|  |
| --- |
|  |
| page1image29306176  Τμήμα Μηχανικών Η/Υ και Πληροφορικής  Πανεπιστήμιο Πατρών  Πολυτεχνική Σχολή  Τομέας Λογικού των Υπολογιστών  Διδάσκων: Χρήστος Μακρής  Ακαδημαϊκό Έτος: 2023 – 2024  Ημ / νία Παράδοσης: 28 / 01 / 2024 |

Ανάκτηση Πληροφορίας

Επιλεγόμενο Μάθημα – CEID\_NE5597

Εργαστηριακή Άσκηση

Χειμερινό Εξάμηνο 2023

Χρυσαυγή Πατέλη | 1084513 | up1084513@ac.upatras.gr

Μηλτιάδης Μαντές | 1084661 | up1084661@ac.upatras.gr

ΠΕΡΙΕΧΟΜΕΝΑ

**1 ΕΙΣΑΓΩΓΗ 1.1** Ερώτημα 1 …………………………………………………………………………………….. 3 **1.2** Ερώτημα 2 …………………………………………………………………………………….. 4 **1.3** Ερώτημα 3 …………………………………………………………………………………….. 7 **1.4** Ερώτημα 4 …………………………………………………………………………………….. 9 **1.5** Περιβάλλον Υλοποίησης και Βιβλιοθήκες ……………………………………………….11

**2 ΥΛΟΠΟΙΗΣΗ 2.1** Ερώτημα 1 …………………………………………………………………………………….14 **2.2** Ερώτημα 2 …………………………………………………………………………………….16  **2.3** Ερώτημα 3 …………………………………………………………………………………….20 **2.4** Ερώτημα 4 …………………………………………………………………………………….23

**3 ΑΠΟΤΕΛΕΣΜΑΤΑ – ΠΑΡΑΤΗΡΗΣΕΙΣ 3.1** Ερώτημα 1.…………………………………………………………………………………….28 **3.2** Ερώτημα 2 …………………………………………………………………………………….  **3.3** Ερώτημα 3 ……………………………………………………………………………………. **3.4** Ερώτημα 4 …………………………………………………………………………………….

**4 ΑΝΑΦΟΡΕΣ 4.1** Αναφορές.…………………………………………………………………………………….

**5 ΠΑΡΑΡΤΗΜΑ 5.1** Κώδικας ....…………………………………………………………………………………….

ΕΙΣΑΓΩΓΗ

* 1. **Ερώτημα 1**

Στο συγκεκριμένο ερώτημα μας ζητείται να υλοποιήσουμε την δομή του **Ανεστραμμένου Ευρετηρίου** (Ιnverted Ιndex) πάνω σε μια συλλογή 1209 κειμένων.

Το ανεστραμμένο ευρετήριο είναι μια δομή δεδομένων, η οποία χρησιμοποιείται σε συστήματα ανάκτησης πληροφοριών, μηχανές αναζήτησης και γενικά σε συστήματα που είναι απαραίτητη η αναζήτηση κειμένων. Στόχος του είναι η γρήγορη αναζήτηση και εύρεση όλων των εγγράφων που περιέχουν έναν συγκεκριμένο όρο ή φράση.

Το αναστραμμένο ευρετήριο αποτελείται από δυο μέρη, το λεξικό που περιέχει τους διακριτούς όρους όλης της συλλογής και για κάθε όρο του λεξικού τις λίστες εμφανίσεων, οι οποίες περιέχουν πληροφορίες για την εμφάνιση των λέξεων στα κείμενα. Οι λίστες αυτές ονομάζονται ανεστραμμένες λίστες. Η πιο απλή μορφή του για κάθε όρο περιέχει το πλήθος των εγγράφων που περιέχεται και τα id των εγγράφων. Υπάρχουν και άλλες εκδοχές, όπου στην ανεστραμμένη λίστα μαζί με το id των κειμένων που περιέχουν την λέξη υπάρχουν και άλλες πληροφορίες σχετικά με την λέξη, η εκδοχή που χρησιμοποιήσαμε εμείς περιέχει την συχνότητα της λέξης μέσα στο συγκεκριμένο κείμενο όπως φαίνεται και στην παρακάτω εικόνα.

Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, οθόνη, γραμματοσειρά

Περιγραφή που δημιουργήθηκε αυτόματα

**Εικόνα 1**: Προσομοίωση Ανεστραμμένου Ευρετηρίου

Το πλεονέκτημα του **Ανεστραμμένου Ευρετηρίου** είναι ότι επιτρέπει γρήγορες, αποδοτικές αναζητήσεις ερωτημάτων και ανάλογα την μορφή μπορεί να απαντήσει σε διάφορα είδη ερωτημάτων (ερωτήματα πολλαπλών λέξεων, μονή λέξης, τομή λίστας, σύνθετα ερωτήματα, ερωτήματα φράσης και εγγύτητας, boolean). Ωστόσο έχουν και μειονεκτήματα, καθώς έχουν υψηλό κόστος συντήρησης, αποθήκευσης και απαιτούν πολύ χώρο στην μνήμη για την αποθήκευση και την ενημέρωσή τους. Γι’ αυτό και συχνά εφαρμόζονται τεχνικές συμπίεσης (κωδικοποίηση δέλτα, κωδικοποίηση γάμμα, η κωδικοποίηση μεταβλητών byte) πάνω στο **Aνεστραμμένο Eυρετήριο**, για περαιτέρω μείωση του απαιτούμενου χώρου.

**1.2 Ερώτημα 2**

Στο συγκεκριμένο ερώτημα μας ζητείται να υλοποιήσουμε το **Μοντέλο Διανυσματικού Χώρου** (Vector Space Model) πάνω στο ανεστραμμένο αρχείο που εξάγαμε στο προηγούμενο ερώτημα.

Το **Μοντέλο Διανυσματικού Χώρου** χρησιμοποιεί την έννοια των διανυσμάτων στον πολυδιάστατο χώρο προκειμένου να αναπαραστήσει αλγεβρικά την φυσική γλώσσα των κειμένων της συλλογής εγγράφων. Πιο συγκεκριμένα, το σκεπτικό πίσω από αυτό το σύστημα ανάκτησης πληροφορίας έγκειται στην ευκολία εύρεσης κοινών όρων (terms) ανάμεσα σε κείμενα και ερωτήματα χρήστη (queries) αν τα ζυγισμένα διανύσματα που τα αναπαριστούν τοποθετηθούν σε διανυσματικό χώρο *V*.

Σε μια πιο γενική περιγραφή του μοντέλου ισχύουν τα εξής:

1. Έστω {t1, t2, …, tn} οι όροι ενός κειμένου di, i = 1, 2, …, m από την συλλογή μας. Υποθέτοντας χωρίς βλάβη γενικότητας ότι για κάθε διακριτό όρο ti υπάρχει στο χώρο ένα μοναδιαίο ορθοκανονικό διάνυσμα **ti** στο οποίο αντιστοιχίζεται, τότε ο διανυσματικός χώρος *V* θα είναι το span{ti} και έτσι το διάνυσμα **di** του κειμένου μπορεί να δοθεί ως γραμμικός συνδυασμός των **ti**.

**di =**

1. Έστω {T1, T2, …, Tn} οι όροι ενός ερωτήματος, οι οποίοι σχετίζονται με το σύνολο {t1, t2, …, tn}. Για κάθε ερώτημα q όμοια προκύπτει ότι το αντίστοιχο του διάνυσμα **q** του διανυσματικού χώρου *V* θα είναι πάλι γραμμικός συνδυασμός των διανυσμάτων **Τi**.

Εικόνα που περιέχει γραμμή

Περιγραφή που δημιουργήθηκε αυτόματα**q =**

Κάθε διακριτός όρος σε κάθε περίπτωση αντιπροσωπεύει μια διάσταση του χώρου μας. Συνεπώς, στόχος μας είναι στο χώρο *V* για κάθε διάνυσμα **q** να υπολογίσουμε την απόστασή του από κάθε διάνυσμα **di**. Για να γίνει αυτό επιδιώκουμε να προσεγγίσουμε τη γωνία που σχηματίζεται ανάμεσά τους και πιο συγκεκριμένα το συνημίτονό της, cos(θ), το οποίο δίνεται από το εσωτερικό γινόμενο.

**θ**

**q**

**di**

cos(θ) =

Στο τέλος, χρησιμοποιούμε ως κριτήριο ταξινόμησης τη τιμή του συνημίτονου που προκύπτει προκειμένου να διατάξουμε τα κείμενα μας με βάση τη σχετικότητα τους ως προς κάθε ερώτημα χρήστη.

