

**Πανεπιστήμιο Δυτικής Αττικής**

**Τμήμα Μηχανικών**

**Τμήμα Μηχανικών και Πληροφορικής Υπολογιστών**

**Ανάκτηση Πληροφορίας**

**Ονοματεπώνυμα: Αικατερίνη Κουμπρόγλου, Μαρία Παπαδοπούλου**

**Αριθμός Μητρώου: 20390118, 20390178**

**Εξάμηνο: 9ο**

**Ημερομηνία Παράδοσης Εργασίας: 14/01/2025**

Περιεχόμενα

[**Βήμα 1. Συλλογή δεδομένων:** 2](#_Toc187083341)

# **Βήμα 1. Συλλογή δεδομένων:**

Για την υλοποίηση της συλλογής δεδομένων χρησιμοποιήθηκε το Wikipedia ως η πηγή των άρθρων. Με χρήση των βιβλιοθηκών requests και BeautifulSoup, ο κώδικας καταφέρνει να συλλέξει 10 άρθρα σχετικά με την θεματολογία: ≪Processor (computing)≫, το καθένα από τα οποία ανακτάται ο τίτλος, ο ιστότοπος και το περιεχόμενο του και αποθηκεύονται στο αρχείο articles.json.

## Δυσκολίες που αντιμετωπίστηκαν

Ο κώδικας είναι σχεδιασμένος να ελέγχει αν ο επόμενος ιστότοπος που θα εξετάσει για να συλλέξει τα δεδομένα του έχει ήδη ελεγχθεί στο παρελθόν. Καταφέρνει να λειτουργήσει σωστά ως επί το πλείστον όμως σε κάθε εκτέλεση θα εξετάσει τον πρώτο ιστότοπο δύο φορές συνεχόμενα το οποίο είναι κάτι που πρέπει να αποφευχθεί. Αυτό συμβαίνει πιθανότατα διότι μετά την επεξεργασία του πρώτου άρθρου, το οποίο είναι αυτό που δίνεται στον κώδικα, αντί να συνεχίσει με την αλληλουχία που του έχει ορισθεί πηγαίνει στο main page του Wikipedia. Αυτή η σελίδα παραλείπεται και δεν επεξεργάζεται περαιτέρω, όμως ίσως από αυτόν τον ιστότοπο ξεκινάει πάλι από την αρχή και πηγαίνει στην πρώτη σελίδα που του δόθηκε και επεξεργάστηκε ήδη.