**Συστήματα Ανάκτησης Πληροφοριών | 8ο Εξάμηνο Πληροφορική ΟΠΑ**

**Γεώργιος Βίκτωρ Κέκας (3210074)– Θανάση Πρίφτι(3210171)**

**1η Φάση Εργασίας**

A screen shot of a computer code

AI-generated content may be incorrect.Όσον αφορά το reprocessing των δεδομένων χρησιμοποιήσαμε την βιβλιοθήκη pandas.

Παρακάτω φαίνεται ο κώδικας με τον οποίο φορτώνουμε τα δεδομένα μας (εικόνα 1), καθώς και η επεξεργασία που γίνεται (εικόνα 2).

A computer screen with text and symbols

AI-generated content may be incorrect.Κατά την προ επεξεργασία των δεδομένων, εφαρμόστηκαν κανονικές εκφράσεις για την ομογενοποίηση του κειμένου. Συγκεκριμένα, μετατρέψαμε όλους τους χαρακτήρες σε πεζά, αφαιρέσαμε ειδικούς χαρακτήρες, αριθμούς και σημεία στίξης. Επιπλέον, εξαλείψαμε τα πολλαπλά διαδοχικά κενά, τα οποία αντικαταστάθηκαν με ένα μόνο κενό.

Η προ επεξεργασία αυτή γίνεται μέσω της συνάρτησης clear\_text(text), η οποία δέχεται σαν μόνο όρισμα το κείμενο στο οποίο θα γίνει η επεξεργασία. Παράλληλα ‘δημιουργούμε’ μια νέα στήλη την clean\_text στην οποία, όπως λέει και το όνομα της, θα περιέχει το «καθαρισμένο κείμενο».

*Τέλος αξίζει να τονιστεί ότι στην αρχή κάθε κειμένου βάλαμε τον τίτλο, ώστε να μην χάνεται τυχών χρησιμη πληροφορία, δεν βάλαμε τον τίτλο σαν επιπλέον Field.*

A screen shot of a computer program

AI-generated content may be incorrect.Όσον αφορά τώρα το κομμάτι του elastic search, την συνάρτηση ομοιότητας και τον analyzer είναι αυτοί που χρησιμοποιήσαμε και στο εργαστήριο. Όσον αφορά το mapping, αντιστοιχούμε το ‘καθαρισμένο κείμενο’ στο text, χρησιμοποιώντας τα settings από πάνω. Ο κώδικας για τα παραπάνω βρίσκεται παρακάτω:

A screenshot of a computer program

AI-generated content may be incorrect.Παραπάνω βλέπουμε τον έλεγχο για την ύπαρξη, ή μη του ευρετηρίου, καθώς και το bulk add, των δεδομένων μας. Αυτό γίνεται μέσω της συνάρτησης helpers.bulk(client, docs, index= index\_name). Όλα τα ορίσματα έχουν αρχικοποιηθεί παραπάνω.

Με τον παραπάνω κώδικα φορτόνουμε τα queries (πάνω block) και γράφουμε τα αποτελέμσατα στο εκάστοτε αρχέιο.

A computer screen shot of code

AI-generated content may be incorrect.Με το παρακάτω κομμάτι κώδικα τρέχουμε όλα τα ερωτήματα μας για το corpus μας, και αποθηκέυουμε τα αποτελέσματα μας στα αρχεία. Για τις διάφορες τιμές του κ, απλώς αλλάξαμε το base\_path και το size.

Για κάθε τιμή του k (20, 30, 50), υπάρχει ενας φάκελος με 50 αρχεία συνολικά, ένα για κάθε query.

Για την αξιολόγηση των αποτελεσμάτων μας , χρησιμοποιήσαμε το εργαλείο trec\_eval. Πιο συγκεκριμένα φτιάξαμε ένα bat αρχείο το οποίο για κάθε αρχείο στον εκάστοτε φάκελο (result\_folder), μας υπολόγιζε τις απαράιτητες μετρικές. Οι μετρικές αυτές είναι: το mean average precision (map), και το average precision για διάφορες τιμές του k (5, 10, 15,20).

Τα αποτελεσματα για κάθε μία από τις τιμές του k (20,30,50) βρίσκοντε στον παραδοτέο φακελο με το όνομα #number\_scores. A screenshot of a computer program

AI-generated content may be incorrect.

*H αναγκαιότητα δημιουργίας αυτού του αρχείου είναι για να μήν τρέχουμε χειροκίνητα την εντολή trec\_eval… για όλα τα αρχεία.*

Τα αποτελέσματα της αξιολόγησης μας, μπορεί να διαφέρουν από ερώτημα σε ερώτημα, επειδή κάθε ερώτημα έχει τις δικές του δυσκολίες. Η τιμή MAP (Μέση Μέση Ακρίβεια) δείχνει πόσο καλά ανακτά το σύστημα τα σχετικά έγγραφα συνολικά — όσο πιο κοντά στο 1.0, τόσο καλύτερη είναι η απόδοση. Οι μετρικές P@K (Ακρίβεια στις K πρώτες θέσεις), όπως P\_5, P\_10 ή P\_20, δείχνουν πόσα σχετικά αποτελέσματα εμφανίζονται στις πρώτες θέσεις. Όσο υψηλότερες είναι αυτές οι τιμές, τόσο καλύτερη είναι η κατάταξη των σχετικών εγγράφων. Οι διαφορές στα σκορ οφείλονται στο ότι κάποια ερωτήματα είναι πιο δύσκολα, λιγότερο σαφή ή αφορούν διαφορετικά θέματα.

Επιπρόσθετα μια ενδιαφέρουσα παρατήρhση είναι ότι, για το ερώτημα 10, όλες οι τιμές των μετρικών , και για Κ=20,30,50 είναι ίσες με 0 !. Αυτό μπορεί να ερμηνευτεί με διάφορους τρόπους. Ένας από αυτούς είναι ότι το query είναι ‘άσχετο’ με το corpus μας, δηλαδή δεν παρουσιάζει καμία συνάφεια με την αναζήτηση μας.

Μια άλλη παρατήρηση είναι ότι δεν παρατηρείται μεγάλη αλλαγη μεταξύ των αποτελεσμάτων για τα queries.

*Τα αποτελέσματα για τα queries καθώς και ο κώδικας και το report αυτό βρίσκοντε και στο github repository :* [*https://github.com/victorkekas/Info-Retrieval.git*](https://github.com/victorkekas/Info-Retrieval.git)*. Τα αποτελέσμαα βρίσκονται στον φάκελο res. Εκεί βρίσκονται και τα αποτελέσματα από το trec\_eval*.