Ειδικότερα τώρα, στη δικιά μας υλοποίηση επιλέγουμε για απλοποίηση τα διανύσματα **di**, **qj** να έχουν ως διάσταση το πλήθος των διακριτών όρων του κειμένου και του ερωτήματος αντίστοιχα και κάθε γραμμή του διανύσματος **di** και **qj** να περιέχει μια boolean τιμή, με το 0 να συμβολίζει την μη – ύπαρξη του συγκεκριμένου όρου στο διάνυσμα κειμένου και το 1 να συμβολίζει προφανώς την ύπαρξή του.

Εικόνα που περιέχει κείμενο, γραμμή, γραμματοσειρά, αριθμός

Περιγραφή που δημιουργήθηκε αυτόματαΤώρα, για την καλύτερη εφαρμογή του **Vector Space Model** είναι απαραίτητη η αξιοποίηση της τεχνικής του *term – weighting*, καθώς οι διαφορετικοί όροι μπορεί να έχουν διαφορετική σημασία και άρα και επίδραση στα συμφραζόμενα του κειμένου. Αρχικά, όσον αφορά το local term – weighting, επιλέγουμε μια από τις ακόλουθες τεχνικές υπολογισμού του term frequency (TF), καθώς μας αφορά μόνο η συχνότητα με την οποία εμφανίζονται οι όροι σε κάθε κείμενο ανεξάρτητα, την οποία και έχουμε αποθηκευμένη στο ανεστραμμένο αρχείο.

**Εικόνα 3**: Τρόποι Υπολογισμού Term Frequency

Η πρώτη μας σκέψη είναι να επιλέξουμε την *Απλή Συχνότητα Εμφάνισης*, η οποία υπολογίζει πόσες φορές εμφανίζεται ένας όρος στο κείμενο. Άλλωστε, όσο περισσότερες φορές εμφανίζεται ένας όρος στο κείμενο, τόσο πιο πιθανό είναι το κείμενο αυτό να είναι σχετικό με το ερώτημά μας. Ωστόσο, το μειονέκτημα αυτού του σκεπτικού είναι κυρίως ότι ευνοεί υπερβολικά τους όρους που εμφανίζονται με αυξημένη συχνότητα, καθώς η μεγάλη συχνότητα εμφάνισης δεν μας εξασφαλίζει υποχρεωτικά ότι οι όροι αυτοί θα είναι και πιο σημαντικοί από κάποιους όρους που εμφανίζονται πιο σπάνια. Επίσης, η χρήση δυαδικών βαρών είναι και αυτή αρκετά περιοριστική. Γι’ αυτό καταλήγουμε στην *Απλή Λογαριθμική Κανονικοποίηση*, η οποία προσπαθεί να αντισταθμίσει την επίδραση της *Απλής Συχνότητας Εμφάνισης* μειώνοντας μέσω του λογαρίθμου την επίδραση πολύ μεγάλων διαφορών στις συχνότητες εμφάνισης, έτσι ώστε να έχουμε δικαιότερη απόδοση βαρών στο τελικό διάνυσμα **di**.

Έπειτα, όσον αφορά το global term – weighting, θέλουμε να εξετάσουμε πώς η κατανομή και η συχνότητα εμφάνισης ενός όρου σε όλα τα κείμενα της συλλογής επηρεάζει το πόσο σημαντικός είναι. Για το λόγο αυτό επιλέγουμε την αποδοτικότερη από τις πιο κάτω τεχνικές για τον υπολογισμό του Inverted Document Frequency (IDF):

Εικόνα που περιέχει κείμενο, γραμμή, γραμματοσειρά, αριθμός

Περιγραφή που δημιουργήθηκε αυτόματα**Εικόνα 4**: Τρόποι Υπολογισμού Inverted Document Frequency

Εδώ επιλέγουμε την τεχνική της Απλής Ανάστροφης Συχνότητας Εμφάνισης, η οποία «λογαριθμίζει» την αναλογία των κειμένων μέσα στη συλλογή μας που περιέχουν ένα συγκεκριμένο όρο. Η τιμή της αυξάνεται όσο το πλήθος των κειμένων στο οποίο εντοπίζεται αυτός ο όρος μειώνεται και έτσι ψηλή τιμή σημαίνει συνήθως ότι ένας όρος εμφανίζεται στο συγκεκριμένο έγγραφο με μεγαλύτερη συχνότητα από το συνηθισμένο.

Έχοντας ολοκληρώσει αυτή τη διαδικασία, πολλαπλασιάζουμε κάθε συντεταγμένη των διανυσμάτων μας με τα τελικά βάρη και τα διανύσματα που προκύπτουν θα είναι:

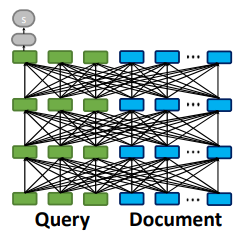
**di** = [di1, di2, di3, …, dik] με dik να συμβολίζει το βάρος του k – όρου στο έγγραφο i **qj** =[qj1, qj2, qj3, …, qjk] με qjk να συμβολίζει το βάρος του k – όρου στο ερώτημα j

και συνεπώς το συνημίτονο της γωνίας που θα σχηματίζουν θα δίνεται από τη σχέση:

cos(**di**, **qj**) =

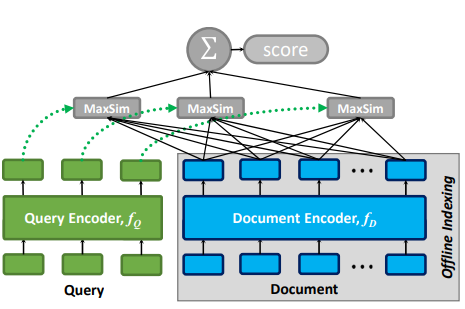
Σημειώνουμε εδώ ότι τα βάρη των όρων που δεν εντοπίζονται μέσα σε κάποιο κείμενο θα παραμείνουν 0 λόγω του πολλαπλασιασμού με τα βάρη tf-idf όπως προείπαμε.

Μια συνηθισμένη τεχνική αποτελεί επίσης η κανονικοποίηση του μήκους των διανυσμάτων **di**, **qi** (length – normalization) διαιρώντας κάθε συντεταγμένη τους με την L2 νόρμα τους, προκειμένου και να γίνουν και αυτά μοναδιαία όπως αυτά που ορίζουν τη βάση του χώρου *V* στον οποίον ανήκουν. Χάριν απλότητας επιλέγουμε να παραλείψουμε αυτό το βήμα, καθώς η χρησιμότητά του είναι κυρίως να κάνει τα βάρη των πολύ μικρών ή πολύ μεγάλων κειμένων ακόμα πιο συγκρίσιμα.

**1.3 Ερώτημα 3**

Μια πολύ κοινή τακτική που χρησιμοποιείται στην κατάταξη κειμένων με βάση την ομοιότητά τους είναι η χρήση αρχιτεκτονικών που αξιοποιούν νευρωνικά δίκτυα, οι οποίες είναι γνωστές και ως transformers. Η πιο γνωστή περίπτωση τέτοιου transformer είναι το **BERT**, το οποίο σε συνδυασμό με τη self – supervised προεκπαίδευση (pretraining) του νευρωνικού δικτύου αποτελεί τη βάση για την λειτουργία του Μοντέλου **ColBERT** που μελετάμε στη συγκεκριμένη εργασία.

**Εικόνα 5**: BERT

Πιο συγκεκριμένα, το **ColBERT** εισάγει για πρώτη φορά την έννοια των αρχιτεκτονικών καθυστερημένης αλληλεπίδρασης (late interaction) ανάμεσα στα έγγραφα και τα ερωτήματα, όπου πρώτα κωδικοποιείται το κείμενο και το ερώτημα σε δύο ανεξάρτητα το ένα από το άλλο σύνολα από contextual embeddings ED kai EQ με τη χρήση **BERT**s και στη συνέχεια μοντελοποιεί αποτελεσματικά και λεπτομερώς τις πολύ μεγάλες ομοιότητες που προκύπτουν. H μοντελοποίηση γίνεται με πολύ γρήγορους υπολογισμούς και στα δύο σετ από embeddings χωρίς να είναι έτσι απαραίτητος ο προσδιορισμός όλων των πιθανών υποψηφίων για την δημιουργία της τελικής κατάταξης σχετικότητας. Με την εισαγωγή της καθυστέρησης προσφέρεται η δυνατότητα ακόμα πιο ακριβούς υπολογισμού των τιμών ομοιότητας ανάμεσα στα κείμενα και τα ερωτήματα, η οποία δίνεται ως το άθροισμα των τελεστών MaxSim από κάθε embedding. Γενικά, το **ColBERT** απομονώνει τους υπολογισμούς ανάμεσα στα κείμενα και τα ερωτήματα επιτρέποντας έτσι τον εκ των προτέρων υπολογισμό των αναπαραστάσεων των εγγράφων offline. Έτσι, η δεικτοδότηση των εγγράφων μέσα στον κωδικοποιητή γίνεται αποδοτικότερα. Τέλος, ο υπολογισμός των τελεστών MaxSim γίνεται με την ακόλουθη διαδικασία: Αρχικά, τα τελικά emdeddings κανονικοποιούνται ώστε καθένα από αυτά να έχει μοναδιαία L2 νόρμα. Έπειτα, βρίσκουμε τα εσωτερικά γινόμενα των δύο emeddings, τα οποία θα ταυτίζονται με το cosine similarity εφόσον τα embeddings έχουν νόρμα 1. Τέλος, από όλες τις τιμές κρατάμε τη μέγιστη τιμή cosine similarity κάθε διανύσματος ερωτήματος q στο EQ με όλα τα διανύσματα **d** στο ED και έπειτα συνδυάζουμε τα αποτελέσματα αυτά σε μία τιμή score μέσω του τελεστή αθροίσματος (Σ).

