

Ανάκτηση Πληροφορίας

Εργαστηριακή Άσκηση Χειμερινό Εξάμηνο 2020

Διδάσκων: Χ. Μακρής

Επικουρικό: Α. Μπομπότας, Γ. Ρόμπολας

Εκφώνηση

Στα πλαίσια της παρούσας εργαστηριακής άσκησης σας ζητείται να υλοποιήσετε μια μηχανή αναζήτησης κινηματογραφικών ταινιών 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 movies.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 τη μετρική ομοιότητας της Elasticsearch (BM25).

Ερώτημα 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 στην ταινία (αν είναι διαθέσιμη) και το μέσο όρο όλων των βαθμολογιών της. Τις βαθμολογίες των χρηστών για την κάθε ταινία θα τις βρείτε στο αρχείο ratings.csv.

Ερώτημα 3

Ένα σημαντικό πρόβλημα που προκύπτει από την προσέγγιση του προηγούμενου ερωτήματος είναι πως οι χρήστες συνήθως βαθμολογούν μόνο ένα υποσύνολο των ταινιών. Σε αυτό το ερώτημα θα προσπαθήσετε να επιλύσετε αυτό το πρόβλημα χωρίζοντας τους χρήστες σε συστάδες (clusters) σύμφωνα με τον τρόπο που βαθμολογούν. Αρχικά, θα υπολογίσετε τον μέσο όρο των βαθμολογιών κάθε κατηγορίας ταινιών για κάθε χρήστη. Στη συνέχεια, θα συσταδοποιήσετε τους χρήστες εφαρμόζοντας τον αλγόριθμο k-means στο σύνολο δεδομένων που προέκυψε. Ακόλουθα, στο αρχικό σύνολο δεδομένων θα συμπληρώσετε για κάθε χρήστη τις βαθμολογίες που του λείπουν χρησιμοποιώντας τον μέσο όρο της ταινίας στην συστάδα στην οποία εκείνος ανήκει. Τέλος, θα πρέπει η μετρική που σχεδιάσατε πριν, να ενημερωθεί κατάλληλα ώστε να χρησιμοποιεί και τη νέα πληροφορία που παραγάγατε.

Ερώτημα 4

Σε αυτό το ερώτημα θα επιχειρήσετε να βελτιώσετε την ποιότητα της ταξινόμησής σας συμπληρώνοντας τις βαθμολογίες που λείπουν με έναν ακόμα τρόπο. Για κάθε χρήστη, πάνω στις ταινίες για τις οποίες υπάρχουν δεδομένα θα εκπαιδεύσετε έναν ένα νευρωνικό δίκτυο το οποίο θα χρησιμοποιήσετε για να μαντέψετε πώς ο συγκεκριμένος χρήστης θα βαθμολογούσε τις υπόλοιπες. Για να εκπαιδεύσετε το μοντέλο σας θα πρέπει να μετασχηματίσετε τα σύνολο δεδομένων που σας δόθηκε μετατρέποντας τους τίτλους των ταινιών σε διανύσματα αξιοποιώντας την τεχνική των Word Embeddings. Στη συνέχεια θα πρέπει να προσθέσετε σε αυτά τα τις κατηγορίες στις οποίες ανήκουν οι ταινίες με την τεχνική one hot encoding. Τα νέα διανύσματα που θα προκύψουν είναι οι εγγραφές του τελικού συνόλου δεδομένων με το οποίο θα δουλέψετε. Προσπαθήσετε να συνδυάσετε τα αποτελέσματα όλων των 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:
 - ο Τα στοιχεία (**AM, ονοματεπώνυμο και email**) του φοιτητή ή των φοιτητών που παραδίδουν την άσκηση.
 - ο Αναλυτική καταγραφή του περιβάλλοντος υλοποίησης (γλώσσα προγραμματισμού, βιβλιοθήκες λογισμικού κτλ.) καθώς και τα βήματα που απαιτούνται για την εγκατάστασή του.
 - ο Σύντομη περιγραφή της διαδικασίας υλοποίησης.
 - ο 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 **ir2020_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.