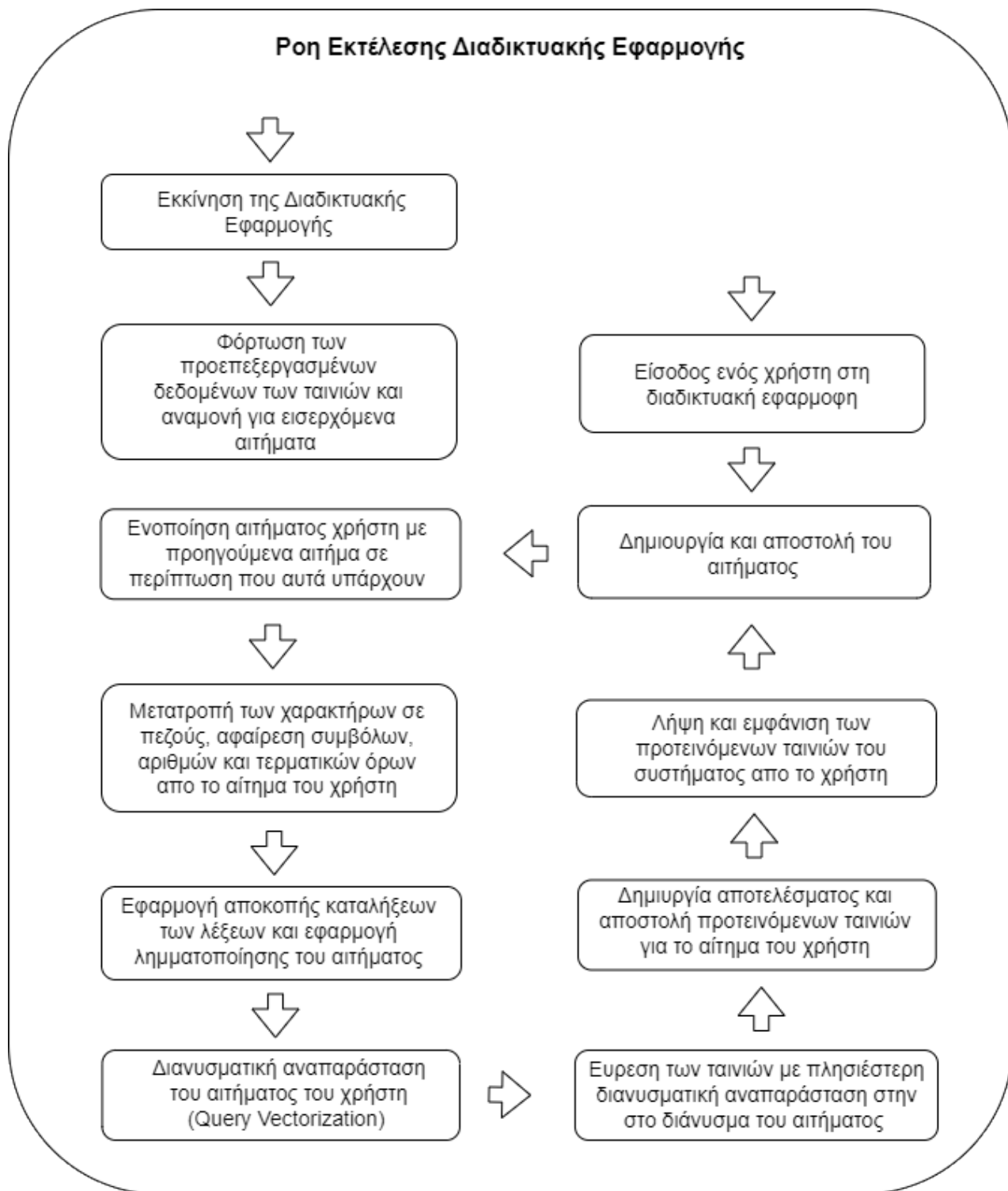


Σχ22. Στο παραπάνω σχήμα παρουσιάζεται ενδεικτικά η διανυσματική αναπαράσταση μερικών κειμένων από κινηματογραφικές ταινίες όπως επίσης και η διανυσματική αναπαράσταση δύο αιτημάτων του χρήστη. Στο συγκεκριμένο σχήμα παρατηρείται ο συνδυασμός των δύο επιμέρους ερωτημάτων και η δημιουργία ενός νέου εμπλουτισμένου ερωτήματος σε διανυσματική μορφή, που περιέχει πληροφορίες από τα δύο διανύσματα άλλα που αποτελούν τις συνιστώσες του. Διαπιστώνεται ότι το νέο διάνυσμα έχει αρκετά διαφορετική μορφή από τα άλλα δύο καθώς και ένα συνδυασμό πληροφοριών από τα άλλα δύο.