Όλη αυτή η διαδικασία φαίνεται στο διπλανό σχήμα:

**Εικόνα 6**: ColBERT

Σε μια γενική περιγραφή οι κωδικοποιητές fQ και fD λειτουργούν ως εξής:

1. **Document Encoder fD**

Αρχικά, κάθε έγγραφο d τμηματοποιείτaι σε διακριτά tokens d1, d2, …, dm στα οποία προστίθενται ένα token εκκίνησης [CLS] του BERT ακολουθούμενο από ένα token [D] που δηλώνει τη θέση του εκάστοτε εγγράφου στην ακολουθία. Έπειτα, περνάμε σαν είσοδο στον κωδικοποιητή την ακολουθία που προκύπτει μετά τις πιο πάνω προσθήκες και αυτός αναλαμβάνει να την φιλτράρει αφαιρώντας τα embeddings που αντιστοιχούν σε σημεία στίξης με βάση μια προκαθορισμένη λίστα. Έτσι, το πλήθος των τελικών embeddings που προκύπτει είναι σημαντικά μικρότερο σε σχέση με πριν, αφού κάνουμε την υπόθεση ότι τα embeddings των σημείων στίξης δεν επηρεάζουν σημαντικά την απόδοση του μοντέλου.

1. **Query Encoder fQ**

Όμοια τμημaτοποιούμε κάθε ερώτημα σε διακριτά tokens q1, q2, …, qn στα οποία προσθέτουμε πάλι το token [CLS] του **BERT** μαζί με το πρόθεμα [Q]. Αν μετά από αυτές τις προσθήκες το ερώτημα συνεχίζει να έχει λιγότερο πλήθος tokens από ένα pre – defined threshold, τότε επεκτείνουμε την ακολουθία με ειδικά tokens [mask] μέχρι να φτάσουμε στο επιθυμητό μήκος. Πάλι περνάμε σαν είσοδο την τελική ακολουθία και ο κωδικοποιητής αναλαμβάνει να πραγματοποιήσει μια αναπαράσταση του κάθε token με βάση το περιεχόμενό του.

**1.4 Ερώτημα 4**

Στόχος μας είναι να συγκρίνουμε την απόδοση των δύο μοντέλων χρησιμοποιώντας ένα πλήθος διαφορετικών μετρικών, προκειμένου να έχουμε μεγαλύτερη αξιοπιστία στη τελική μας σύγκριση. Επιλέγουμε να χρησιμοποιήσουμε μετρικές αξιολόγησης σχετικότητας και πιο συγκεκριμένα τις πιο κάτω μετρικές:

1. **Precision – Recall Curve**

**1.1** Ηακρίβεια προσδιορίζει την ικανότητα του μοντέλου ανάκτησης πληροφορίας που διαθέτουμε να επιστρέφει τα κορυφαία σχετικά κείμενα με βάση το score τους κατά τον έλεγχο σχετικότητας. Στη προκειμένη περίπτωση, τα score της τελικής κατάταξης των ανακτημένων εγγράφων από τη συλλογή προκύπτουν από την εφαρμογή των **Vector Space Model** και **ColBERT**. Με άλλα λόγια, εκφράζει το ποσοστό των ανακτημένων κειμένων, τα οποία είναι όντως σχετικά με το ερώτημα του χρήστη. Το πλήθος και ο βαθμός σχετικότητας των σχετικών αυτών κειμένων καθορίζεται από κάποιους ειδικούς, οι οποίοι τα κατατάσσουν αυτά σε κλίμακα εύρους 0 – 2, με 0 να σημαίνει «καθόλου σχετικό» και 2 «αρκετά σχετικό».

Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, γραμματοσειρά, κύκλος

Περιγραφή που δημιουργήθηκε αυτόματα**1.2** Η ανάκληση εκφράζει την ικανότητα του μοντέλου μας να επιστρέψει όλη τη συλλογή με τα σχετικά κείμενα, την οποία αναφέραμε ακριβώς πιο πάνω, δηλαδή το ποσοστό των σχετικών κειμένων τα οποία απαντάνε στα ερωτήματα του χρήστη.

**Εικόνα 7**: Σχέση ανάμεσα σε Ανάκληση και Ακρίβεια

1. **Mean Average Precision (MAP)**

Η Μέση Ακρίβεια (AP) δίνει έμφαση στην τοποθέτηση των πιο σχετικών εγγράφων στις πιο υψηλές θέσεις της κατάταξης. Πιο συγκεκριμένα, υπολογίζουμε το μέσο όρο των τιμών του precision από όλα τα σχετικά κείμενα της συλλογής για κάθε ερώτημα χρήστη. Συνεπώς, για τον υπολογισμό του MAP θα αξιοποιήσουμε την έννοια της Μέσης Ακρίβειας (AP), καθώς τώρα υπολογίζουμε τη μέση τιμή όλων των τιμών του AP από όλα τα ερωτήματα χρήστη. Δίνεται από τη σχέση:

**MAP =**

**Qj**: ο αριθμός των σχετικών κειμένων για το ερώτημα j

**N**: ο αριθμός των ερωτημάτων

**P(doci)**: η τιμή της ακρίβειας του σχετικού κειμένου i

1. **precision@k**

Υπάρχουν περιπτώσεις στις οποίες είναι απαραίτητο να συγκρίνουμε την απόδοση των μοντέλων ανάκτησης για ατομικές πληροφοριακές ανάγκες. Ο λόγος που μας οδηγεί σε αυτή την απόφαση είναι ότι η χρήση Mέσων Tιμών (MAP) που προκύπτουν από την εκτέλεση διαφόρων ερωτημάτων μπορεί να αποκρύπτει σημαντικές ανωμαλίες στο μοντέλο ανάκτησης που εξετάζεται, οπότε πρέπει για μεγαλύτερη ασφάλεια στην εκτίμηση απόδοσης να μελετήσουμε αν το ένα μοντέλο είναι καλύτερο από το άλλο για κάθε μία από τις πρότυπες πληροφοριακές ανάγκες. Στις περιπτώσεις αυτές, μόνο μία τιμή precision υπολογίζεται για κάθε ερώτημα και η τιμή αυτή θεωρούμε ότι λειτουργεί ως σύνοψη του συνολικού Διαγράμματος Ακρίβειας – Ανάκλησης. Συνήθως αυτή η τιμή είναι η ακρίβεια σε κάποιο προκαθορισμένο επίπεδο k στην ουρά ανάκλησης. Δίνεται από τη σχέση:

**precision@k =**

**rj** = 1 αν το κείμενο στη θέση j είναι σχετικό αλλιώς 0.

Οι μετρικές 2, 3 μας επιστρέφουν μια μοναδική τιμή, οπότε για να συγκρίνουμε την απόδοση των δύο μοντέλων αρκεί να συγκρίνουμε τις τιμές των **MAP** και **precision@k**. Όσο μεγαλύτερες είναι αυτές οι τιμές, τόσο πιο αποδοτικά ανακτά το μοντέλο μας τα σχετικά έγγραφα. Για τη 1η περίπτωση των γραφικών precision – recall για να συγκρίνουμε τις αποδόσεις, μια καλή τεχνική θα ήταν να υπολογίσουμε το εμβαδόν της περιοχής που περικλείεται ανάμεσα στις γραφικές και στον άξονα x. Όμοια με πριν, όσο μεγαλύτερη είναι η τιμή του εμβαδού που υπολογίζουμε, τόσο πιο αποδοτικό είναι το μοντέλο μας. Ωστόσο, επειδή υπάρχει περίπτωση για κάποια ερωτήματα να λειτουργεί καλύτερα το ένα μοντέλο (**VSM**) και για κάποια το άλλο (**ColBERT**), υπολογίζουμε τη μέση τιμή των εμβαδών για όλα τα ερωτήματα, προκειμένου να έχουμε μια συνολική εικόνα για το ποιο από τα δύο μοντέλα λειτουργεί καλύτερα σε γενικές γραμμές.

