Τμήμα Μηχανικών Η/Υ & Πληροφορικής
Πανεπιστήμιο Πατρών

Τεχνικές Διαχείρισης και Εξόρυξης για Δεδομένα Μεγάλου Όγκου

Εργαστηριακή Άσκηση
Εαρινό Εξάμηνο 2022-23

Διδάσκοντες:
Αναπληρωτής Καθηγητής Χ. Μακρής

Περιβάλλον Υλοποίησης

Η άσκηση πρέπει να υλοποιηθεί χρησιμοποιώντας τη μηχανή επεξεργασίας δεδομένων μεγάλου όγκου Apache Spark και της βιβλιοθήκης Spark NLP. Είστε ελεύθεροι να χρησιμοποιήσετε όποια επιπλέον βιβλιοθήκη επιθυμείτε αρκεί να την συμπεριλάβετε στην αναφορά σας, ενώ για γλώσσα υλοποίησης μπορείτε να επιλέξετε μεταξύ της scala και της python. Η εγκατάσταση του Apache Spark για τους σκοπούς της άσκησης αρκεί να γίνει τοπικά και δεν χρειάζεται να δημιουργήσετε κάποιου είδους υπολογιστικού cluster.

Εκφώνηση

Σας δίνεται ένα σύνολο δεδομένων που αποτελείται από 144438 ιατρικές περιλήψεις και είναι διαθέσιμο στον σύνδεσμο https://github.com/sebischair/Medical-Abstracts-TC-Corpus. Σκοπός σας είναι να προσπαθήσετε να μαντέψετε την ασθένεια στην οποία αναφέρεται η κάθε περίληψη χρησιμοποιώντας τεχνικές μηχανικής μάθησης. Για να μετασχηματίσετε το σύνολο δεδομένων που σας δόθηκε έτσι ώστε να μπορέσετε να το εισάγετε στους κατηγοριοποιητές σας θα πρέπει να μετατρέψετε τα κείμενα σε διανύσματα χρησιμοποιώντας δύο διαφορετικές τεχνικές:

1. Θα αφαιρέσετε από το κείμενο σας εκείνες τις λέξεις που δεν προσφέρουν σημασιολογική πληροφορία (stopwords) και στο κείμενο που θα προκύψει θα εφαρμόσετε την τεχνική των Word Embeddings.

2. Θα χρησιμοποιήσετε τον προεκπαιδευμένο transformer BERT για να εξάγετε διανύσματα για την κάθε λέξη του κειμένου.

Στην συνέχεια, θα πρέπει να συνδυάσετε τα διανύσματα των λέξεων για να δημιουργήσετε διανύσματα προτάσεων τόσο με τη μέθοδο της άθροισης όσο και αυτή των μέσων όρων. Μετά τη δημιουργία των τεσσάρων τελικών μητρώων, καλείστε να εκπαιδεύσετε πάνω στο training dataset ένα SVM και ένα νευρωνικό δίκτυο. Οι παράμετροι εισόδου του SVM και οι λεπτομέρειες της αρχιτεκτονικής του νευρωνικού δικτύου, είναι στην ευχέρεια σας. Αξιολογήστε πάνω στο test dataset την απόδοσή των μοντέλων που εκπαιδεύσατε σύμφωνα με τις μετρικές accuracy, f1 score, precision και recall.

Επιπρόσθετα, σας ζητείται να ομαδοποιήσετε τα δεδομένα σας χρησιμοποιώντας την τεχνική LDA (Latent Dirichlet Allocation) με σκοπό την εξαγωγή θεματικών. Συμφωνούν οι θεματικές που εξαγάγατε με τις προκαθορισμένες κατηγορίες του συνόλου δεδομένων και αν ναι, σε ποιο βαθμό; Διερευνήστε (η διερεύνηση μπορεί να είναι και με θεωρητικά επιχειρήματα) χρήση του LDA ως υποβοηθητικό εργαλείο στον μηχανισμό πρόβλεψης σε συνδυασμό με τα word embeddings (κοιτάξτε σχετικό χρήσιμο υλικό σε https://web.stanford.edu/~jurafsky/slp3/ed3book_jan72023.pdf)

Παραδοτέα

1. Τα αρχεία κώδικα που υλοποιούν τα ζητούμενα των ασκήσεων.
2. Μια αναφορά σε μορφή pdf η οποία θα πρέπει να περιέχει τα ακόλουθα:
    - Αναλυτική καταγραφή του περιβάλλοντος υλοποίησης (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 άσκηση μπορεί να υλοποιηθεί είτε **ατομικά** είτε από **ομάδες δύο ατόμων**.
2. Η άσκηση μπορεί να υποβληθεί έως και **την 28/06/2023 στις 23:59**.
3. Η άσκηση θα εξεταστεί προφορικά στις 30/6/2023
4. Η υποβολή της άσκησης πρέπει να γίνει μέσω του eclass του μαθήματος.
5. Η άσκηση μπορεί να αποσταλεί πολλές φορές αλλά θα βαθμολογηθεί μόνο η τελευταία της υποβολή.
6. Για την εργασία μπορείτε να απευθύνετε απορίες και στον υποψήφιο διδάκτορα κ. Αγοράκη Μπομπότα, mpompotas@ceid.upatras.gr