Τμήμα Μηχανικών Η/Υ & Πληροφορικής

Πανεπιστήμιο Πατρών

Ανάκτηση Πληροφορίας

Εργαστηριακή Άσκηση

Χειμερινό Εξάμηνο 2023

Διδάσκων: Χ. Μακρής

Στόχος

Στα πλαίσια της παρούσας εργασίας ζητείται η μελέτη, η υλοποίηση και η σύγκριση μηχανών αναζήτησης. Θα κληθείτε να επεξεργαστείτε μία συλλογή να εφαρμόσετε μοντέλα ανάκτησης πληροφορίας σε αυτή με δεδομένα ερωτήματα και λίστες σχετικών κειμένων. Με βάση, λοιπόν, τα δεδομένα αυτά θα εφαρμόσετε τις κατάλληλες μετρικές αξιολόγησης και θα καταγράψετε τις παρατηρήσεις σας. Δεν ορίζεται γλώσσα υλοποίησης αλλά προτείνεται η χρήση της Python με τις κατάλληλες βιβλιοθήκες. Οι μετρικές και τα μοντέλα που δεν θα χρησιμοποιησετε κάποια έτοιμη συνάρτηση απο καποια βιβλιοθήκη θα αναφέρεται ρητά στο αντίστοιχο ερώτημα. Για τα υπόλοιπα έχετε την ελευθερία να τα αντιμετωπίσετε όπως εσεις κρίνετε.

Η συλλογή

Η συλλογή που σας έχει δοθεί είναι η Cystic Fibrosis (C.F)[1] [2] και περιλαμβάνει 1209 κείμενα και 100 ερωτήματα. Απο τα 100 ερωτήματα αυτά σας έχει δοθεί ένα υποσύνολο αυτων (20 ερωτήματα) που μπορείτε να χρησιμοποιήσετε. Κάθε αρχείο έχει σαν όνομα το ID του κειμένου και σαν περιεχόμενο τον τίτλο του και το κείμενο. Τα ερωτήματα περιλαμβάνουν μία λίστα κειμένων που θεωρούνται σχετικά με το ερώτημα καθως και ο βαθμός σχετικότητας τους απο ειδικούς, που δύναται να εφαρμοστεί σε κατάλληλες μετρικές.

[1] https://people.ischool.berkeley.edu/~hearst/irbook/cfc.html

[2] Shaw, W.M. & Wood, J.B. & Wood, R.E. & Tibbo, H.R. The Cystic Fibrosis Database: Content and Research Opportunities. LISR 13, pp. 347-366, 1991.

Ερωτήματα

Ερώτημα 1 - Ανάγνωση και Επεξεργασία της Συλλογής (Μονάδα 2)

Στο ερώτημα αυτό θα πρέπει να διαπεράσετε τη συλλογή ώστε να παραχθεί ένα ανεστραμμένο ευρετήριο που θα περιλαμβάνει την πληροφορία που θα χρειαστούν τα μοντέλα στα επόμενα ερωτήματα. Στο σημείο αυτό μπορείτε να κάνετε και οποιαδήποτε μορφή προεπεξεργασίας εσείς κρίνετε απαραίτητη.

Ερώτημα 2 - Υλοποίηση Vector Space μοντέλου (Μονάδες 3)

Στο ερώτημα αυτό θα υλοποιήσετε το μοντέλο ανάκτησης πληροφορίας διανυσματικού χώρου, ***Vector Space Model (VSM)***, χρησιμοποιώντας ως εισοδο το ευρετήριο που δημιουργήσατε στο πρώτο ερώτημα. Επιλεξτε τις δύο καλύτερες εναλλακτικές από τους διάφορους τρόπους που παρέχονται στη βιβλιογραφία[3] (Πίνακες 1 και 2) για τον υπολογισμό των βαρών στα διανύσματα που θα κατασκευάσετε. Σαν απάντηση του ερωτήματος **δεν επιτρέπεται** να παραδώσετε έτοιμες λύσεις που προσφέρουν διάφορες βιβλιοθήκες (π.χ sklearn) αλλά μπορείτε να τις χρησιμοποιήσετε για σύγκριση/επαλήθευση με τη δική σας υλοποίηση.

