Εργασία: Μηχανή αναζήτησης άρθρων σχετικών με τον COVID-19

Φάση 1: Αρχικός σχεδιασμός και συλλογή δεδομένων

Κωνσταντίνος Κρανάς, 2278

Συλλογή εγγράφων: Τα έγγραφα προέρχονται από το dataset του Our World in Data (owid github). Τα έγγραφα αυτά αποτελούνται από περίπου 80,000 πλειάδες με ημερήσια καταγραφή για κάθε χώρα των συνολικών κρουσμάτων, θανάτων, πλήθος ασθενών σε ΜΕΘ, πλήθος εμβολιασμών κλπ.

Προεπεξεργασία του dataset:

Το Dataset θα φιλτραριστεί και χωριστεί με ένα script σε 500 αρχεία των 160 πλειάδων. Κάθε χώρα έχει περίπου 500 καταγραφές, οπότε για να μην καταλήξει κάθε χώρα να έχει 3 αρχεία δικά της, κάθε πλειάδα θα κατανεμηθεί τυχαία σε κάποιο από τα 500 αρχεία. Τα αρχεία στο github θα είναι populated το καθένα από 2-3 χώρες, σε ανεξάρτητο σύνολο από τα λοιπά που θα κατανεμηθούν με το script. Επίσης, καθώς το csv αρχείο του dataset δεν ξεχώριζε τις τιμές με “” από τα κόμματα, το dataset υπέστη εξαγωγή από το excel αρχείο ως csv διαχωρισμένο με semicolon (;).

Τα κυριότερα χαρακτηριστικά της συλλογής είναι: **iso_code, continent, location, date, total_cases, new_cases, total_deaths, new_deaths, total_cases_per_million, total_deaths_per_million, reproduction_rate, icu_patients, icu_patients_per_million, weekly_icu_admissions, new_tests, total_tests, total_vaccinations..**

Το dataset είναι ήδη ταξινομημένο ως προς τον iso_code κάθε χώρας, οπότε ανατρέχοντας το μία φόρα (με script) μπορούμε εύκολα να προσκομίσουμε το λεξιλόγιο (**iso_code**, **continents**, **location**). Επίσης, μπορούμε στο λεξιλόγιο να προσθέσουμε τις χρονολογίες 2020 και 2021.

Tokens κατά κύριο λόγο δεν χρειάζονται να χωριστούν, εκτός από κάποιες περιπτώσεις όπου ο κωδικός χώρας έχει πρόθημα “OWID_”.

Αντεστραμμένα ευρετήρια με τους **iso_code**, τις **continents** και **location** και χρονιές.

Possible queries with this dataset:

1. Country/Region specific queries.
2. Country code like queries, where the country’s code is exactly 3 sequential alphabetical characters.
3. Only dates query σε αριθμούς που είναι της μορφής *yyyy-mm-dd.*

4. Non-dates query σε αριθμούς που δεν είναι της μορφής *yyyy-mm-dd.*
5. Range queries σε αριθμούς (και σε συνδυασμό με τα 3, 4).

PS: αν θεωρείτε ότι δεν αρκεί η συλλογή, ενημερώστε με στο cs02278 at uoi.gr ώστε να αλλάξω σε μία πιο text-based.