**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

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 πληροφορία, δεν βάλαμε τον τίτλο σαν επιπλέον Field.*

Όσον αφορά τώρα το κομμάτι του elastic search, ο κώδικας βρίσκεται παρακάτω: A computer screen with text

AI-generated content may be incorrect.

A screen shot of a computer program

AI-generated content may be incorrect.

Η συνάρτηση ομοιότητας και ο analyzer είναι αυτοί που χρησιμοποιήσαμε και στο εργαστήριο. Όσον αφορά το mapping, αντιστοιχούμε το ‘καθαρισμένο κείμενο’ στο text, χρησιμοποιώντας τα settings από πάνω.

A screen shot of a computer screen

AI-generated content may be incorrect.

Παραπάνω βλέπουμε έναν 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. A black background with white text

AI-generated content may be incorrect.

*Αξίζει να τονιστεί ότι για κάθε «καθαρό κείμενο» περνάμε και το εκάστοτε μοναδικό id, σαν αναγνωριστικό.*

Παράλληλα εάν πάμε στο site της elasticsearch, και δούμε το index μας, παρατηρούμε ότι τα δεδομένα έχουν περάσει με τον επιθυμητό τρόπο.

A screenshot of a computer

AI-generated content may be incorrect.

Github repository : <https://github.com/victorkekas/Info-Retrieval.git>