Πανεπιστήμιο Δυτικής Αττικής

Τμήμα Μηχανικών Πληροφορικής και Υπολογιστών

Σωτήριος Σουκαράς, 21390206

Θεοφάνης Κουνιάκης, 21390103

Link του κώδικα:

A blue logo with text

Description automatically generated

Ανακτηση ΠΛηροφοριασ

ΈΚΘΕΣΗ ΑΝΑΦΟΡΑΣ

# Βήμα 1. Συλλογή δεδομένων:

Στο αρχείο **‘wiki\_scraper.py’** πραγματοποιείται συλλογή των δεδομένων από άρθρα του Wikipedia.

Η βασική συνάρτηση του αρχείου είναι η **‘fetch\_wikipedia’** η οποία δέχεται ως όρισμα τα **URL**, **depth\_limit** και **depth. (fetch\_wikipedia(URL, depth\_limit, depth = 1))**

* Η μεταβλητή **URL** αντιστοιχεί στον αρχικό σύνδεσμο του άρθρου που θα εξεταστεί.
* Το **depth\_limit** δηλώνει το μέγιστο βάθος της αναδρομής.
* Η μεταβλητή **depth** είναι το τρέχον επίπεδο της αναδρομής και έχει default τιμή **1**.

Η συνάρτηση αρχικά στέλνει **HTTP** αίτημα από το **URL** και ανακτά to **HTML**. Στην συνέχεια κρατά το κείμενο που περιέχεται στις παραγράφους του και αφαιρεί τα **html tags** τους.

Κάθε παράγραφος τοποθετείται στην λίστα **parsed\_paragraphs**.

Έπειτα γίνεται έλεγχος αν το **depth** είναι μεγαλύτερο ή ίσο του **depth\_limit**. Αν η συνθήκη ισχύει τότε η συνάρτηση επιστρέφει την λίστα **parsed\_paragraphs**.

Στην περίπτωση που η συνθήκη είναι ψευδής τότε γίνεται αναδρομή για κάθε link **(href)** καλώντας την **wiki\_scraper** με όρισμα **depth+1**. Η επιστροφή κάθε αναδρομής αποθηκεύεται στην λίστα **new\_paragraphs** η οποία χρησιμοποιείται για την ανανέωση της λίστας **parsed\_paragraphs**.

Με το τέλος των αναδρομών η συνάρτηση επιστρέφει την λίστα **parsed\_paragraphs**.

Για την συλλογή των δεδομένων η συνάρτηση κλήθηκε με το **URL** της σελίδας της Wikipedia του δευτέρου παγκοσμίου πολέμου με βάθος **depth = 2**. Η έξοδος της συνάρτησης τοποθετήθηκε στην λίστα **results**.

Τέλος η λίστα αυτή αποθηκεύεται στο αρχείο **wiki\_scrape.json** .

# Βήμα 2. Προεπεξεργασία κειμένου (Text Processing):

Στο αρχείο **‘preprocessor.py’** γίνεται η προεπεξεργασία του κειμένου που συλλέχθηκε από το **wiki\_scraper.**

Βασική συνάρτηση του αρχείου είναι η ‘**clear\_special\_char’** και αφαιρεί τα εξής:

* χαρακτήρες αναφοράς του τύπου ‘[1][2]’ με την χρήση regex
* μη αλφαριθμητικούς χαρακτήρες
* τον χαρακτήρα ‘-’

δδδ