Τμήμα Μηχανικών Η/Υ & Πληροφορικής
Πανεπιστήμιο Πατρών

Ανάκτηση Πληροφορίας

Εργαστηριακή Άσκηση
Χειμερινό Εξάμηνο 2022

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.

Ερώτημα 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[1] και θα τις εισάγει στην Elasticsearch. Στη συνέχεια, θα πρέπει να γράψετε ένα πρόγραμμα το οποίο θα δέχεται ως είσοδο (είτε ως όρισμα γραμμής εντολών είτε κατά τη διάρκεια της εκτέλεσής του) ένα αλφαριθμητικό και έναν ακέραιο αριθμό, το αναγνωριστικό του χρήστη. Το πρόγραμμα αυτό θα επιστρέφει την λίστα των ταινιών που ταιριάζουν με το αλφαριθμητικό εισόδου διατεταγμένη σε φθίνουσα σειρά σύμφωνα με μια μετρική την οποία θα δημιουργήσετε εσείς και η οποία θα συνυπολογίζει την μετρική ομοιότητας της Elasticsearch και τη βαθμολογία που έχει βάλει ο χρήστης στο βιβλίο (αν είναι διαθέσιμη). Η επιστρεφόμενη λίστα θα πρέπει να εμφανίζει **μόνο το 10% των βιβλίων** με το καλύτερο ταίριασμα. Τις βαθμολογίες των χρηστών για το κάθε βιβλίο θα τις βρείτε στο αρχείο BX-Book-Ratings.csv.

Ερώτημα 2

Σε αυτό το ερώτημα σας ζητείται να ομαδοποιήσετε τους χρήστες σε συστάδες με χρήση του αλγόριθμου k-means βάσει της χώρας στην οποία κατοικούν και της ηλικίας τους. Στην

[1] https://www.kaggle.com/datasets/ruchi798/bookcrossing-dataset

συνέχεια για τον κάθε χρήστη θα συμπληρώσετε τις βαθμολογίες των βιβλίων που δεν έχει αξιολογήσει με το μέσο όρο των βαθμολογιών που έχουν τα συγκεκριμένα βιβλία στην συστάδα στην οποία ανήκει ο χρήστης. Τι επίδραση είχαν οι παραπάνω ενέργειες στην ταξινόμηση των αποτελεσμάτων με την χρήση της μετρικής του Ερωτήματος 2; Πληροφορίες για τους χρήστες θα βρείτε στο αρχείο BX-Users.csv.

Ερώτημα 3

Σε αυτό το ερώτημα θα επιχειρήσετε να βελτιώσετε την ποιότητα της ταξινόμησής σας συμπληρώνοντας τις βαθμολογίες που ακόμα λείπουν. **Για κάθε συστάδα χρηστών**, πάνω στα βιβλία για τα οποία υπάρχουν δεδομένα θα εκπαιδεύσετε έναν ένα νευρωνικό δίκτυο το οποίο θα χρησιμοποιήσετε για να μαντέψετε πώς η συγκεκριμένη ομάδα χρηστών 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.

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.
    - Περιγραφή της διαδικασία υλοποίησης.
    - Σχολιασμό των τελικών αποτελεσμάτων.

Διαδικαστικά

1. Επιλέγετε ή την υλοποιητική ή 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 **ir2022_ΑΜ1_ΑΜ2.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 (και κοινοποίηση σε makri@ceid.upatras.gr, kalogeropo@ceid.upatras.gr)