* 1. **Περιβάλλον Υλοποίησης και Βιβλιοθήκες**

Όλα τα παραπάνω τα οποία εν συντομία αναπτύχθηκαν θεωρητικά, καλούμαστε τώρα να τα υλοποιήσουμε και σε κώδικα. Για τα ερωτήματα 1, 2, 3 επιλέγουμε να χρησιμοποιήσουμε ως περιβάλλον ανάπτυξης αυτό της *PyCharm*, το οποίο διαθέτουμε ήδη εγκατεστημένο από το *JetBrains*. Η έκδοση που χρησιμοποιούμε είναι η 2023.3.2 και επιπλέον οι βιβλιοθήκες που εισάγουμε κατά την εκτέλεση του προγράμματος είναι οι εξής:

1. **numpy**

Από αυτή τη βιβλιοθήκη αξιοποιούμε τις συναρτήσεις **np.linalg.norm( )** και **np.dot( )**, προκειμένου να εφαρμόσουμε στο Ερώτημα 2 πάνω στα dataframes μας την πράξη της Ευκλείδιας νόρμας (L2) και του εσωτερικού γινομένου στοιχείο προς στοιχείο.

1. **pandas**

Από αυτή τη βιβλιοθήκη αξιοποιούμε την **pd.Dataframe( )**, προκειμένου να κατασκευάσουμε όλα τα dataframes στα οποία θα αποθηκεύουμε τις διάφορες τιμές που προκύπτουν κατά τη διαδικασία παραγωγής των τελικών scores του cosine similarity στο Ερώτημα 2.

1. **csv**

Από αυτή τη βιβλιοθήκη αξιοποιούμε τη συνάρτηση **csv.reader(csvfile, dialect='excel', \*\*fmtparams)** προκειμένου να διαβάσουμε στο Ερώτημα 4 τα ανακτηθέντα έγγραφα που μας επιστρέφει το μοντέλο **ColBERT**, τα οποία έχουμε μεταβιβάσει σε ένα αρχείο csv από το περιβάλλον του *Google Collab* σε αυτό της *PyCharm*.

1. **math**

Από αυτή τη βιβλιοθήκη αξιοποιούμε τη συνάρτηση **math.log10(*x*)** για τον υπολογισμό του δεκαδικού λογαρίθμου στις συχνότητες TF και IDF όταν καθορίζουμε τα βάρη των διανυσμάτων μας στο Ερώτημα 2.

1. **pyplot**

Από αυτή τη βιβλιοθήκη αξιοποιούμε όλες τις απαραίτητες συναρτήσεις για το Ερώτημα 4 προκειμένου να χαράξουμε σωστά τις γραφικές παραστάσεις των δύο μοντέλων στη μετρική Precision – Recall. Τέτοιες συναρτήσεις είναι οι: plt.figure( ), …

1. **collections**

Από αυτή τη βιβλιοθήκη εισάγουμε τη κλάση **class collections.defaultdict(default\_factory=None, /[, ...])**, η οποία είναι ιδιαίτερα χρήσιμη για την κατασκευή του **Ανεστραμμένου Ευρετηρίου** στο Ερώτημα 1, καθώς μας επιτρέπει να προσθέτουμε αντικείμενα (δηλαδή δυάδες της μορφής [ID κειμένου, πλήθος εμφανίσεων όρου]) στη λίστα που αντιστοιχεί στο εκάστοτε key (δηλαδή κάθε μοναδικός όρος) χωρίς να χρειάζεται να ελέγξουμε αν αυτό υπάρχει ήδη στο λεξικό.

1. **os**

Από αυτή τη βιβλιοθήκη αξιοποιούμε τη συνάρτηση **os.chdir(path)** προκειμένου να μεταβούμε στο Ερώτημα 1 στο φάκελο docs που περιέχει τα 1209 κείμενα της συλλογής και στη συνέχεια να ακολουθήσουμε τη διαδικασία που περιγράφεται στην υποενότητα **2.1** για την κατασκευή του **Ανεστραμμένου** **Ευρετηρίου**.

1. **ast**

Από αυτή τη βιβλιοθήκη αξιοποιούμε τη συνάρτηση **ast.literal\_eval(node\_or\_string)**, καθώς σε κάθε ερώτημα στο αρχείο CSV όπου έχουμε αποθηκευμένα τα αποτελέσματα του **ColBERT** αντιστοιχεί ένα αλφαριθμητικό το οποίο αντιπροσωπεύει τη λίστα με τα ταξινομημένα ανακτηθέντα κείμενα. Συνεπώς, είναι απαραίτητο κάθε ένα από τα αλφαριθμητικά αυτά να μετατραπούν σε πραγματικές λίστες με περιεχόμενο ακεραίους αριθμούς, δηλαδή τα IDs των κειμένων.

1. **re**

Από αυτή τη βιβλιοθήκη αξιοποιούμε τη συνάρτηση

Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, γραμματοσειρά, αριθμός

Περιγραφή που δημιουργήθηκε αυτόματαΕπιπλέον, για την υλοποίηση του Μοντέλου **ColBERT** εργαστήκαμε στο περιβάλλον του *Google Collab* εφαρμόζοντας τις εξής ρυθμίσεις:

Αρχικά, εγκαθιστούμε το πακέτο virtualenv, το οποίο είναι απαραίτητο για την δημιουργία ενός εικονικού περιβάλλοντος. Έπειτα, δημιουργούμε ένα καινούριο εικονικό περιβάλλον που το ονομάζουμε myenv και το ενεργοποιούμε ώστε να απομονώσει το περιβάλλον της *Python* για να εμποδίσει τυχόν επικαλύψεις με άλλα πακέτα.

Οι βιβλιοθήκες που χρησιμοποιούμε είναι:

**torch**: είναι συντομογραφία για την ανοικτού κώδικα βιβλιοθήκη μηχανικής μάθησης Pytorch η οποία έχει σχεδιαστεί για την εκτέλεση deep learning και μηχανικής μάθησης μέσω νευρωνικών δικτύων συνεπώς είναι απαραίτητη για την υλοποίηση του **ColBERT**.

**sys**: χρησιμοποιείται η εντολή **sys.path.insert(0, 'ColBERT/')** για να προσθέσει το μονοπάτι του **ColBERT** στο περιβάλλον της *Python* που χρησιμοποιούμε.

**faiss**: είναι ανοικτού κώδικα βιβλιοθήκη και έχει σχεδιαστεί για αποτελεσματική αναζήτηση ομοιότητας και ομαδοποίηση συνόλων δεδομένων σε χώρους υψηλών διαστάσεων που και αυτό είναι απαραίτητο για το **ColBERT**.

**transformers**: είναι ανοικτού κώδικα βιβλιοθήκη, η οποία έχει έτοιμα προεκπαιδευμένα μοντέλα και εργαλεία τα οποία τα χρειαζόμαστε στην επεξεργασία φυσικής γλώσσα (NLP), η οποία πραγματοποιείται στο **ColBERT**.

**cuda**: χρησιμοποιείται για την επιτάχυνση της GPU.

**colbert**: χρησιμοποιείται ώστε να πάρουμε όλες τις απαραίτητες βιβλιοθήκες για την σωστή εκτέλεση του μοντέλου. Θα χρησιμοποιήσουμε τις βιβλιοθήκες Indexer, Searcher, Run, RunConfig, ColBERTConfig, Queries και Collection.

Επίσης, συνδέουμε το *Google Collab* με το drive μας, ώστε να έχει πρόσβαση στα απαραίτητα αρχεία και φακέλους της συλλογής μας.

ΥΛΟΠΟΙΗΣΗ

**2.1 Ερώτημα 1**

Αρχικά ορίζουμε - αρχικοποιούμε 3 μεταβλητές που θα χρησιμοποιηθούν στην συνέχεια:

1. **inverted\_index**: είναι ένα defaultdict, το οποίο αρχικοποιείται ως κενή λίστα. Το deafultdict είναι μια υποκλάση της κλάσης dictionary και επιστρέφει ένα αντικείμενο που μοιάζει με λεξικό. Χρησιμοποιείται ώστε να δημιουργήσει το ανεστραμμένο ευρετήριο, όπου κάθε λέξη θα συσχετίζεται με μια λίστα πλειάδων που θα περιέχουν τα id των κειμένων και τον αριθμό εμφανίσεων της λέξης στο αντίστοιχο κείμενο.
2. **document\_count:** είναι dictionary και θα χρησιμοποιηθεί για να αποθηκεύσουμε τα id των κειμένων που εμφανίζεται κάθε λέξη, για να υπολογίζεται το πλήθος των κειμένων σε επόμενα ερωτήματα.
3. **path:** μεταβλητή η οποία περιέχει την διαδρομή που βρίσκεται ο φάκελος με την συλλογή των κειμένων (docs).

