Τμήμα Μηχανικών Η/Υ & Πληροφορικής

Πανεπιστήμιο Πατρών

Ανάκτηση Πληροφορίας

Εργαστηριακή Άσκηση

Χειμερινό Εξάμηνο 2024

Διδάσκων: Χ. Μακρής

Διδακτορικοι Επιβλεποντες: Μπομπότας Αγοράκης, Καλογερόπουλος Νικήτας-Ρήγας

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. Για την **ευρετηρίαση**, τα **μοντέλα** και τις **μετρικες** που θα υλοποιήσετε **δε πρέπει** να χρησιμοποιήσετε έτοιμες λύσεις, αν και μπορούν να χρησιμοποιηθούν για επαλήθευση. Παράλληλα θα υλοποιήσετε ένα **chatBot** το οποίο θα δέχεται ερωτήσεις και θα απαντά με βαση τα σχετικά κείμενα. Δεν ορίζεται γλώσσα υλοποίησης αλλά προτείνεται η χρήση της Python με τις κατάλληλες βιβλιοθήκες.

Η συλλογή

Η συλλογή που σας έχει δοθεί είναι η Cystic Fibrosis (C.F) και περιλαμβάνει 1239 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 από ειδικούς, που δύναται να εφαρμοστεί σε κατάλληλες μετρικές.

Ερωτήματα

Ερώτημα 1 - Ανάγνωση και Επεξεργασία της Συλλογής (Μονάδες 1)

Στο ερώτημα αυτό θα πρέπει να διαπεράσετε τη συλλογή ώστε να παραχθούν **τα κατάλληλα** (δύο) **ανεστραμμένα ευρετήρια** που θα περιλαμβάνουν την πληροφορία που θα χρειαστούν **τα μοντέλα στα επόμενα ερωτήματα.** Στο σημείο αυτό μπορείτε να κάνετε και οποιαδήποτε μορφή

προεπεξεργασίας εσείς κρίνετε απαραίτητη. Στη συλλογή θα παρατηρήσετε ότι υπάρχουν κείμενα τα οποία **απουσιάζουν**. Θα πρέπει να διαχειριστείτε το πρόβλημα αυτό ώστε να **μην υπάρχει** κάποια αναντιστοιχία μεταξύ των id των κειμένων και αυτών που βρίσκονται στις λίστες σχετικών ερωτημάτων.

Στην αναφορά σας, εκτος άλλων, θα πρέπει να συμπεριλάβετε:

- *Τη δομή κάθε ευρετηρίου.*
- *Αιτιολόγηση της δομής που επιλέχθηκε.*
- *Τεχνικές προεπεξεργασίας και αιτιολόγηση τους, αν εφαρμόστηκαν.*
- *Τον τρόπο διαχείρισης των κειμένων που απουσιάζουν.*
- *Αν θέλετε να συμπεριλάβετε κώδικα στην αναφορά σας διαμορφώστε μία ενότητα με τίτλο "Παράρτημα" και παρουσίαστε τον σε εκείνο το σημείο μαζί με μία σύντομη περιγραφή.*

Ερώτημα 2 - Υλοποίηση Boolean μοντέλου (Μονάδες 2)

Στο ερώτημα αυτό, θα υλοποιήσετε το μοντέλο **ανάκτησης πληροφορίας Boolean**[1], χρησιμοποιώντας ως είσοδο το **ευρετήριο που δημιουργήσατε στο πρώτο ερώτημα**. Στόχος είναι η ανάκτηση εγγράφων που ικανοποιούν λογικά ερωτήματα με χρήση τελεστών όπως **AND, OR, και NOT**. Θα πρέπει να υλοποιήσετε έναν μηχανισμό που λαμβάνει ως είσοδο **ένα λογικό ερώτημα** και επιστρέφει τα **σχετικά έγγραφα από το ευρετήριο**.

Στην αναφορά σας, εκτος άλλων, θα πρέπει να συμπεριλάβετε:

- *Μια αλγοριθμικη περιγραφή της υλοποίησης.*
- *Τον τρόπο που διαμορφώσατε τα ερωτήματα.*
- *Μία περιγραφή των προβλημάτων που συναντήσατε.*
- *Αν θέλετε να συμπεριλάβετε κώδικα στην αναφορά σας διαμορφώστε μία ενότητα με τίτλο "Παράρτημα" και παρουσίαστε τον σε εκείνο το σημείο μαζί με μία σύντομη περιγραφή.*

Ερώτημα 3 - Υλοποίηση Μοντέλου Διανυσματικού Χώρου (Vector Space Model - VSM) (Μονάδες 2)

Στο ερώτημα αυτό, θα υλοποιήσετε το μοντέλο ανάκτησης πληροφορίας διανυσματικού χώρου[2][3] (Vector Space Model - VSM) χρησιμοποιώντας ως είσοδο το ευρετήριο που δημιουργήσατε στο πρώτο

---

[1] *Lashkari, A.H.; Mahdavi, F.; Ghomi, V. (2009), "A Boolean Model in Information Retrieval for Search Engines", 2009 International Conference on Information Management and Engineering, pp. 385–389, doi:10.1109/ICIME.2009.101,*

[2] *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.*

[3] *G. Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. Commun. ACM 18, 11 (Nov. 1975), 613–620. https://doi.org/10.1145/361219.361220*

ερώτημα. Στόχος είναι η αξιολόγηση της ομοιότητας μεταξύ ενός ερωτήματος και των εγγράφων, βασιζόμενοι στη συσχέτιση των όρων.

