ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ & ΠΛΗΡΟΦΟΡΙΚΗΣ

Ανάκτηση Πληροφορίας

Χειμερινό εξάμηνο 2021-22

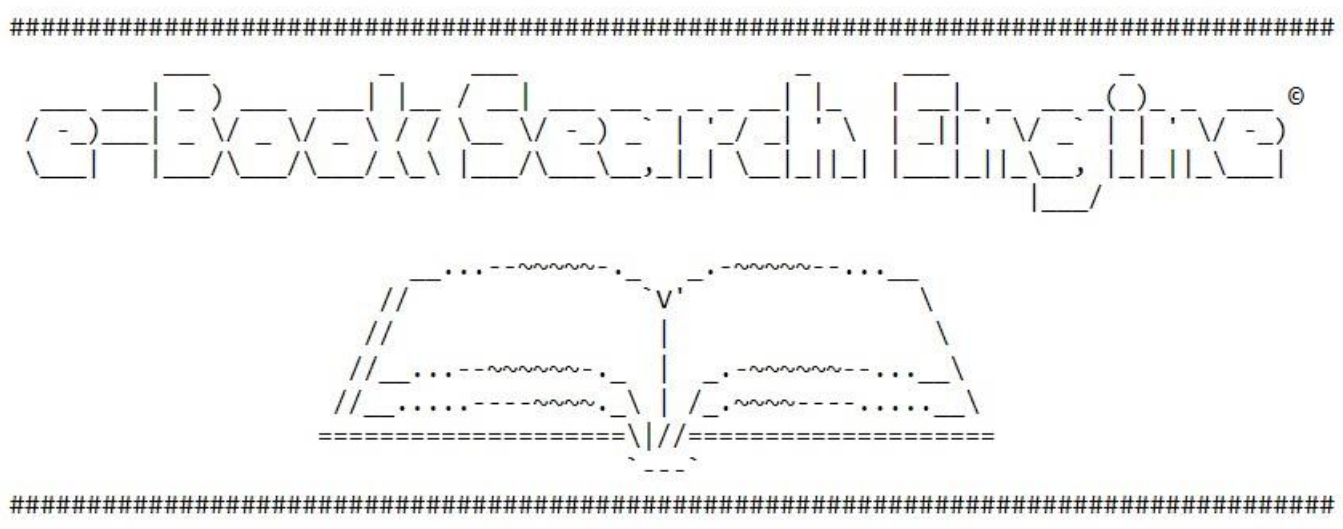
Αναφορά Υλοποιητικού Project

Στοιχεία ομάδας:

Παπαγεωργίου Νικηφόρος – Γεώργιος	1059633	st1059633@ceid.upatras.gr
Σταμόπουλος Νικόλαος	1057764	st1057764@ceid.upatras.gr

ΙΑΝΟΥΑΡΙΟΣ 2022 | ΠΑΤΡΑ

Εισαγωγή



Στα πλαίσια της παρούσας εργασίας, κληθήκαμε να υλοποιήσουμε μια μηχανή αναζήτησης βιβλίων, η οποία θα βασίζεται στην Elasticsearch και θα αποφασίζει τη σειρά παρουσίασης των αποτελεσμάτων χρησιμοποιώντας τεχνικές μηχανικής μάθησης. Σαν γλώσσα προγραμματισμού επιλέξαμε την **Python** και την έκδοση **3.6.8** αυτής. Χαρακτηριστικές βιβλιοθήκες που χρησιμοποιήσαμε είναι οι elasticsearch, pandas, numpy, tensorflow, keras, sklearn, nltk κ.α. Το σύνολο των ερωτημάτων της εργασίας υλοποιήθηκαν στο περιβάλλον του **Jupyter Notebook** και τα αρχεία κώδικα που παραδίδουμε είναι της μορφής ".ipynb". Ακολουθεί μία σύντομη περιγραφή της υλοποίησης κάθε ερωτήματος. Αναλυτικότερη περιγραφή παρέχεται μέσω σχολίων στα αντίστοιχα αρχεία κώδικα.