Υπολογισμός Αναπαράστασης Αιτημάτων Χρήστη



Σχ23. Στο σχήμα παρατηρείται ότι τα δύο διαφορετικά διανύσματα που αποτελούν τα αιτήματα του χρήστη, είναι πλησιέστερα σε διανυσματικές αναπαραστάσεις διαφορετικών κινηματογραφικών ταινιών, ενώ ο συνδυασμός τους είναι πλησιέστερος σε άλλη διανυσματική αναπαράσταση κινηματογραφικής ταινίας. Παρατηρείται ότι το νέο διάνυσμα που προκύπτει από τον συνδυασμό των άλλων δύο αποτελεί ένα επαυξημένο διάνυσμα αναπαράστασης. Το συγκεκριμένο διάνυσμα περιλαμβάνει πληροφορία από τα άλλα δύο διανύσματα που αποτελούν και τις συνιστώσες του.



Σχ24. Εδώ παρουσιάζεται σχηματικά τη δομή και τα βασικά επιμέρους τμήματα της διαδικτυακής εφαρμογής του συστήματος πρότασης κινηματογραφικών ταινιών όπως επίσης και η ροή εκτέλεσης από την έναρξή της.

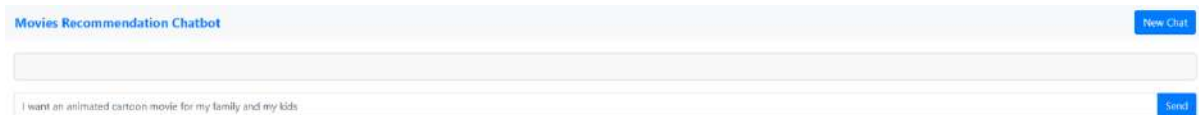
Παρακάτω παρουσιάζεται ένα δείγμα της εφαρμογής αμέσως μετά την έναρξή της:



The screenshot shows the 'Movies Recommendation Chatbot' interface. At the top, there is a header bar with the title 'Movies Recommendation Chatbot' on the left and a 'New Chat' button on the right. Below the header is a large, empty text input field. At the bottom of the input field, there is a placeholder text 'Type your message...'. To the right of the input field is a 'Send' button.

Σχ25. Στην παραπάνω εικόνα παρουσιάζεται ένα στιγμιότυπο της εφαρμογής κατά την έναρξή της, όπου αναμένεται από το χρήστη να υποβάλει το αίτημά του.

Παρακάτω παρουσιάζονται μερικά δείγματα της εφαρμογής αμέσως μετά την έναρξή της όπου ο χρήστης δίνει ως είσοδο την προτίμηση του στην εφαρμογή:



The screenshot shows the 'Movies Recommendation Chatbot' interface. At the top, there is a header bar with the title 'Movies Recommendation Chatbot' on the left and a 'New Chat' button on the right. Below the header is a large, empty text input field. At the bottom of the input field, there is a placeholder text 'Type your message...'. To the right of the input field is a 'Send' button.

Σχ26. Στο παραπάνω στιγμιότυπο ο χρήστης έχει γράψει το αίτημά του και αναμένεται να το υποβάλει στην εφαρμογή για να λάβει προτάσεις κινηματογραφικών ταινιών.

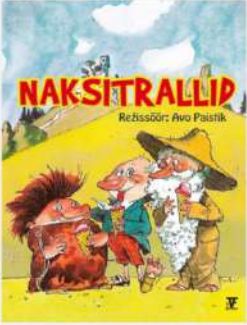
Παρακάτω παρουσιάζονται μερικά δείγματα της εφαρμογής αμέσως μετά την έναρξή της όπου ο χρήστης έχει δώσει ως είσοδο την προτίμηση του στην εφαρμογή και έχει λάβει απάντηση με τις κινηματογραφικές ταινίες που προτείνει το σύστημα:

Movies Recommendation Chatbot

New Chat

You I want an animated cartoon movie for my family and my kids

Recommended movies for you



Naksitrallid
Režissöör: Avo Paistik

Description: The animated cartoon tells about adventures three naksitrall look like gnomes.