# Question 1  
# defaultdict to store the inverted index  
inverted\_index = defaultdict(list)  
  
# Dictionary to store the documents each word appears in  
document\_count = {}  
  
path = (r"C:\Users\me\PycharmProjects\pythonProject1\docs")  
os.chdir(path)

Στην συνέχεια ξεκινάει η διαδικασία δημιουργίας του Ανεστραμμένου Ευρετηρίου. Η διαδικασία αρχίζει με μια επανάληψη που διατρέχει κάθε αρχείο μέσα στο φάκελο docs με την χρήση της εντολής **os.listdir(path)** και αποθηκεύουμε στην μεταβλητή **file\_path** την διαδρομή για το συγκεκριμένο αρχείο (**os.path.join(path, file)**). Ανοίγουμε, διαβάζουμε το αρχείο και στη μεταβλητή **dictionary** αποθηκεύονται οι λέξεις που περιέχει το συγκεκριμένο κείμενο. Επίσης, δημιουργούμε ένα ακόμη λεξικό (**count**) που θα χρησιμοποιηθεί για να αποθηκεύουμε την συχνότητα εμφάνισης κάθε λέξης στο τρέχων κείμενο. Ύστερα, διατρέχουμε κάθε λέξη του dictionary και αν βρεθεί ανανεώνουμε το **count** για την συγκεκριμένη λέξη. Επιπλέον, για κάθε λέξη που υπάρχει στο τρέχων κείμενο εισάγουμε στο **inverted\_index** το id (είναι το όνομα του αρχείου) και την συχνότητα εμφάνισης (**count**) αν η λέξη υπάρχει ήδη, αλλιώς εισάγεται η λέξη μαζί με τις αντίστοιχες πληροφορίες. Τέλος, στον ίδιο βρόγχο επανάληψης ανανεώνουμε το **document\_count** για κάθε λέξη του κειμένου αν υπάρχει στο dictionary προσθέτουμε το id του, αλλιώς προσθέτουμε όλη την λέξη μαζί με το id.

for file in os.listdir(path):  
 file\_path = os.path.join(path, file)  
 with open(file\_path, 'r') as folder:  
 text = folder.read()  
  
 dictionary = text.split()  
  
 # Create a dictionary to store word counts for each document  
 count = {}  
  
 for word in dictionary:  
 count[word] = count.get(word, 0) + 1  
  
 # Update the inverted index with word counts for the current document  
 for word, count in count.items():  
 inverted\_index[word].append((file, count))  
  
 # Update the document count for the current word  
 if word in document\_count:  
 document\_count[word].add(file)  
 else:  
 document\_count[word] = {file}

Τέλος εκτυπώνουμε το τελικό **inverted\_index** και το αποθηκεύουμε σε ένα αρχείο CSV.

inverted\_index\_csv = r"C:\Users\me\PycharmProjects\pythonProject1\inverted\_index.csv"  
  
# Print the inverted index  
for word, documents in inverted\_index.items():  
 print(f"{word}: {documents}")  
  
# Save the inverted index into a CSV file  
with open(inverted\_index\_csv, 'w', newline='', encoding='utf-8') as csvfile:  
 csv\_writer = csv.writer(csvfile)  
 csv\_writer.writerow(['Word', 'Documents'])  
  
 for word, documents in inverted\_index.items():  
 csv\_writer.writerow([word, documents])

**2.2 Ερώτημα 2**

Ξεκινάμε με τον υπολογισμό του παράγοντα IDF. Αρχικά, για κάθε λέξη **word** που έχουμε στο ευρετήριό μας διατρέχουμε το **document\_count**, το οποίο είναι λεξικό με κλειδί τη συγκεκριμένη λέξη και περιεχόμενο τη λίστα των κειμένων στο οποίο εμφανίζεται αυτή η λέξη. Αποθηκεύουμε μετά τη σάρωση στη λίστα **documents** τα κείμενα όπου εμφανίζεται η κάθε λέξη και υπολογίζουμε το **idf** ως τον δεκαδικό λογάριθμο του μήκους του φακέλου που έχουμε όλα τα έγγραφα της συλλογής διά το μήκος της λίστας **documents**. Έτσι καταφέρνουμε να βρούμε τον λόγο από όλα τα κείμενα της συλλογής προς τα κείμενα που περιλαμβάνουν τη κάθε λέξη. Το **idf** είναι επίσης λεξικό με κλειδί το συγκεκριμένο term και περιεχόμενο κλειδιού τη τιμή του IDF του κλειδιού. Αφού ολοκληρώσουμε την επεξεργασία και γεμίσουμε το λεξικό με τιμές, το μετατρέπουμε σε dataframe και το αποθηκεύουμε σε ένα αρχείο CSV για να μπορούμε να το επεξεργαστούμε πιο εύκολα.

# Question 2  
# Calculate IDF for each word and store it in a dictionary  
idf = {}  
for word, documents in document\_count.items():  
 idf\_value = math.log10((len([f for f in os.listdir(path)])) / len(documents))  
 idf[word] = idf\_value  
  
# Create a DataFrame with IDF values  
idf = pd.DataFrame.from\_dict(idf, orient='index', columns=['IDF'])  
  
csv\_file\_path = r"C:\Users\me\PycharmProjects\pythonProject1\idf.csv"  
# Save the DataFrame to a CSV file  
idf.to\_csv(csv\_file\_path, index=False)

Με όμοιο τρόπο ακριβώς κατασκευάζουμε και το dataframe **tf** με τη μόνη διαφορά ότι τώρα ορίζουμε ένα μετρητή **count**, ο οποίος μεταβαίνει στο Ανεστραμμένο Ευρετήριο και εντοπίζει για κάθε κείμενο **document** το πλήθος των εμφανίσεων κάθε λέξης του **word** που είναι αποθηκευμένο στη λίστα **occurrences**. Έτσι, υπολογίζουμε το TF κάθε term κάθε κειμένου ξεχωριστά ως το δεκαδικό λογάριθμο της τιμής **count** συν 1. Στο τελικό dataframe που προκύπτει συμπληρώνουμε μηδενικά για τα terms τα οποία δεν υπάρχουν καθόλου μέσω της συνάρτησης **fillna( )**.

# Create a dictionary to store TF values for each word in each document  
tf = {}  
for word, occurrences in inverted\_index.items():  
 for document, count in occurrences:  
 if document in tf:  
 tf[document][word] = 1 + math.log10(count)  
 else:  
 tf[document] = {word: 1 + math.log10(count)}  
  
# Create a DataFrame where rows are document titles and columns are words with their TF values  
tf = pd.DataFrame.from\_dict(tf, orient='index')  
tf.fillna(0, inplace=True) # Fill missing values with 0  
  
csv\_file\_path = r"C:\Users\me\PycharmProjects\pythonProject1\doc\_tf.csv"  
# Save the DataFrame to a CSV file  
tf.to\_csv(csv\_file\_path, index=False)

Για τον υπολογισμό του βάρους tf-idf πολλαπλασιάζουμε στη συνέχεια κάθε στήλη (term) του dataframe **tf** με την αντίστοιχη τιμή του dataframe **idf** και το τελικό αποτέλεσμα το αποθηκεύουμε στο dataframe **tfidf**, το οποίο επίσης αποθηκεύουμε σε ένα CSV αρχείο για διευκόλυνση.

# Iterate through the TF and IDF DataFrames to calculate TF-IDF  
tfidf = tf.copy()  
# Multiply each TF value by the corresponding IDF value  
for column in tfidf.columns:  
 tfidf[column] = tfidf[column] \* idf.loc[column, 'IDF']  
  
csv\_file\_path = r"C:\Users\me\PycharmProjects\pythonProject\doc\_weights.csv"  
# Save the DataFrame to a CSV file  
tfidf.to\_csv(csv\_file\_path, index=False)