Ερώτημα 1

Αρχικά, κατεβάσαμε την Elasticsearch από την επίσημη σελίδα της Elastic, και αφού κάναμε unzip τον φάκελο, εκκινούμε την Elasticsearch στο σύστημά μας μέσω της εντολής "bin\elasticsearch" στη γραμμή εντολών (command prompt).

Έχοντας πραγματοποιήσει επιτυχή σύνδεση με την Elasticsearch, δημιουργούμε ένα νέο index ονόματι "books", στο οποίο ανεβάζουμε τις εγγραφές του αρχείου "BX-Books.csv". Στη συνέχεια, ζητείται από τον χρήστη να εισάγει ένα λήμμα αναζήτησης, το οποίο τρέχουμε σαν query στην Elasticsearch για να πάρουμε όλα τα σχετικά με αυτό αποτελέσματα. Ορίσαμε το πλήθος των αποτελεσμάτων που θα επιστρέψει η Elasticsearch να είναι 10.000, όσο είναι δηλαδή το max_result_window. Για την καλύτερη απεικόνιση των αποτελεσμάτων, χρησιμοποιούμε έναν πίνακα BeautifulTable στον οποίο αποθηκεύουμε τα ονόματα των βιβλίων και τα αντίστοιχα scores τους βάσει της προκαθορισμένης μετρικής ομοιότητας της Elasticsearch, BM25. Παρακάτω, παραθέτουμε ένα στιγμιότυπο εκτέλεσης για το λήμμα αναζήτησης "clara".

BOOK RESULTS	BM25 SCORE
Clara Callan	12.645
Henry and Clara	11.622
Clara Mondschein's Melancholia	11.622
Clara : A Novel	11.622
Clara Bow: Runnin' Wild	10.752
Clara Joins the Circus	10.752
Clara Barton (Women of Achievement)	10.003
Sweet Clara and the Freedom Quilt	9.351
Dancing With Clara (Signet Regency Romance)	9.351
Driven to Kill: The Clara Harris Story	8.779
Clara and the Bookwagon (I Can Read Book 3)	7.823
This Old House: The Story of Clara Rust Alaska Pioneer	7.419
Clara Barton : Founder Of The American Red Cross (Childhood Of Famous Americans)	6.724
The Glory Cloak : A Novel of Louisa May Alcott and Clara Barton	6.724
South Bay Trails: Outdoor Adventures in & Around Santa Clara Valley : From the Diablo Range to the Pacific Ocean	5.249

Ερώτημα 2

Μια βασική παραδοχή που κάνουμε γι' αυτό το ερώτημα, είναι ότι θεωρούμε ως έγκυρη βαθμολογία στο αρχείο "BX-Book-Ratings.csv" όποια βαθμολογία ανήκει στο διάστημα $[1, 10]$, δηλαδή ότι η χαμηλότερη βαθμολογία που μπορεί να δώσει ένας χρήστης είναι 1, ενώ η υψηλότερη 10. Επομένως, όπου υπάρχουν μηδενικά στο csv αρχείο, θεωρούμε πως δεν υπάρχει βαθμολογία.

Όπως και στο προηγούμενο ερώτημα, συνδεόμαστε με την Elasticsearch, καθώς από εκεί θα αντλήσουμε τα αποτελέσματα της αναζήτησης όπως και τη μετρική BM25 που είναι απαραίτητη για τον υπολογισμό της δικής μας personalized μετρικής. Ακολούθως, από τον χρήστη ζητείται να δώσει το ID του και, όπως και προηγουμένως, ένα λήμμα αναζήτησης. Παράλληλα, έχοντας φορτώσει σε dataframe το csv αρχείο με τις βαθμολογίες, καθαρίζουμε τις μηδενικές βαθμολογίες βάσει της παραδοχής και δημιουργούμε δύο νέα dataframes από αυτό: ένα με τη μέση βαθμολογία κάθε βιβλίου και ένα με τις προσωπικές βαθμολογίες του χρήστη. Αφού επιστραφούν τα αποτελέσματα της αναζήτησης από την Elasticsearch, για κάθε ένα από τα αποτελέσματα υπολογίζουμε την personalized μετρική λαμβάνοντας υπόψη την BM25 μετρική, τη μέση και την προσωπική βαθμολογία του χρήστη για το εκάστοτε βιβλίο (αν υπάρχουν). Αναλόγως με τον αν υπάρχουν ή δεν υπάρχουν όλα ή κάποιο από αυτά τα δεδομένα, ή ακόμα και αν η βαθμολογία είναι αρνητική (≤ 5) ή θετική (> 5), λαμβάνουμε ειδικές υποπεριπτώσεις για τον υπολογισμό της τελικής τιμής της μετρικής. Τέλος, ανάμεσα σε μέση και προσωπική βαθμολογία δίνουμε ένα μεγαλύτερο ποσοστό στην προσωπική (περίπου 60-40), καθώς θεωρούμε πως κάτι που άρεσε σε κάποιο χρήστη πρέπει να εμφανίζεται πιο ψηλά στα αποτελέσματα. Παρακάτω, παραθέτουμε ένα στιγμιότυπο εκτέλεσης για τον χρήστη με ID "11676" και λήμμα αναζήτησης "clara".

