Κατασκευή real dataset

Αποφασίσαμε να κατασκευάσουμε το δικό μας dataset με πραγματικά δεδομένα επιστημόνων, αντλούμενα από τη βάση δεδομένων της Wikipedia. Στο αρχείο **“download\_data.py”**, πραγματοποιείται ανάκτηση και επεξεργασία πληροφοριών σχετικά με επιστήμονες της επιστήμης των υπολογιστών, από τη Wikipedia. Ο κώδικας στοχεύει στη δημιουργία ενός dataset, που για κάθε επιστήμονα περιέχει το επώνυμό του, τον αριθμό βραβείων που έχει λάβει, καθώς και πληροφορίες σχετικά με την εκπαίδευσή του.

Αρχικά, με τη συνάρτηση **“get\_urls”**, ανακτώνται όλοι οι σύνδεσμοι (URLs) των επιστημόνων που περιέχονται στην κεντρική σελίδα με τη λίστα επιστημόνων της επιστήμης των υπολογιστών: <https://en.wikipedia.org/wiki/List_of_computer_scientists>. Καθένας από αυτούς τους συνδέσμους αποθηκεύεται σε μία λίστα, που είναι και αυτό που επιστρέφει η συνάρτηση.

Στη συνέχεια, καλούμε τη συνάρτηση **“get\_scientist\_info”** για κάθε σελίδα επιστήμονα, προκειμένου να εξάγουμε τις ζητούμενες πληροφορίες για καθέναν από τους επιστήμονες της λίστας. Για το σκοπό αυτό, κάνουμε χρήση της βιβλιοθήκης **BeautifulSoup**, με την οποία μπορούμε να επεξεργαστούμε το HTML περιεχόμενο κάθε σελίδας προκειμένου να βρούμε και να εξάγουμε συγκεκριμένα στοιχεία, όπως συνδέσμους, κείμενο και άλλα χαρακτηριστικά. Είναι ένα απαραίτητο εργαλείο γι’ αυτόν τον κώδικα, καθώς παρέχει τη δυνατότητα ευέλικτης και αποτελεσματικής ανάλυσης του HTML περιεχομένου, καθιστώντας την ανάκτηση και επεξεργασία των πληροφοριών ευκολότερη και πιο αυτοματοποιημένη.

Όταν έχουν ανακτηθεί όλα τα δεδομένα, δημιουργείται ένα DataFrame, το οποίο περιέχει τις πληροφορίες για κάθε επιστήμονα. Ακολούθως, ο κώδικας πραγματοποιεί διορθώσεις στο DataFrame, βάσει των πληροφοριών που περιέχονται στο αρχείο **“corrections.txt”**. Το αρχείο αυτό περιέχει εγγραφές με το επώνυμο του επιστήμονα, τον αριθμό των βραβείων του, καθώς και ενδεχομένως τον αριθμό γραμμής (index) του συγκεκριμένου επιστήμονα στο DataFrame. Σκοπός αυτής της διαδικασίας είναι να διορθώσουμε ορισμένες λανθασμένες καταχωρήσεις βραβείων που έγιναν κατά την εξαγωγή των δεδομένων, έτσι ώστε να ανταποκρίνονται στην πραγματικότητα. Οι λανθασμένες καταχωρήσεις οφείλονται στο γεγονός ότι δεν είναι όλες οι σελίδες της Wikipedia δομημένες κατά τον ίδιο τρόπο όσον αφορά το HTML περιεχόμενό τους.

Αφού εφαρμοστούν όλες οι απαραίτητες διορθώσεις στο DataFrame, τα δεδομένα εξάγονται σε ένα αρχείο CSV με την ονομασία **“scientists\_data.csv”**, που είναι και το τελικό dataset που χρησιμοποιήσαμε για την υλοποίηση της εργασίας. Το τελικό dataset περιέχει δεδομένα 254 επιστημόνων, που είναι αυτοί για τους οποίους υπήρχε κείμενο σχετικά με την εκπαίδευσή τους στην αντίστοιχη σελίδα τους.

Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, γραμματοσειρά, αριθμός

Περιγραφή που δημιουργήθηκε αυτόματα