Στη συνέχεια, θέλουμε να υπολογίσουμε τα αντίστοιχα βάρη για κάθε ερώτημα από τη συλλογή ερωτημάτων χρήστη, οπότε πάλι θα αποθηκεύσουμε την απόλυτη διαδρομή του αρχείου με τα ερωτήματα στη μεταβλητή **query\_path**, θα διαβάσουμε κάθε ερώτημα που αντιστοιχεί σε μία γραμμή του αρχείου μέσω της **readlines( )** και θα το αποθηκεύσουμε στη λίστα **questions**. Έπειτα, κατασκευάζουμε μια λίστα **all\_words** με όλες τις λέξεις του Ανεστραμμένου Ευρετηρίου ως περιεχόμενο και ένα κενό dataframe **df**, όπου θα αποθηκεύουμε τους παράγοντες TF κάθε ερωτήματος. Αρχικοποιούμε πάλι με 0 το μετρητή **count** όπως και πριν και ξεκινάμε να διατρέχουμε τη λίστα με τα ερωτήματα. Έτσι, για κάθε ερώτημα **question** της λίστας, το μετατρέπουμε αρχικά με την **upper( )** σε κεφαλαία γράμματα για να ταιριάζει με το case sensitivity των terms του Ανεστραμμένου Ευρετηρίου και αποθηκεύουμε όλες τις διακριτές του λέξεις μέσω της **split( )** στη λίστα **words**. Στη συνέχεια, για κάθε ερώτημα βρίσκουμε τις κοινές λέξεις του Ανεστραμμένου Ευρετηρίου και του ερωτήματος παίρνοντας τη τομή των δύο συνόλων **words** και **all\_words** και τις αποθηκεύουμε στη λίστα **common\_words**. Ορίζουμε και μία λίστα **word\_counts** που διατηρεί τις τιμές του μετρητή για κάθε κοινή λέξη, την οποία αρχικοποιούμε με 0. Τέλος, για κάθε κοινή λέξη ψάχνουμε στο αλφαριθμητικό **question** πόσες φορές εμφανίζεται μέσω της **findall( )**, η οποία ωστόσο μας επιστρέφει τη λέξη όσες φορές αυτή εμφανίζεται στην πρόταση, γι’ αυτό και παίρνουμε το μήκος αυτής της λίστας, μέσω της **len( ),** ως τιμή για το μετρητή **count**. Το TF κάθε λέξης του ερωτήματος δίνεται πάλι ως ο δεκαδικός λογάριθμος της τιμής count συν 1 και αποθηκεύεται στο αρχικό dataframe **df**.

query\_path = r"C:\Users\me\PycharmProjects\pythonProject1\Queries\_20"  
# Read questions from the file  
with open(query\_path, 'r') as file:  
 questions = file.readlines()  
  
# Create a list of words from the inverted index  
all\_words = list(inverted\_index.keys())  
  
# Create a DataFrame with zeros  
df = pd.DataFrame(0, index=range(len(questions)), columns=all\_words)  
  
# Iterate through questions and update the DataFrame  
count = 0  
for i, question in enumerate(questions):  
 words = question.upper().split()  
 # Find common words between the question and the inverted index  
 common\_words = set(words) & set(inverted\_index)  
 word\_counts = {word: 0 for word in common\_words}  
  
 for word in common\_words:  
 # Count the number of occurrences of the word in the question  
 pattern = r'\b' + re.escape(word) + r'\b'  
 count = len(re.findall(pattern, question.upper()))  
 # Update the DataFrame with the log-transformed count  
 if count > 0:  
 df.at[i, word] = 1 + math.log10(count)  
  
df = df.fillna(0)  
csv\_file\_path = r"C:\Users\me\PycharmProjects\pythonProject1\query\_tf.csv"  
# Save the DataFrame to a CSV file  
df.to\_csv(csv\_file\_path, index=False)

Για τον υπολογισμό του βάρους tf-idf των ερωτημάτων όμοια με πριν πολλαπλασιάζουμε στη συνέχεια κάθε στήλη (term) του dataframe **df** με την αντίστοιχη τιμή του dataframe **idf** και το τελικό αποτέλεσμα το αποθηκεύουμε στο dataframe **query\_tfidf**, το οποίο επίσης αποθηκεύουμε σε ένα CSV αρχείο.

query\_tfidf = df.multiply(idf['IDF'], axis=1)  
csv\_file\_path = r"C:\Users\me\PycharmProjects\pythonProject1\query\_weights.csv"  
# Save the DataFrame to a CSV file  
query\_tfidf.to\_csv(csv\_file\_path, index=False)

Το τελικό μας βήμα αποτελεί ο υπολογισμός των Ευκλείδειων νορμών κάθε γραμμής των **tfidf** και **query\_tfidf**, το οποίο πραγματοποιείται μέσω της συνάρτησης **linalg.norm( )** της βιβλιοθήκης numpy. Τα τελικά αποτελέσματα για τις νόρμες κειμένων και ερωτημάτων τα αποθηκεύουμε στα αντίστοιχα dataframes **euclidean\_norms\_d** και **euclidean\_norms\_q**. Το εσωτερικό γινόμενο υπολογίζεται μέσω της συνάρτησης **dot( )** με ορίσματα τις τιμές των αντίστοιχων γραμμών των dataframes **tfidf** και **query\_tfidf.Τ**. Ο λόγος για τον οποίο εφαρμόζουμε τον τελεστή αναστροφής (Τ) στο διάνυσμα των βαρών των ερωτημάτων είναι για να υπολογιστεί σωστά το εσωτερικό γινόμενο, καθώς όπως γνωρίζουμε και από τη Γραμμική Άλγεβρα ισχύει ότι: <**u**, **w**> := **uw**T. Τα τελικά αποτελέσματα αποθηκεύονται και αυτά σε ένα dataframe **dot\_product**, ενώ τα γινόμενα των νορμών **euclidean\_norms\_d** και **euclidean\_norms\_q** αποθηκεύονται στο datafame **norm\_product**, όπου πολλαπλασιάζουμε κάθε γραμμή του πρώτου με κάθε στήλη του δεύτερου. Προφανώς, οι διαστάσεις και για τα δύο τελευταία dataframes που αναφέραμε είναι 1209x20, γεγονός που επιβεβαιώνει ότι έχουμε πράξει ορθά, καθώς κάθε κελί των δύο πινάκων αντιπροσωπεύει ένα συγκεκριμένο κείμενο ως προς ένα συγκεκριμένο ερώτημα. Για να υπολογίσουμε λοιπόν το dataframe **cosine\_similarity** απλώς διαιρούμε το **dot\_product** με το **norm\_product** μέσω της **div( )** και το τελικό dataframe το οποίο προκύπτει θα είναι και το ζητούμενο.

# Calculate the Euclidean norm for each row of doc\_weights  
euclidean\_norms = np.linalg.norm(tfidf.values, axis=1)  
# Create a DataFrame with the Euclidean norm values  
euclidean\_norms\_d = pd.DataFrame({'Euclidean Norm': euclidean\_norms}, index=tfidf.index)  
  
# Calculate the Euclidean norm for each row of query\_weights  
euclidean\_norms = np.linalg.norm(query\_tfidf.values, axis=1)  
# Create a DataFrame with the Euclidean norm values  
euclidean\_norms\_q = pd.DataFrame({'Euclidean Norm': euclidean\_norms}, index=query\_tfidf.index)  
  
# Calculate the dot product  
dot\_products = np.dot(tfidf.values, query\_tfidf.values.T)  
# Create a DataFrame with the dot product values  
dot\_product = pd.DataFrame(dot\_products, index=tfidf.index, columns=query\_tfidf.index)  
  
# Calculate the product of the Euclidean norms  
norm\_product = euclidean\_norms\_d.values @ euclidean\_norms\_q.values.T  
# Create a DataFrame with the result  
norm\_product = pd.DataFrame(norm\_product, index=euclidean\_norms\_d.index, columns=euclidean\_norms\_q.index)  
  
# Calculate the cosine similarity  
cosine\_similarity = dot\_product.div(norm\_product)  
print(cosine\_similarity)

**2.3 Ερώτημα 3**

Αρχικά, όλα τα κείμενα και τα απαραίτητα αρχεία είναι αποθηκευμένα στο φάκελο docs μέσα στο drive και είναι σε συμπιεσμένα σε .zip μορφή, οπότε τον αποσυμπιέζουμε και τα περνάμε στο μονοπάτι **“/content/docs”**.

Δημιουργείται έτσι η συνάρτηση **read\_documents\_from\_directory( ),** η οποία διαβάζει κάθε αρχείο που βρίσκεται μέσα στο φάκελο docs με τον ίδιο τρόπο όπως και στο **Vector Space Model** και αναπαριστά κάθε αρχείο ως λίστα με id το όνομα του αρχείου και το περιεχόμενο του εγγράφου.

Η συνάρτηση **read\_queries\_from\_file( ),** διαβάζει το αρχείο που περιέχει τα ερωτήματα και τα αποθηκεύει σε λίστα.

Στην συνέχεια ορίζουμε τα configurations του **ColBERT** ως εξής:

* **nbits**: κωδικοποιεί κάθε διάσταση με 2 bit.
* **max\_id** : Ορίζουμε το μέγιστο id των αρχείων που είναι 1209 και συμβολίζει ότι στα 1209 id θα χρησιμοποιηθεί ο indexer.
* **index\_name** : το όνομα του δείκτη.
* **checkpoint**: περιέχει το μονοπάτι που βρίσκεται το checkpoint.