BOOK RESULTS	PERSONALIZED SCORE	BM25 SCORE	AVERAGE RATING	PERSONAL RATING
Clara Callan	20.935	12.645	7.667	8.0
Henry and Clara	18.014	11.622	7.5	6.0
Clara Bow: Runnin' Wild	13.619	10.752	8.0	0
Sweet Clara and the Freedom Quilt	11.845	9.351	8.0	0
This Old House: The Story of Clara Rust Alaska Pioneer	9.397	7.419	8.0	0
Clara Mondschein's Melancholia	9.169	11.622	0	0
Clara : A Novel	9.169	11.622	0	0
Clara Barton : Founder Of The American Red Cross (Childhood Of Famous Americans)	8.741	6.724	9.0	0
Clara Joins the Circus	8.377	10.752	0	0
Clara Barton (Women of Achievement)	7.7	10.003	0	0
Dancing With Clara (Signet Regency Romance)	7.013	9.351	4.0	0
South Bay Trails: Outdoor Adventures in & Around Santa Clara Valley : From the Diablo Range to the Pacific Ocean	6.824	5.249	9.0	0
Driven to Kill: The Clara Harris Story	6.607	8.779	0	0
Clara and the Bookwagon (I Can Read Book 3)	6.258	7.823	5.0	0
The Glory Cloak : A Novel of Louisa May Alcott and Clara Barton	4.818	6.724	0	0

Παρατηρούμε πως, σε σύγκριση με το προηγούμενο ερώτημα, η ταξινόμηση έχει βελτιωθεί, καθώς πιο ψηλά στα αποτελέσματα πλέον εμφανίζονται βιβλία με καλή μέση ή/και προσωπική βαθμολογία. Αντιθέτως, βιβλία που ήταν προηγουμένως ψηλότερα λόγω της τιμής της μετρικής BM25, αλλά δεν έχουν ούτε μέση ούτε προσωπική βαθμολογία, «τιμωρούνται» στη τρέχουσα κατάταξη χάνοντας από βιβλία που προηγουμένως ήταν χαμηλότερα. Όπου εμφανίζεται 0 σε βαθμολογία, σημαίνει ότι δεν υπάρχει.

Ερώτημα 3

Παρομοίως, και σ' αυτό το ερώτημα, πραγματοποιείται, αρχικά, σύνδεση με την Elasticsearch και ζητείται από τον χρήστη να εισάγει το ID του και ένα λήμμα αναζήτησης. Για να εκπαιδεύσουμε το νευρωνικό μας δίκτυο, χρησιμοποιούμε τα summaries των βιβλίων, τα οποία έχει βαθμολογήσει ο χρήστης, και τις βαθμολογίες αυτών ως τα αντίστοιχα labels. Προτού, όμως, δοθούν τα summaries προς εκπαίδευση του νευρωνικού, υπόκεινται σε προεπεξεργασία, κατά την οποία αφαιρούνται ειδικοί χαρακτήρες, stop-words και γίνεται στελεχοποίηση (stemming) των λέξεων που περιέχονται στα summaries.