Τρόπος Υπολογισμου	**TF_{ij}**
Δυαδικό	$\{0, 1\}$
Απλή συχνότητα εμφάνισης	F_{ij}
Απλή λογαριθμικη κανονικοποίηση	$1 + log\ F_{ij}$
Διπλή 0,5 κανονικοποίηση	$0.5 + 0.5 \cdot \frac{F_{ij}}{MAX(F_{ij})}$
Διπλή K κανονικοποίηση	$K + (1 - K) \cdot \frac{F_{ij}}{MAX(F_{ij})}$

***Πίνακας 1**: Υπολογισμός του όρου TF_{ij}*

Τρόπος Υπολογισμου	**IDF_i**
Μοναδιαίο	1
Απλη ανάστροφη συχνότητα εμφάνισης	$log(\frac{N}{n_i})$
Απλή λογαριθμική κανονικοποίηση	$log(1 + \frac{N}{n_i})$
Ανάστροφη κανονικοποίηση περισσότερων εμφανίσεων	$log(1 + \frac{MAX(n_i)}{n_i})$

***Πίνακας 2**: Υπολογισμός του όρου IDF_i*

[3] *Gerard Salton, Christopher Buckley,Term-weighting approaches in automatic text retrieval, Information Processing & Management, Volume 24, Issue 5, 1988,Pages 513-523,ISSN 0306-4573,https://doi.org/10.1016/0306-4573(88)90021-0.*

Ερώτημα 3 - Υλοποίηση colBERT μοντέλου (Μονάδες 2)

Ένα από τα πλεονεκτήματα των μοντέλων που χρησιμοποιούν διανυσματικες αναπαραστάσεις είναι ότι περιέχουν σημασιολογική πληροφορία η οποία έχει προκύψει ως αποτέλεσμα της εκπαίδευσης τους πάνω σε συλλογές κειμένων μεγάλου όγκου. Στο ερώτημα αυτό θα κληθείτε να εφαρμόσετε το μοντέλο *colBERT* έτσι ώστε να απαντάει σε ερωτήματα πάνω στην συλλογή που σας δίνεται **χωρίς** να το **εκπαιδεύσετε** απο την αρχή. Θα πρέπει να προσαρμόσετε την αρχική συλλογή ώστε να μπορεί να τη διαχειριστεί το προεκπαιδευμένο μοντέλο. Ως υλοποίηση μπορει να χρησιμοποιηθεί αυτή του *Stanford*[4] ενώ ένα παράδειγμα εκτέλεσης φαίνεται στο *Notebook*[5] που προσφέρεται.

Ερώτημα 4 - Συγκρίσεις (Μονάδες 3)

Να συγκρίνετε μεταξύ τους τα μοντέλα που αναπτύχθηκαν στα προηγούμενα ερωτήματα με τουλάχιστον 2 μετρικές αξιολόγησης και να ερμηνεύσετε τα αποτελέσματα. Για να ελέγξετε την απόδοση των μοντέλων βάσει των μετρικών που θα επιλέξετε, χρησιμοποιήστε την λίστα ερωτημάτων-σχετικών κειμένων που σας παρέχεται μαζί με την συλλογή στα αρχεία: queries_20, relevant_20, cfquery_detailed.

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 (επισυνάπτεται δείγμα):
   a. Τα στοιχεία (**ΑΜ, ονοματεπώνυμο και email**) του φοιτητή ή των φοιτητών που παραδίδουν την άσκηση.
   b. Αναλυτική καταγραφή του περιβάλλοντος υλοποίησης (γλώσσα προγραμματισμού, βιβλιοθήκες λογισμικού κτλ.) καθώς και τα βήματα που απαιτούνται για την εγκατάστασή του.
   c. Περιγραφή της διαδικασία υλοποίησης.
   d. 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.

[4] https://github.com/stanford-futuredata/ColBERT

[5] https://colab.research.google.com/github/stanford-futuredata/ColBERT/blob/main/docs/intro2new.ipynb

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 **ir2024_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 αποστείλετε μέσω email στη διεύθυνση mpompotas@ceid.upatras.gr (και κοινοποίηση σε makri@ceid.upatras.gr , kalogeropo@ceid.upatras.gr)