Βασικά σημεία που πρέπει να περιλαμβάνει η υλοποίησή σας:

- Αναπαράσταση εγγράφων και ερωτημάτων ως **διανύσματα**: Κάθε έγγραφο και το ερώτημα θα πρέπει να αναπαριστώνται ως διανύσματα με βάση τους όρους που περιέχουν.
- Συντελεστές βαρύτητας όρων: Υπολογισμός των συντελεστών βαρύτητας των όρων, χρησιμοποιώντας τεχνικές όπως το **TF-IDF** (Term Frequency - Inverse Document Frequency).
- Υπολογισμός συσχέτισης: Εφαρμογή του **συνημίτονου της γωνίας (cosine similarit**y) για την εύρεση της ομοιότητας μεταξύ του ερωτήματος και των εγγράφων.
- Ταξινόμηση εγγράφων: Ταξινόμηση των εγγράφων **με βάση τη βαθμολογία ομοιότητας** και επιστροφή των πιο σχετικών εγγράφων.

Στην αναφορά σας, εκτος άλλων, θα πρέπει να συμπεριλάβετε:

- *Μια αλγοριθμικη περιγραφή της υλοποίησης.*
- *Το τρόπο υπολογισμού του βάρους μαζί με μία σύντομη αιτιολόγηση για την συγκεκριμένη επιλογή.*
- *Μία περιγραφή των προβλημάτων που συναντήσατε.*
- *Αν θέλετε να συμπεριλάβετε κώδικα στην αναφορά σας διαμορφώστε μία ενότητα με τίτλο "Παράρτημα" και παρουσίαστε τον σε εκείνο το σημείο μαζί με μία σύντομη περιγραφή.*

Ερώτημα 4 - Συγκρίσεις (Μονάδες 3)

Σε αυτό το ερώτημα, θα συγκρίνετε πρακτικά τα δύο μοντέλα ανάκτησης πληροφορίας που υλοποιήσατε (Boolean και Vector Space) χρησιμοποιώντας το ίδιο σύνολο δεδομένων και ερωτημάτων. Η σύγκριση θα πρέπει να γίνει **βάσει συγκεκριμένων μετρικών που θα βασίζονται** στις εξης:

- **Ακρίβεια (Precision):** Πόσα από τα έγγραφα που ανακτήθηκαν είναι σχετικά με το ερώτημα;
- **Πληρότητα - Ανάκληση (Recall):** Πόσα από τα σχετικά έγγραφα ανακτήθηκαν συνολικά;
- **Χρόνος εκτέλεσης:** Πόσος χρόνος απαιτείται για την εκτέλεση κάθε μοντέλου;

*Η σύγκριση, που θα παρουσιάσετε στην αναφορά σας θα πρέπει να περιλαμβάνει τις **μετρικές** και τους **σχολιασμούς** για την απόδοση των μοντέλων:*

- *Για τα πρώτα κείμενα της επιστρεφόμενης λίστας αλλά και συνολικα για κάθε ερώτημα.*
- *Για όλα τα ερωτήματα της συλλογής.*
- *Για τον χρόνο εκτέλεσης του μοντέλου αλλα και της ευρετηρίασης του.*
- *Αν θέλετε να συμπεριλάβετε κώδικα στην αναφορά σας διαμορφώστε μία ενότητα με τίτλο "Παράρτημα" και παρουσίαστε τον σε εκείνο το σημείο μαζί με μία σύντομη περιγραφή.*

Για να ελέγξετε την απόδοση των μοντέλων βάσει των μετρικών, χρησιμοποιήστε την λίστα ερωτημάτων-σχετικών κειμένων που σας παρέχεται μαζί με την συλλογή στα αρχεία: queries_20, relevant_20.