Stars: Aarne Oksula, Jüri Kijukov, Maria Kleriskaja, Ita Ever

Director: Avo Paistik

Votes: 59.0

Genre: Animation

Ratings: 7.6

Year: 1990.0



JG and the BC Kids

Description: JG and the BC Kids is an animated feature film created by Janet Hubert bringing street smart kids and book smart kids together which brings life smart kids based on math, science, 2D animation.

Stars: Cheryl Alexander, Blaze Berdehl, Raven Goodwin, Jasmine Hanes

Director: Indra Narayan Diatta, Antony Leo, Ron Myrick

Votes: nan

Genre: Animation

Ratings: nan

Year: 2020.0



Kniamon Bedtime Tales
By Sue Ellen Randolph

Description: Based on the children's book "Kniamon Bedtime Tales" This film tells the kids 8 bedtime stories. It also teaches the kids about the phonics, the planets, constellations, what animal comes ... See full summary »

Stars: Sue Ellen Randolph

Directors: Sue Ellen Randolph

Votes: nan

Genre: Animation

Ratings: nan

Year: 2019.0



Mars

Description: An upcoming animated feature film, written by and starring The Whitest Kids U' Know.

Stars: Zach Cregger, Trevor Moore, Timmy Williams, Sam Brown

Director: Sevan Najarlan

Votes: nan

Genre: Animation, Comedy, Fantasy

Ratings: nan

Year: nan

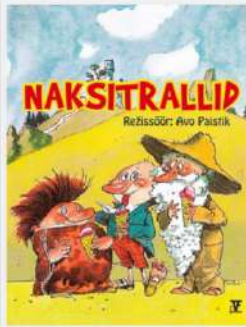
Type your message...

Send

Σχ27. Στο στιγμιότυπο φαίνεται ότι ο χρήστης έχει υποβάλλει το αίτημά του και έχει λάβει ως απάντηση τα αποτελέσματα των κινηματογραφικών ταινιών που του έχει προτείνει το σύστημα πρότασης ταινιών.

You: You want an animated cartoon movie for my family and my kids

Recommended movies for you



Naksitrallid

Description: The animated cartoon tells about adventures three naksitrall look like gnomes.

Stars: Aarne Ölskõla, Jõni Kijaku, Maria Klenskaja, Ita Ever

Director: Avo Paistik

Votes: 59.0

Genre: Animation

Ratings: 7.6

Year: 1990.0



JG and the BC Kids

Description: JG and the BC Kids is an animated feature film created by Janet Hubert bringing street smart kids and book smart kids together which brings life smart kids based on math, science, 2D animation.

Stars: Cheryl Alexander, Blaze Berdahl, Raven Goodwin, Jasmine Hanes

Director: Indra Narayan Datta, Antony Leo, Ron Myrick

Votes: nan

Genre: Animation

Ratings: nan

Year: 2020.0



Kniamon Bedtime Tales

Description: Based on the children's book "Kniamon Bedtime Tales" This film tells the kids 8 bedtime stories. It also teaches the kids about the phonics, the planets, constellations, what animal comes ... See full summary »

Stars: Sue Ellen Randolph

Director: Sue Ellen Randolph

Votes: nan

Genre: Animation

Ratings: nan

Year: 2019.0



Mars

Description: An upcoming animated feature film, written by and starring The Whitest Kids U' Know.

Stars: Zach Cregger, Trevor Moore, Timmy Williams, Sam Brown

Director: Sevan Najarian

Votes: nan

Genre: Animation, Comedy, Fantasy

Ratings: nan

Year: nan

It's christmas so I want a Christmas movie with with santa

Send

Σχ28. Στο στιγμιότυπο παρουσιάζεται ότι ο χρήστης έχει λάβει τα αποτελέσματα του συστήματος και μπορεί να συντάξει ένα νέο ερώτημα ή αίτημα για να επαναπροσδιορίσει το αρχικό αίτημα, επιδιώκοντας μια πιο ακριβή απάντηση. Με αυτή τη δυνατότητα, ο χρήστης μπορεί να καθοδηγήσει καλύτερα την εφαρμογή για να λάβει πιο στοχευμένες και εξατομικευμένες προτάσεις ταινιών.

