Τμήμα Μηχανικών Η/Υ & Πληροφορικής
Πανεπιστήμιο Πατρών

Ανάκτηση Πληροφορίας

Εργαστηριακή Άσκηση
Χειμερινό Εξάμηνο 2021

Διδάσκων: Χ. Μακρής

Εκφώνηση

Στα πλαίσια της παρούσας εργαστηριακής άσκησης σας ζητείται να υλοποιήσετε μια μηχανή αναζήτησης βιβλίων η οποία θα βασίζεται στην Elasticsearch και θα αποφασίζει τη σειρά παρουσίασης των αποτελεσμάτων χρησιμοποιώντας τεχνικές μηχανικής μάθησης. Δεν ορίζεται γλώσσα υλοποίησης αλλά προτείνεται η χρήση της Python και των βιβλιοθηκών pandas, scikit-learn, tensorflow και keras.

[redacted] 1

Αρχικά, θα πρέπει να εγκαταστήσετε στο σύστημα σας την Elasticsearch και να γράψετε ένα μικρό πρόγραμμα το οποίο θα διαβάζει τις εγγραφές που περιέχονται στο αρχείο BX-Books.csv και θα τις εισάγει στην Elasticsearch. Στη συνέχεια, θα πρέπει να γράψετε ένα δεύτερο πρόγραμμα το οποίο θα δέχεται ως είσοδο (είτε ως όρισμα γραμμής εντολών είτε κατά τη διάρκεια της εκτέλεσής του) ένα αλφαριθμητικό και θα επιστρέφει την λίστα των ταινιών που ταιριάζουν με αυτό διατεταγμένη σε φθίνουσα σειρά σύμφωνα με την προκαθορισμένη μετρική ομοιότητας της Elasticsearch.

[redacted] 2

Σας ζητείται να τροποποιήσετε το δεύτερο πρόγραμμα του ερωτήματος 1 έτσι ώστε να δέχεται ως επιπρόσθετη είσοδο έναν ακέραιο αριθμό, το αναγνωριστικό του χρήστη. Ακόμα θα πρέπει να πρέπει να αλλάξετε τον τρόπο ταξινόμησης της λίστας των αποτελεσμάτων. Πλέον τα αποτελέσματα θα εμφανίζονται σύμφωνα με μια νέα μετρική την οποία θα δημιουργήσετε εσείς και η οποία θα συνυπολογίζει την μετρική ομοιότητας της Elasticsearch, τη βαθμολογία που έχει βάλει ο χρήστης στο βιβλίο (αν είναι διαθέσιμη) και το μέσο όρο όλων των βαθμολογιών που έχουν βάλει οι υπόλοιποι χρήστες. Τις βαθμολογίες των χρηστών για την κάθε ταινία θα τις βρείτε στο αρχείο BX-Book-Ratings.csv.

[redacted] [redacted] 3

Σε αυτό το ερώτημα θα επιχειρήσετε να βελτιώσετε την ποιότητα της ταξινόμησής σας συμπληρώνοντας τις βαθμολογίες που λείπουν. Για κάθε χρήστη, πάνω στα βιβλία για τα οποία υπάρχουν δεδομένα θα εκπαιδεύσετε έναν ένα νευρωνικό δίκτυο το οποίο θα χρησιμοποιήσετε για να μαντέψετε πώς ο συγκεκριμένος χρήστης θα βαθμολογούσε τα υπόλοιπα. Για να εκπαιδεύσετε το μοντέλο σας θα πρέπει να μετασχηματίσετε τα σύνολο δεδομένων που σας δόθηκε μετατρέποντας τις περιλήψεις των βιβλίων σε διανύσματα αξιοποιώντας την τεχνική των Word Embeddings. Προσπαθήσετε να συνδυάσετε τα αποτελέσματα των προηγούμενων ερωτημάτων για να πετύχετε την καλύτερη ταξινόμηση.

[redacted] [redacted] 4

Για το τελευταίο ερώτημα σας ζητείται να χωρίσετε τα βιβλία σε συστάδες με βάση την πλοκή τους. Για να το πετύχετε αυτό θα χρησιμοποιήσετε τα διανύσματα των περιλήψεων που δημιουργήσατε στο προηγούμενο ερώτημα τα οποία τώρα θα δώσετε ως είσοδο στον k-means. Ο αλγόριθμος θα πρέπει να τροποποιηθεί με τέτοιο τρόπο έτσι ώστε να χρησιμοποιεί ως μετρική απόστασης την ομοιότητα συνημίτονου (δύο στοιχεία με υψηλή ομοιότητα πρέπει να καταλήγουν στην ίδια συστάδα). Τέλος, προσπαθήστε να ανακαλύψετε συσχετίσεις μεταξύ των συστάδων που δημιουργήσατε και του τρόπου που βαθμολογούν οι χρήστες με βάση τα δημογραφικά τους στοιχεία που θα βρείτε στο αρχείο BX-Users.csv.

Παραδοτέα

1. Τα αρχεία κώδικα που υλοποιούν τα ζητούμενα της εκφώνησης.
2. Μια αναφορά σε μορφή pdf η οποία θα πρέπει να περιέχει τα ακόλουθα:
    - Τα στοιχεία (**ΑΜ, ονοματεπώνυμο και email**) του φοιτητή ή των φοιτητών που παραδίδουν την άσκηση.
    - Αναλυτική καταγραφή του περιβάλλοντος υλοποίησης (γλώσσα προγραμματισμού, βιβλιοθήκες λογισμικού κτλ.) καθώς και τα βήματα που απαιτούνται για την εγκατάστασή του.
    - Σύντομη περιγραφή της διαδικασία υλοποίησης.
    - Σχολιασμό των τελικών αποτελεσμάτων.

Διαδικαστικά

1. Η υλοποίηση της εργαστηριακής άσκησης απαλλάσσει τους φοιτητές από την υποχρέωση να παραδώσουν την θεωρητική εργασία.
2. Η άσκηση μπορεί να υλοποιηθεί είτε **ατομικά** είτε σε **ομάδες των δύο**.
3. Ως **ημερομηνία υποβολής** ορίζεται η **ημερομηνία τρεις ημέρες πριν την γραπτή εξέταση** του μαθήματος στις **23:59**.
4. Η άσκηση θα εξεταστεί προφορικά σε ημερομηνία που θα ανακοινωθεί μετά την ανακοίνωση του προγράμματος της εξεταστικής.

5. Η υποβολή της άσκησης πρέπει να γίνει μέσω του eclass του μαθήματος. Τα παραδοτέα της άσκησης θα πρέπει να περιέχονται σε ένα συνημμένο αρχείο με όνομα της μορφής **ir2021_AM1_AM2.zip**
6. Η άσκηση μπορεί να αποσταλεί πολλές φορές αλλά θα βαθμολογηθεί μόνο η τελευταία της υποβολή.
7. Τις σχετικές με την υλοποιητική εργασία απορίες σας μπορείτε να τις αποστέλλετε μέσω email στη διεύθυνση mpompotas@ceid.upatras.gr.