Ερώτημα 5 - Δημιουργία Chatbot με βάση τα Αποτελέσματα των Boolean και Vector Space Μοντέλων (Μονάδες 1)

Σε αυτό το ερώτημα, θα δημιουργήσετε δύο εκδοχές ενός chatbot, χρησιμοποιώντας τα σχετικά κείμενα που ανακτήθηκαν από κάθε μοντέλο (Boolean και Vector Space Model - VSM). Ο αριθμός των κειμένων που θα επιλέξετε εξαρτάται από εσάς. Το chatbot θα επιτρέπει στους χρήστες να κάνουν ερωτήσεις πάνω στα κείμενα που ανακτήθηκαν και θα πρέπει να επιστρέφει σχετικές απαντήσεις, βασισμένο στις πληροφορίες που παρέχουν τα κείμενα.
Για την υλοποίηση του συστήματος προτείνεται η χρήση ενός μεγάλου γλωσσικού μοντέλου από τη βιβλιοθήκη του OLLAMA[4], σε συνδυασμό με το πλαίσιο εργασίας LangChain. Ακολουθήστε τα παρακάτω βήματα για την προετοιμασία του περιβάλλοντος:

1. Εγκατάσταση του OLLAMA: Ξεκινήστε εγκαθιστώντας το OLLAMA στο σύστημά σας, σύμφωνα με τις οδηγίες που παρέχονται στο GitHub του OLLAMA[5].

2. Επιλογή Μοντέλου: Μετά την εγκατάσταση, επιλέξτε το κατάλληλο γλωσσικό μοντέλο που ταιριάζει στις ανάγκες και δυνατότητες του συστήματος. Βεβαιωθείτε ότι το μοντέλο έχει το σωστό μέγεθος και αριθμό παραμέτρων. Προσοχή: η επιλογή πολύ μεγάλου μοντέλου μπορεί να οδηγήσει σε δυσκολίες κατά την υλοποίηση, καθώς ενδέχεται να απαιτεί υπερβολικούς πόρους.

3. Κατέβασμα του Μοντέλου: Αφού επιλέξετε το κατάλληλο μοντέλο και βεβαιωθείτε ότι το OLLAMA λειτουργεί στο παρασκήνιο, χρησιμοποιήστε την εντολή ollama **pull** <όνομα μοντέλου> για να κατεβάσετε το μοντέλο τοπικά. Για παράδειγμα: ollama pull llama3

4. Εγκατάσταση των βιβλιοθηκών LangChain: Μετά την εγκατάσταση του OLLAMA και την τοπική αποθήκευση του επιλεγμένου μοντέλου, θα χρειαστεί να εγκαταστήσετε τις βιβλιοθήκες που σχετίζονται με το LangChain framework. Προτείνεται να εγκαταστήσετε τις παρακάτω βιβλιοθήκες:
    - langchain
    - langchain-community
    - langchain-core

[4] https://ollama.com/download
[5] https://github.com/ollama/ollama

5. Ενδεικτικό Παράδειγμα: Για να ξεκινήσετε με την υλοποίηση, μπορείτε να συμβουλευτείτε ένα μικρό παράδειγμα κώδικα στο σύνδεσμο.

Ακολουθώντας αυτά τα βήματα, θα είστε έτοιμοι να ξεκινήσετε με την ανάπτυξη και υλοποίηση του συστήματος. Η διεπαφή του χρήστη με το σύστημα **δε χρειάζεται** να είναι **σύνθετη** και μπορεί να γίνει **μέσω γραμμής εντολών**.

Στην αναφορά σας να παρουσιάσετε εικόνες εκτέλεσης του chatbot και να απαντήσετε στα ερωτήματα:

- *Ποιο chatbot δίνει καλύτερες απαντήσεις;*
- *Είναι οι απαντήσεις του Boolean chatbot πιο ακριβείς αλλά περιορισμένες, ενώ του VSM chatbot πιο ευέλικτες αλλά ίσως λιγότερο ακριβείς;*
- *Τι προβλήματα συναντήσατε;*
- *Ποια μειονεκτήματα ή περιορισμούς πιστεύετε ότι έχει η υλοποίηση σας;*
- *Αν θέλετε να συμπεριλάβετε κώδικα στην αναφορά σας διαμορφώστε μία ενότητα με τίτλο "Παράρτημα" και παρουσίαστε τον σε εκείνο το σημείο μαζί με μία σύντομη περιγραφή.*

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 μεταγενέστερη της ανακοίνωσης του προγράμματος της εξεταστικής. Κατα την εξεταση καλείστε να αιτιολογήσετε **οτιδήποτε** έχετε γράψει στην **αναφορά** σας, αλλά και στον **κωδικά** σας. Σε περίπτωση αδυναμίας επαρκούς τεκμηρίωσης, θα αφαιρούνται οι μονάδες που αντιστοιχούν στο εκάστοτε ερώτημα.
5. Η αναφορά σας βαθμολογείται ως προς την πληρότητα με **1 μονάδα**.

6. 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 **ir2025_AM1_AM2.zip**
7. Η άσκηση μπορεί να αποσταλεί πολλές φορές αλλά θα βαθμολογηθεί **μόνο η τελευταία** της υποβολή.
8. 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)