Recommended movies for you



You: It's Christmas so I want a Christmas movie with with Santa



Send

Σχ29. Στο παραπάνω στιγμιότυπο παρουσιάζεται ότι ο χρήστης έχει λάβει τα αποτελέσματα του νέου ερωτήματος και έχει λάβει μια βελτιωμένη απάντηση με πιο ακριβείς προτάσεις ταινιών, σε σχέση με το αρχικό του ερώτημα.

Κεφάλαιο 6 - Συμπεράσματα

Συνοψίζοντας, η παρούσα διπλωματική εργασία ανέδειξε τη δυνατότητα χρήσης μεθόδων επεξεργασίας φυσικής γλώσσας (Natural Language Processing ή αλλιώς NLP) για τη δημιουργία μιας διαδραστικής πλατφόρμας συνομιλίας που προτείνει κινηματογραφικές ταινίες στους χρήστες της. Μέσω της χρήσης διανυσματικών αναπαραστάσεων λέξεων με το μοντέλο GloVe, καθώς και της διανυσματικής αναπαράστασης των κειμένων και των αιτημάτων των χρηστών, σε συνδυασμό με μεθόδους όπως η αφαίρεση τερματικών όρων, η λημματοποίηση, και η αποκοπή καταλήξεων, επιτεύχθηκε η αποτελεσματική επεξεργασία κειμένων φυσικής γλώσσας. Οι περιγραφές των κινηματογραφικών ταινιών, όπως επίσης και τα αιτήματα των χρηστών, μετατράπηκαν σε μία αντίστοιχη διανυσματική μορφή, επιτρέποντας έτσι την εφαρμογή μεθόδων και τεχνικών αναζήτησης για την εύρεση παραπλήσιων αποτελεσμάτων σύμφωνα με τα ερωτήματα των χρηστών.

Η δυνατότητα επέκτασης και συνένωσης των διαδοχικών αιτημάτων του χρήστη αποδείχθηκε ένα χρήσιμο εργαλείο για την παροχή πιο ακριβών και εξατομικευμένων αιτημάτων ή αλλιώς ερωτήσεων για τη βελτίωση των αποτελεσμάτων. Η διαρκής ανατροφοδότηση και προσαρμογή του διανύσματος αναπαράστασης βάσει των νέων αιτημάτων του χρήστη βελτίωσε σημαντικά την ακρίβεια και τη σχετικότητα των αποτελεσμάτων των προτεινόμενων κινηματογραφικών ταινιών. Αυτή η προσέγγιση επέτρεψε τη διατήρηση της σημασιολογικής πληροφορίας σε διαδοχικά αιτήματα του χρήστη, προσφέροντας μια πιο συγκεντρωτική και ολοκληρωμένη εικόνα των προτιμήσεων του.

Τέλος, η εφαρμογή αυτής της πλατφόρμας συνομιλίας απέδειξε ότι οι τεχνικές επεξεργασίας φυσικής γλώσσας και μηχανικής μάθησης μπορούν να εφαρμοστούν αποτελεσματικά και σε πραγματικά σενάρια για την παροχή εξατομικευμένων αποτελεσμάτων και προτάσεων. Οι χρήστες έχουν τη δυνατότητα να υποβάλλουν επαυξημένα ερωτήματα μέσω μιας συνεχούς και διαδραστικής διαδικασίας αναζήτησης, όπου τα αιτήματά τους εμπλουτίζονται διαδοχικά, προσφέροντας έτσι μια πιο συντομευμένη και προσαρμοσμένη συλλογή αποτελεσμάτων. Αυτή η διαδικασία ενισχύει την ακρίβεια και τη σχετικότητα των προτάσεων κινηματογραφικών ταινιών, επιδεικνύοντας τις δυνατότητες των σύγχρονων τεχνολογιών επεξεργασίας φυσικής γλώσσας (Natural Language Processing ή NLP) και μηχανικής εκμάθησης (machine Learning ή ML).