Όσον αφορά το indexing, γίνεται με βάση το προεκπαιδευμένο μοντέλο και με βάση τα configurations που ορίσαμε παραπάνω. Δημιουργεί έναν δείκτη και στην συνέχεια δεικτοδοτεί τα 1209 έγγραφα.

Στην συνέχεια δημιουργούμε τοv Searcher και για κάθε query εκτυπώνουμε τα ανακτημένα κείμενα αντιστοιχίζοντας το passage\_id με το σωστό document\_id. Επίσης αποθηκεύομε σε ένα αρχείο csv τα query και για κάθε query τα ανακτημένα κείμενα, ώστε να μπορούμε να τα επεξεργαστούμε στο ερώτημα 4.

**2.4 Ερώτημα 4**

Για το συγκεκριμένο ερώτημα αρχικά θα ταξινομήσουμε σε φθίνουσα σειρά τα αποτελέσματα του **Vector Space Model** με βάση το cosine similarity, τα αποθηκεύουμε στο dataframe **sorted\_doc\_ids** και τα εξάγουμε και σto CSV “**cosine\_similarity.csv”**. Στην συνέχεια, αντικαθιστούμε τις τιμές του dataframe **cosine similarity** με το id κάθε κειμένου στο οποίο αντιστοιχούν, το οποίο το μετατρέπουμε σε ακέραιο για κάθε ερώτημα, ώστε τα πιο σχετικά id να βρίσκονται στην κορυφή της λίστας και τα λιγότερο σχετικά στο τέλος. Τα αποτελέσματα τα αποθηκεύουμε σε μια λίστα **retrieved\_docs** στην οποία κλειδί είναι το id του ερωτήματος και το περιεχόμενο είναι λίστα που περιέχει τα id των κειμένων.

# Question 4  
# Precision - Recall and MAP Metrics  
# Sort each column and replace values with doc IDs  
sorted\_doc\_ids = cosine\_similarity.apply(lambda col: col.sort\_values(ascending=False).index)  
csv\_file\_path = r"C:\Users\me\PycharmProjects\pythonProject1\cosine\_similarity.csv"  
# Save the DataFrame to a CSV file  
sorted\_doc\_ids.to\_csv(csv\_file\_path, index=False)  
# Convert document IDs to integers and get the retrieved docs  
retrieved\_docs = {query\_id: [int(doc\_id) for doc\_id in doc\_ids] for query\_id, doc\_ids in sorted\_doc\_ids.items()}

Σχετικά με τα αποτελέσματα του **ColBERT** τα έχουμε μεταβιβάσει από το περιβάλλον του *Google Colab* και αποθηκεύσει σε CSV, του οποίου η πρώτη στήλη περιέχει τα ερωτήματα και η δεύτερη τα id των κειμένων που είναι ήδη τοποθετημένα σε φθίνουσα σειρά. Οπότε διατρέχουμε κάθε γραμμή του CSV και ελέγχουμε αν στη δεύτερη στήλη τα id των κειμένων είναι ακέραιοι και σε λίστα. Αν είναι τότε τα αποθηκεύουμε στο λεξικό **query\_docs** όπου το κλειδί είναι το id των ερωτημάτων και τα περιεχόμενα είναι λίστα με όλα τα κείμενα.

colbert\_path = r"C:\Users\me\PycharmProjects\pythonProject1\colbert\_result.csv"  
query\_docs = {}  
# Read the ColBERT results into a DataFrame  
with open(colbert\_path, 'r') as csvfile:  
 csvreader = csv.reader(csvfile)  
 next(csvreader) # Skip the header row  
  
 for query\_id, row in enumerate(csvreader):  
 # Parse the string representation of the list into an actual list  
 doc\_ids = ast.literal\_eval(row[1])  
 # Ensure doc\_ids is a list and contains integers  
 if isinstance(doc\_ids, list) and all(isinstance(id, int) for id in doc\_ids):  
 query\_docs[query\_id] = doc\_ids

Αναφορικά με το αρχείο CSV **“Relevant\_20”**, κάθε γραμμή του περιέχει τα σχετικά κείμενα για κάθε ερώτημα, οπότε διατρέχουμε το αρχείο, μετατρέπουμε τα id των κειμένων σε ακεραίους και τα αποθηκεύουμε στην λίστα **relevant\_docs**, όπου το κλειδί είναι το id των ερωτημάτων και τα περιεχόμενα είναι λίστα με όλα τα κείμενα.

# Get the relevant docs  
file\_path = r"C:\Users\me\PycharmProjects\pythonProject1\Relevant\_20"  
relevant\_docs = {}  
with open(file\_path, 'r') as file:  
 for query\_id, line in enumerate(file):  
 # Convert all space-separated numbers in the line to integers and store them as a set  
 doc\_ids = set(map(int, line.split()))  
 # Assign this set to the corresponding query ID  
 relevant\_docs[query\_id] = doc\_ids

Τώρα θα υλοποιήσουμε τις μετρικές precision - recall, preciosion@k και Mean Average Length (MAP). Οι τιμές των precision, recall και η MAP θα υπολογιστούν στην ίδια συνάρτηση **calculate\_metrics( )**, ενώ η τιμή precision@k υπολογίζεται από την συνάρτηση **precision\_at\_k( )**.

Όσον αφορά την συνάρτηση **calculate\_metrics( )**, δέχεται σαν όρισμα την λίστα με τα σχετικά κείμενα και την λίστα με τα ανακτημένα. Αρχικά ορίζονται δυο μεταβλητές που θα χρησιμοποιηθούν σαν μετρητές, η **ret\_count** που είναι μετρητής για τα ανακτημένα κείμενα και η **rel\_count**, η οποία είναι μετρητής για τα σχετικά. Επίσης ορίζονται οι λίστες **precisions** και **recalls**. Ύστερα, διατρέχουμε με μια επανάληψη όλα τα κείμενα που βρίσκονται μέσα στη λίστα **retrieved\_docs** και κάθε φορά που μεταβαίνουμε σε καινούριο κείμενο αυξάνεται ο μετρητή **ret\_count** κατά 1. Ελέγχουμε αν το κείμενο που εξετάζουμε αυτή τη στιγμή βρίσκεται και στην λίστα **relevant\_docs** και αν βρίσκεται, τότε σημαίνει ότι είναι σχετικό κείμενο, οπότε αυξάνουμε και τον μετρητή **rel\_count** κατά 1. Υπολογίζουμε τις τιμές precision και το recall για το συγκεκριμένο κείμενο εφαρμόζοντας τους τύπους που έχουν αναφερθεί στην υποενότητα **1.4** και τα αποθηκεύουμε στις λίστες **precisions** και **recalls** αντίστοιχα, οι οποίες επιστρέφονται όταν καλούμε την συνάρτηση. Για τον υπολογισμό της MAP θα υπολογίσουμε έξω από τον βρόχο επανάληψης τον μέσο όρο των precisions (**avg\_precision**) και το αποτέλεσμα επιστρέφεται επίσης μαζί με τα υπόλοιπα.

def calculate\_metrics(relevant\_docs, retrieved\_docs):  
 ret\_count = 0  
 rel\_count = 0  
 precisions = []  
 recalls = []  
  
 # Iterate through each relevant document  
 for doc in retrieved\_docs:  
 ret\_count += 1  
 # Check if the relevant document is in the retrieved documents  
 if doc in relevant\_docs:  
 rel\_count += 1  
 precision = rel\_count / ret\_count  
 precisions.append(precision)  
 recall = rel\_count / len(relevant\_docs)  
 recalls.append(recall)  
 avg\_precision = sum(precisions) / len(relevant\_docs)  
  
 return precisions, recalls, avg\_precision

Για να μπορούμε να συγκρίνουμε την απόδοση των δύο μοντέλων με βάση το γράφημα θα δημιουργήσουμε μια συνάρτηση **calculate\_area( )**, η οποία θα υπολογίζει το εμβαδόν που περικλείεται ανάμεσα στον άξονα x και κάθε γραφική για όλα τα ερωτήματα. Ο υπολογισμός του εμβαδού γίνεται με βάση τον ορισμό του Riemann, όπου προσεγγίζουμε το εμβαδό αθροίζοντας τα εμβαδά μικρότερων ορθογωνίων παραλληλογράμμων με πλάτος την απόσταση δύο διαδοχικών τιμών της λίστας **recalls** (x – axis) και ύψος το ημιάθροισμα δύο διαδοχικών τιμών της λίστας **precisions** (y – axis). Το πλήθος των ορθογωνίων που θα χρησιμοποιήσουμε θα είναι ίσο με το πλήθος των τιμών της λίστας recalls, δηλαδή για κάθε ερώτημα χωρίζουμε τον άξονα των τετμημένων με βάση το πλήθος των τιμών ανάκλησης σε αντίστοιχα υποδιαστήματα. Τέλος, το εμβαδό **area** που θα προκύπτει κάθε φορά θα είναι το άθροισμα των εμβαδών των μέχρι τώρα ορθογωνίων που έχουμε υπολογίσει με το εμβαδόν του ορθογωνίου που υπολογίζουμε τη συγκεκριμένη χρονική στιγμή. Η συνάρτηση θα δέχεται σαν όρισμα την λίστα **recalls**, τη λίστα **precisions** που θα έχει δημιουργηθεί από την προηγούμενη συνάρτηση και θα επιστρέφει τη τιμή της μεταβλητής **area**.