Στη συνέχεια, μέσω της Tokenizer κλάσης που παρέχεται από το keras, μετατρέπουμε κάθε summary σε μια ακολουθία από ακέραιους αριθμούς που αντιστοιχούν στις αντίστοιχες λέξεις του summary. Επειδή θέλουμε όλα τα διανύσματα να έχουν το ίδιο μήκος, εφαρμόζουμε zero padding. Χρησιμοποιώντας το GloVe, και πιο συγκεκριμένα το txt αρχείο "glove.6B.100d", το οποίο περιέχει διανύσματα 100 διαστάσεων, δημιουργούμε τα word embeddings για όλες τις μοναδικές λέξεις των summaries. Ακολούθως, χρησιμοποιούμε το embedding_matrix που παράγουμε από το GloVe για να το φορτώσουμε στο Embedding layer του νευρωνικού μας.

Αφού ολοκληρωθεί η εκπαίδευση του νευρωνικού πάνω στις υπάρχουσες βαθμολογίες του χρήστη, προχωράμε επιστρέφοντας από την Elasticsearch τα αποτελέσματα της αναζήτησης του χρήστη. Τα summaries των αποτελεσμάτων αφού υπόκεινται και αυτά σε προεπεξεργασία, δίνονται σαν είσοδος στο νευρωνικό για να προσπαθήσει να προβλέψει τη βαθμολογία του χρήστη. Τέλος, υπολογίζεται πάλι η personalized μετρική από το 2ο ερώτημα, μόνο που αυτή τη φορά θα υπάρχει πάντα προσωπική βαθμολογία του χρήστη για κάθε βιβλίο, είτε αυτή έχει ήδη δοθεί από εκείνον, είτε έχει προβλεφθεί από το νευρωνικό. Παρακάτω, παραθέτουμε το ίδιο στιγμιότυπο εκτέλεσης για τον χρήστη με ID "11676" και λήμμα αναζήτησης "clara".

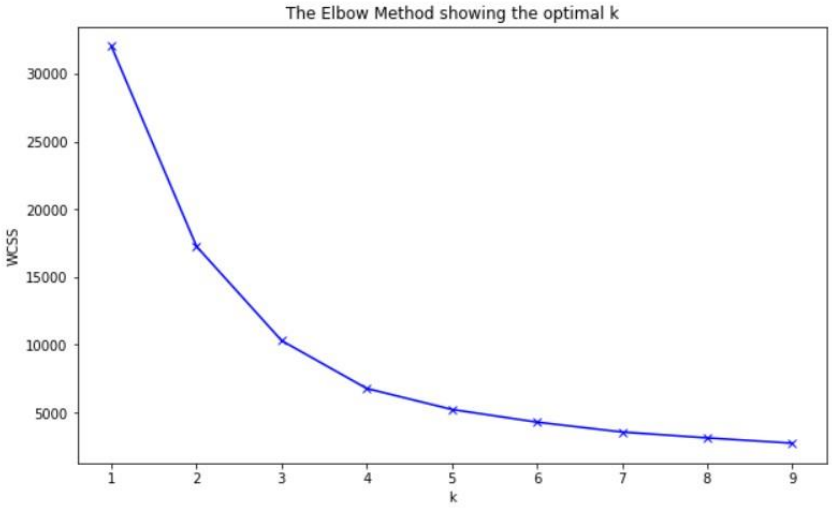
BOOK RESULTS	PERSONALIZED SCORE	BM25 SCORE	AVERAGE RATING	PERSONAL RATING
Clara Callan	20.935	12.645	7.667	8.0
Henry and Clara	18.014	11.622	7.5	6.0
Clara Bow: Runnin' Wild	17.382	10.752	8.0	7.0
Sweet Clara and the Freedom Quilt	16.52	9.351	8.0	10.0
Clara : A Novel	15.69	11.622	0	7.0
Clara Mondschein's Melancholia	15.109	11.622	0	6.0
Clara Joins the Circus	14.515	10.752	0	7.0
Clara Barton (Women of Achievement)	14.504	10.003	0	9.0
This Old House: The Story of Clara Rust Alaska Pioneer	12.364	7.419	8.0	8.0
Driven to Kill: The Clara Harris Story	12.291	8.779	0	8.0
Clara Barton : Founder Of The American Red Cross (Childhood Of Famous Americans)	11.094	6.724	9.0	7.0
Dancing With Clara (Signet Regency Romance)	9.819	9.351	4.0	6.0
Clara and the Bookwagon (I Can Read Book 3)	9.778	7.823	5.0	9.0
The Glory Cloak : A Novel of Louisa May Alcott and Clara Barton	8.741	6.724	0	6.0
South Bay Trails: Outdoor Adventures in & Around Santa Clara Valley : From the Diablo Range to the Pacific Ocean	8.661	5.249	9.0	7.0