Κεφάλαιο 7 - Βελτιώσεις και μελλοντικές προσθήκες

Για περαιτέρω βελτιώσεις και επέκταση της πλατφόρμας πρότασης ταινιών της συγκεκριμένης διπλωματικής εργασίας, υπάρχει η δυνατότητα να ενσωματωθούν μέθοδοι και τεχνικές όπως η Ονομαστική Αναγνώριση Οντοτήτων (Named Entity Recognition ή αλλιώς NER). Η χρήση της ονομαστικής αναγνώρισης οντοτήτων μπορεί να ενισχύσει σημαντικά την ακρίβεια των αναζητήσεων και των προτάσεων ταινιών, καθώς επιτρέπει την αναγνώριση και εξαγωγή συγκεκριμένων ονομαστικών πληροφοριών από τα αιτήματα των χρηστών, όπως για παράδειγμα τα ονόματα των ηθοποιών, των σκηνοθετών ή τους τίτλους των κινηματογραφικών ταινιών. Με αυτό τον τρόπο, η πλατφόρμα της εφαρμογής έχει τη δυνατότητα να ανταποκρίνεται με μεγαλύτερη ακρίβεια στα αιτήματα ή αλλιώς τα ερωτήματα των χρηστών και να προσφέρει πιο στοχευμένες ή προσωποποιημένες προτάσεις και αποτελέσματα στις προτεινόμενες ταινίες που παρέχει.

Επιπλέον, η επέκταση της εφαρμογής με την ενσωμάτωση ενός συστήματος προτάσεων βάσει των προτιμήσεων άλλων χρηστών με παρόμοια ή παραπλήσια ενδιαφέροντα μπορεί να βελτιώσει αισθητά την ακρίβεια και τη σχετικότητα των αποτελεσμάτων που παρέχονται. Αυτό μπορεί να επιτευχθεί με τη χρήση αλγορίθμων μηχανικής μάθησης που αναλύουν τα πρότυπα ή αλλιώς μοτίβα των αιτημάτων και τις τάσεις των χρηστών, προσαρμόζοντας τις προτάσεις των κινηματογραφικών ταινιών ανάλογα με τις προτιμήσεις άλλων χρηστών με παρόμοια ενδιαφέροντα. Ένα σύστημα προτάσεων που λαμβάνει υπόψη τις προτιμήσεις χρηστών με παρόμοια ενδιαφέροντα είναι δυνατό να βελτιώσει τα αποτελέσματα των προτάσεων, παρέχοντας πιο στοχευμένα και εξατομικευμένα αποτελέσματα στα ενδιαφέροντα του χρήστη.

Τέλος, υπάρχει η δυνατότητα προσθήκης περαιτέρω λειτουργιών, όπως η δυνατότητα για τους χρήστες να αξιολογούν τα αποτελέσματά που παρέχονται και να δίνουν ανατροφοδότηση στο σύστημα πρότασης ταινιών. Η ανατροφοδότηση αυτή μπορεί να χρησιμοποιηθεί για τη συνεχή παραμετροποίηση και τη βελτίωση των αποτελεσμάτων που παρέχει η εφαρμογή. Με αυτές τις προσθήκες, η εφαρμογή έχει τη δυνατότητα να παρέχει καλύτερα αποτελέσματα στους χρήστες της λαμβάνοντας υπόψιν ονομαστικές οντότητες καθώς επίσης και προτιμήσεις άλλων χρηστών με παραπλήσια συμπεριφορά ή ενδιαφέροντα.

Βιβλιογραφία

API. (2024). In *Wikipedia*.

<https://en.wikipedia.org/w/index.php?title=API&oldid=1226019698>

Chatbot history: Key bot characters through the ages. (n.d.). Retrieved June 3, 2024, from <https://www.whoson.com/chatbots-ai/chatbot-history-a-countdown-through-the-ages/>

ChatGPT. (2024). In *Wikipedia*.