def calculate\_area(recalls, precisions):  
 area = 0.0  
 for i in range(1, len(recalls)):  
 # Calculate the area of the trapezoid  
 width = recalls[i] - recalls[i-1]  
 height = (precisions[i] + precisions[i-1]) / 2  
 area += width \* height  
 return area

Για την μετρική precision@k, όπως έχουμε αναφέρει ήδη δημιουργούμε την συνάρτηση **precision\_at\_k( )**, οποία παίρνει σαν ορίσματα την λίστα με τα σχετικά κείμενα, την λίστα με τα ανακτημένα και το όριο **k**. Ελέγχει άμα η λίστα των ανακτηθέντων κειμένων είναι μεγαλύτερη του ορίου **k** και άμα είναι, τότε η λίστα «κόβεται» και πλέον περιέχει μόνο τα πρώτα **k** κείμενα. Στην συνέχεια, διατρέχουμε την καινούρια λίστα και για κάθε κείμενο ελέγχουμε αν είναι σχετικό και επιστρέφουμε τη τιμή 1 αν είναι. Υπολογίζουμε το άθροισμα των σχετικών κειμένων **relevant\_count** και η συνάρτηση επιστρέφει το αποτέλεσμα της διαίρεσης του αθροίσματος προς το όριο **k**.

def precision\_at\_k(relevant\_docs, retrieved\_docs, k):  
 if len(retrieved\_docs) > k:  
 retrieved\_docs = retrieved\_docs[:k]  
 relevant\_count = sum([1 for doc in retrieved\_docs if doc in relevant\_docs])  
 return relevant\_count / k

Αφού τελειώσαμε με την ανάλυση των συναρτήσεων τώρα θα εξηγήσουμε τον τρόπο εφαρμογής τους. Αρχικά, ορίζουμε 4 μεταβλητές που μας είναι απαραίτητες:

* **queries** : το πλήθος των ερωτημάτων.
* **top\_k**: το όριο k που χρειάζεται για την μετρική precision@k.
* **avg\_precisions**: λίστα που θα περιέχει τους μέσους όρους των precisions για κάθε ερώτημα για το μοντέλο Vector Space.
* **avg\_precisions\_colbert**: λίστα που θα περιέχει τους μέσους όρους των precisions για κάθε ερώτημα για το μοντέλο ColBERT.
* **areas\_cosine\_similarity**: λίστα που θα περιέχει τa εμβαδά για κάθε ερώτημα για το μοντέλο Vector Space.
* **areas\_colbert**: λίστα που θα περιέχει τa εμβαδά για κάθε ερώτημα για το μοντέλο ColBERT.

Στην συνέχεια, διατρέχουμε κάθε γραμμή (ερώτημα) της λίστας **relevant\_docs** και για κάθε γραμμή αποθηκεύουμε στη λίστα **relevant\_docs\_set** τα σχετικά κείμενα που αντιστοιχούν στο συγκεκριμένο **query\_id**. Επίσης, με βάση το id των ερωτημάτων βρίσκουμε και αποθηκεύουμε σε λίστες τα ανακτηθέντα κείμενα για το Vector Space Model (**retrieved\_docs\_list**) και για ColBERT (**retrieved\_docs\_list\_colbert**). Για κάθε μοντέλο ξεχωριστά καλούμε την συνάρτηση **calculate\_metrics( )** και την τρίτη παράμετρο που μας επιστρέφει, που είναι ο μέσος όρος των precisions, την προσθέτουμε στην λίστα **avg\_precisions** ή **avg\_precisions\_colbert** ανάλογα με το μοντέλο. Επιπλέον, υπολογίζουμε την τιμή του MAP που θα είναι το άθροισμα των παραπάνω λιστών προς το πλήθος των ερωτημάτων. Ύστερα, καλούμε και την συνάρτηση **precision\_at\_k( )** και για τα δύο μοντέλα και εκτυπώνουμε τα αποτελέσματά της για κάθε ερώτημα. Τις δύο τιμές που μας επιστρέφει η **calculate\_metrics( )**, που είναι το recall και το precision, θα τις αποτυπώσουμε σε ένα γράφημα, όπου ο άξονας x θα είναι το recall και ο y το precision. Τέλος, υπολογίζουμε τα εμβαδά για κάθε γραφική καλώντας την συνάρτηση **calculate\_area( )**, τα οποία αποθηκεύουμε σε αντίστοιχες λίστες και τα εκτυπώνουμε.

queries = 20  
top\_k = 400  
avg\_precisions = []  
avg\_precisions\_colbert = []  
areas\_cosine\_similarity = []  
areas\_colbert = []  
  
for query\_id, relevant\_docs\_set in relevant\_docs.items():  
 retrieved\_docs\_list = retrieved\_docs.get(query\_id, [])  
 retrieved\_docs\_list\_colbert = query\_docs.get(query\_id, [])  
 precisions, recalls, avg\_precision = calculate\_metrics(relevant\_docs\_set, retrieved\_docs\_list)  
 avg\_precisions.append(avg\_precision)  
 map\_values = sum(avg\_precisions) / queries  
  
 precisions\_colbert, recalls\_colbert, avg\_precision\_colbert = calculate\_metrics(relevant\_docs\_set, retrieved\_docs\_list\_colbert)  
 avg\_precisions\_colbert.append(avg\_precision\_colbert)  
 map\_values\_colbert = sum(avg\_precisions\_colbert) / queries  
  
 # Calculate precision@k for Cosine Similarity and ColBERT  
 precision\_at\_k\_cosine = precision\_at\_k(relevant\_docs\_set, retrieved\_docs\_list, top\_k)  
 precision\_at\_k\_colbert = precision\_at\_k(relevant\_docs\_set, retrieved\_docs\_list\_colbert, top\_k)  
  
 # Print or store these values for comparison  
 print(f"Precision@{top\_k} for Cosine Similarity (Query {query\_id + 1}): {precision\_at\_k\_cosine}")  
 print(f"Precision@{top\_k} for ColBERT (Query {query\_id + 1}): {precision\_at\_k\_colbert}")  
  
 plt.figure()  
 plt.plot(recalls, precisions, marker='o', label = 'Cosine Similarity')  
 plt.plot(recalls\_colbert, precisions\_colbert, marker='o', label = 'ColBERT')  
 plt.title(f"Precision - Recall for Query {query\_id + 1}")  
 plt.xlabel("Recall")  
 plt.ylabel("Precision")  
 plt.grid(True)  
 plt.legend()  
 plt.show()  
  
 # After calculating precisions and recalls for each query  
 area\_cosine\_similarity = calculate\_area(recalls, precisions)  
 areas\_cosine\_similarity.append(area\_cosine\_similarity)  
 area\_colbert = calculate\_area(recalls\_colbert, precisions\_colbert)  
 areas\_colbert.append(area\_colbert)  
  
 # Now you can print or plot the AUC values  
 print(f"Area for Cosine Similarity (Query {query\_id + 1}): {area\_cosine\_similarity}")  
 print(f"Area for ColBERT (Query {query\_id + 1}): {area\_colbert}")

Μετά το τέλος του βρόγχου επανάληψης και αφού έχουμε επεξεργαστεί όλα τα ερωτήματα τυπώνουμε το τελικό αποτέλεσμα και της MAP και για τα δύο μοντέλα και στη συνέχεια υπολογίζουμε το μέσο όρο των εμβαδών κάθε μοντέλου και εκτυπώνουμε και αυτό το αποτέλεσμα.

# Calculate the mean area to compare the general efficiency of both models  
mean\_area\_cosine\_similarity = sum(areas\_cosine\_similarity) / queries  
mean\_area\_colbert = sum(areas\_colbert) / queries  
  
print(f"Mean Area under Precision-Recall Curve for Cosine Similarity: {mean\_area\_cosine\_similarity}")  
print(f"Mean Area under Precision-Recall Curve for ColBERT: {mean\_area\_colbert}")

ΑΠΟΤΕΛΕΣΜΑΤΑ – ΠΑΡΑΤΗΡΗΣΕΙΣ

**3.1 Ερώτημα 1**

Η συλλογή πάνω στην οποία εργαζόμαστε είναι η Cystic Fibosis. Το Ανεστραμμένο Ευρετήριο

ΑΝΑΦΟΡΕΣ