Παρατηρούμε πως, σε σύγκριση με το προηγούμενο ερώτημα, η ταξινόμηση έχει βελτιωθεί κι άλλο, καθώς βιβλία για τα οποία προηγουμένως δεν διαθέταμε καμία βαθμολογία, πλέον μέσω πρόβλεψης τους δίνεται μια βαθμολογία βάσει παρόμοιων χαρακτηριστικών με βιβλία που έχουν βαθμολογηθεί ήδη από τον χρήστη. Οι βαθμολογίες αυτές ποικίλουν ανάλογα με τα δεδομένα που δίνονται κάθε φορά για εκπαίδευση του νευρωνικού δικτύου.

Ερώτημα 4

Αρχικά, φορτώνουμε όλο το αρχείο "BX-Books.csv" σε dataframe και προεπεξεργαζόμαστε τη στήλη με τα summaries, αφαιρώντας ειδικούς χαρακτήρες, σημεία στίξης και stop-words. Έπειτα, «σπάμε» κάθε summary σε μια λίστα από λέξεις, την οποία δίνουμε σαν είσοδο στο Word2Vec μοντέλο που δημιουργούμε για να υπολογίσουμε το διάνυσμα κάθε summary.

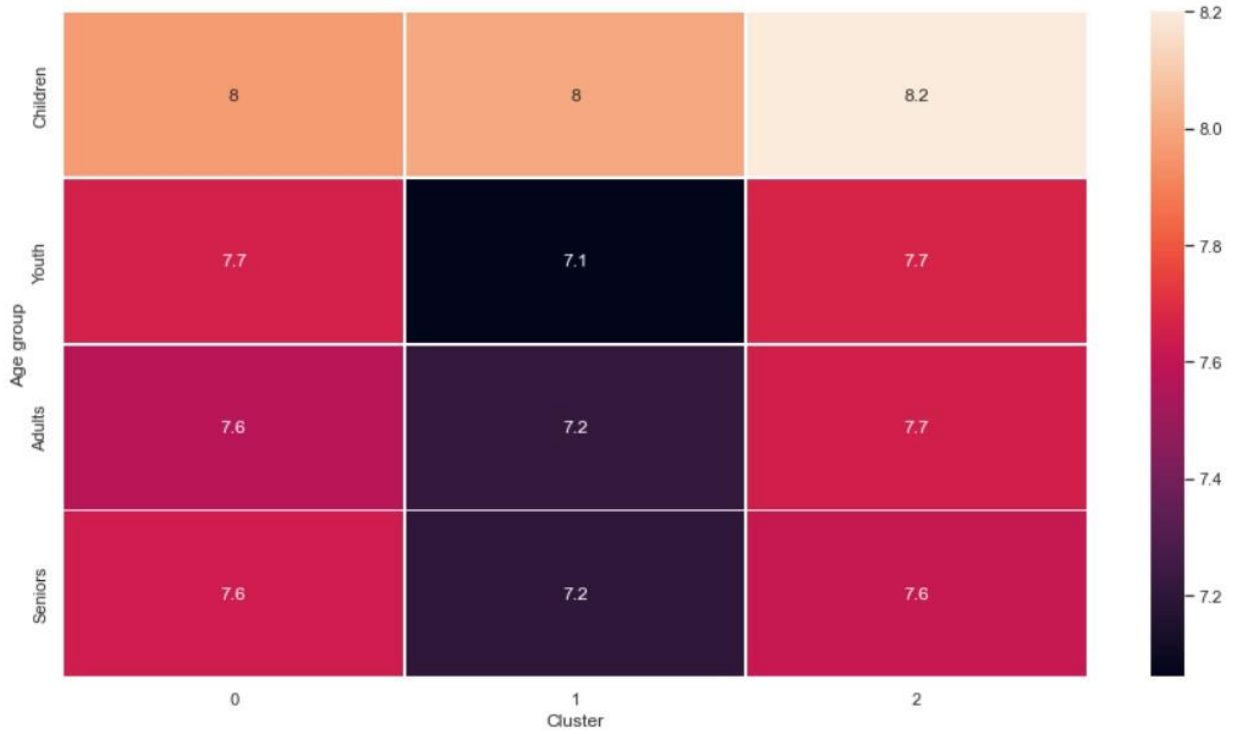
Για τον υπολογισμό των συστάδων μέσω του k-means με ομοιότητα συνημιτόνου, πραγματοποιούμε L2 κανονικοποίηση στα διανύσματα των summaries και στη συνέχεια καλούμε τον k-means, ο οποίος χρησιμοποιεί Ευκλείδεια απόσταση. Αυτή η διαδικασία, δίνει τα ίδια αποτελέσματα με τη μετρική συνημιτόνου (*πηγή:* Wikipedia). Επίσης, πριν καλέσουμε τον k-means, χρησιμοποιούμε τη μέθοδο PCA, για να μειώσουμε τη διαστατικότητα των διανυσμάτων σε 2D. Για να βρούμε το κατάλληλο k για τον k-means, κάνουμε χρήση του Elbow Method, τρέχοντας τον k-means για 1 έως και 10 συστάδες. Παρατηρούμε πως το κατάλληλο k είναι το 3.