<https://en.wikipedia.org/w/index.php?title=ChatGPT&oldid=1226877307>

Cosine similarity. (2024). In *Wikipedia*.

https://en.wikipedia.org/w/index.php?title=Cosine_similarity&oldid=1224774490

Euclidean distance. (2024). In *Wikipedia*.

https://en.wikipedia.org/w/index.php?title=Euclidean_distance&oldid=121427356

6

Gensim: Topic modelling for humans. (n.d.). Retrieved June 4, 2024, from

<https://radimrehurek.com/gensim/index.html?fbclid=IwAR1oCPwe7OIB6RjzJTWN CVCFNGihBMoP92sny2o5jiAgp4Duslbn53jgi3w>

GloVe: Global Vectors for Word Representation. (n.d.). Retrieved June 4, 2024, from

https://nlp.stanford.edu/projects/glove/?fbclid=IwAR1mptRMHHWvotJihQEZrlpng _18MrsWdJMQhVms00YHGLuUNmVCEO06ozA

Lemmatization. (2023). In *Wikipedia*.

<https://en.wikipedia.org/w/index.php?title=Lemmatization&oldid=1188111640>

pandas—Python Data Analysis Library. (n.d.). Retrieved June 4, 2024, from <https://pandas.pydata.org/>

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

<https://doi.org/10.3115/v1/D14-1162>

Pierce, D. (2024, February 28). *From Eliza to ChatGPT: Why people spent 60 years building chatbots*. The Verge.

<https://www.theverge.com/24054603/chatbot-chatgpt-eliza-history-ai-assistants-video>

Ramos, J. (n.d.). *Using TF-IDF to Determine Word Relevance in Document Queries*.

Scikit-learn: Machine learning in Python—Scikit-learn 1.5.0 documentation.

(n.d.). Retrieved June 4, 2024, from [https://scikit-](https://scikit-learn.org/stable/?fbclid=IwAR2tleFO-f8Fh_96S1WSLb0QNIJfi0xBv_Ls2mM5oq3F9iZuh9RMk7_6cRo)

[learn.org/stable/?fbclid=IwAR2tleFO-](https://scikit-learn.org/stable/?fbclid=IwAR2tleFO-f8Fh_96S1WSLb0QNIJfi0xBv_Ls2mM5oq3F9iZuh9RMk7_6cRo)

[f8Fh_96S1WSLb0QNIJfi0xBv_Ls2mM5oq3F9iZuh9RMk7_6cRo](https://scikit-learn.org/stable/?fbclid=IwAR2tleFO-f8Fh_96S1WSLb0QNIJfi0xBv_Ls2mM5oq3F9iZuh9RMk7_6cRo)

Stemming. (2024). In *Wikipedia*.

<https://en.wikipedia.org/w/index.php?title=Stemming&oldid=1196894403>

Stop word. (2024). In *Wikipedia*.

https://en.wikipedia.org/w/index.php?title=Stop_word&oldid=1216542266

Tf-idf. (2024). In *Wikipedia*.

<https://en.wikipedia.org/w/index.php?title=Tf%E2%80%93idf&oldid=1214201167>

Vector Operations. (n.d.). Retrieved June 4, 2024, from

[https://people.math.harvard.edu/~jjchen/math21a/handouts/vector-](https://people.math.harvard.edu/~jjchen/math21a/handouts/vector-ops.html?fbclid=IwAR1X-3sMhjFnGjOi9lQqrioRqsGNCWvUwcVFz85jKfXIN8OYYKIwRSxGaGs)

[ops.html?fbclid=IwAR1X-](https://people.math.harvard.edu/~jjchen/math21a/handouts/vector-ops.html?fbclid=IwAR1X-3sMhjFnGjOi9lQqrioRqsGNCWvUwcVFz85jKfXIN8OYYKIwRSxGaGs)

[3sMhjFnGjOi9lQqrioRqsGNCWvUwcVFz85jKfXIN8OYYKIwRSxGaGs](https://people.math.harvard.edu/~jjchen/math21a/handouts/vector-ops.html?fbclid=IwAR1X-3sMhjFnGjOi9lQqrioRqsGNCWvUwcVFz85jKfXIN8OYYKIwRSxGaGs)

Wang, C., Nulty, P., & Lillis, D. (2020). *A Comparative Study on Word Embeddings in Deep Learning for Text Classification*. 37–46.

<https://doi.org/10.1145/3443279.3443304>

What is a chatbot + how does it work? The ultimate guide. (2020, October 22).

Zendesk. <https://www.zendesk.com/blog/what-is-a-chatbot/>

Word embedding. (2024). In *Wikipedia*.

https://en.wikipedia.org/w/index.php?title=Word_embedding&oldid=1225649017