Συνεπώς, έχοντας χωρίσει τα βιβλία σε τρεις συστάδες, φορτώνουμε σε dataframe το αρχείο "BX-Users.csv", όπου περιέχονται τα δημογραφικά στοιχεία των χρηστών, και κρατάμε μόνο αυτά για τα οποία υπάρχουν δοθείσες βαθμολογίες πέρα από 0, στο "BX-Book-Ratings.csv". Στη συνέχεια, αποφασίζουμε για κάθε συστάδα να παραθέσουμε μέσο όρο βαθμολογιών των χρηστών βάσει της ηλικίας τους και βάσει της τοποθεσίας τους και να συγκρίνουμε μεταξύ τους τις συστάδες:

I. Μέσος όρος βαθμολογιών βάσει ηλικιακής ομάδας

Θεωρούμε πως μπορούμε να κατανείμουμε τους χρήστες σε 4 ηλικιακές ομάδες, οι οποίες είναι τα παιδιά (Children), η νεολαία (Youth), οι ενήλικες (Adults) και οι ηλικιωμένοι (Seniors). Για κάθε μία από αυτές τις ομάδες υπολογίζουμε μέσο όρο βαθμολογίας σε κάθε συστάδα για να αποκτήσουμε μια εικόνα για τον τρόπο με τον οποίο βαθμολογούν τα βιβλία. Αναπαριστούμε τα τελικά δεδομένα και σε μορφή πίνακα, αλλά και σε heatmap.



Παρατηρούμε πως η πρώτη και η τρίτη συστάδα έχουν αρκετά παρόμοια αποτελέσματα, ενώ αντιθέτως η δεύτερη διαφοροποιείται από τις υπόλοιπες δύο, στις ηλικιακές ομάδες πέραν των παιδιών. Τέλος, παρατηρούμε πως τα παιδιά βαθμολογούν υψηλότερα σε σύγκριση με τις υπόλοιπες ηλικιακές ομάδες.

II. Μέσος όρος βαθμολογιών βάσει τοποθεσίας

Σ' αυτή την περίπτωση, θα θεωρήσουμε μια λίστα με τις 10 κορυφαίες χώρες σε πλήθος ratings, και όλες τις υπόλοιπες ως «Άλλες χώρες». Για κάθε μία από αυτές τις χώρες, υπολογίζουμε το μέσο όρο βαθμολογιών σε κάθε συστάδα και αναπαριστούμε πάλι τα τελικά δεδομένα σε μορφή πίνακα και heatmap.

Στο παραπάνω heatmap που παράγεται, μπορούμε να παρατηρήσουμε πώς διαφέρει ο μέσος όρος βαθμολογιών σε κάθε χώρα ανά συστάδα. Με "unknown" συμβολίζουμε τις χώρες για τις οποίες δεν είχαν δοθεί ακριβή στοιχεία από τους χρήστες.