PP1 – Programming Project

Προσληφθήκατε σε μια εταιρία ανάπτυξης λογισμικού και το πρώτο σας project είναι να κατασκευάσετε μια μηχανή αναζήτησης πάνω σε μια συλλογή κειμένων. Καθώς είχατε διαφημίσει την εξειδίκευσή σας στις Βάσεις Δεδομένων (κακώς, όπως θα δείτε παρακάτω ;-)), πρέπει να υλοποιήσετε τη λύση σας χρησιμοποιώντας τις δυνατότητες της PostgreSQL πάνω στην Αναζήτηση Ελεύθερου Κειμένου (Free Text Search). Επιπλέον του υλικού που είδαμε στο μάθημα, δείτε και αυτό το link.

[ΜΕΡΟΣ Α – 30 μονάδες]
Σας δίνεται ένα αρχείο json[1] που περιέχει περίπου 100000 άρθρα από το PubMed Central. Κάθε γραμμή του αρχείου αντιστοιχίζεται σε ένα άρθρο και κάθε άρθρο έχει ένα title, ένα abstract, καθώς και άλλα βιβλιογραφικά μεταδεδομένα (λίστα συγγραφέων, πληροφορίες για το περιοδικό όπου δημοσιεύτηκε το άρθρο, κλπ).

Η δουλειά σας είναι να δημιουργήσετε και να γεμίσετε έναν πίνακα με το ακόλουθο σχήμα:

docs (**id** int generated by default as identity primary key, **title** text, **abstract** text)

Η πρώτη στήλη θα παίρνει τις τιμές της αυτόματα, ενώ το περιεχόμενο των επόμενων δυο στηλών θα προέλθει από το αρχείο json. Χρησιμοποιήστε τις json δυνατότητες της PostgreSQL για να φορτώσετε το αρχείο json απευθείας σε έναν temporary table και μετά εξάγετε τα απαιτούμενα πεδία json στις κατάλληλες στήλες του **docs** (το PostgreSQL manual και η αναζήτηση στο web είναι φίλοι σας, ειδικά για να βρίσκετε απαντήσεις στα λάθη που κάνετε).

[ΜΕΡΟΣ Β – 70 μονάδες]
Σε περίπτωση που δεν καταφέρετε να ολοκληρώσετε το ΜΕΡΟΣ Α, σας δίνεται ένα sql dump[2] του πίνακα **docs** ώστε να το εισάγετε στον PostgreSQL server σας και να προχωρήσετε στα επόμενα βήματα.

Αφού φορτώσετε τα δεδομένα στον server σας (πάλι πρέπει να ψάξετε και να βρείτε πως γίνεται αυτό), προσθέστε δυο επιπλέον στήλες ώστε το σχήμα του πίνακα να τροποποιηθεί ως εξής:

docs (**id** int generated by default as identity primary key, **title** text, **abstract** text,
title_tsv tsvector, **abstract_tsv** tsvector)

Δημιουργήσετε τα κατάλληλα ευρετήρια και γεμίστε τις νέες στήλες χρησιμοποιώντας τη συνάρτηση to_tsvector() (δείτε τον PostgreSQL κώδικα από τη Διάλεξη 04).

Τώρα μπορείτε να απαντήσετε τα παρακάτω αιτήματα:

1. Βρείτε το πλήθος των docs που στο title περιέχουν τους όρους 'rat' ή 'liver'.
2. Βρείτε το πλήθος των docs που στο abstract περιέχουν τους όρους 'rat' ή 'liver'.
3. Βρείτε το πλήθος των docs που στο title ή στο abstract περιέχουν τους όρους 'rat' ή 'liver'.
4. Βρείτε το πλήθος των docs που και στο title και στο abstract περιέχουν τους όρους 'rat' ή 'liver'.
5. Απαντήστε τα παραπάνω τέσσερα αιτήματα όταν θέλουμε να περιέχονται οι όροι 'rat' και 'liver'.
6. Βρείτε όλα τα docs που περιέχουν στο abstract τους όρους 'cancer' και 'liver'. Κατατάξτε τις απαντήσεις με τη χρήση της συνάρτησης ts_rank_cd σε φθίνουσα κατάταξη.
7. Βρείτε τους top 10 όρους (terms) ως προς το document frequency.
8. Βρείτε τους top 10 όρους (terms) ως προς το collection frequency.

[EXTRA CREDIT – 20 μονάδες]
Φτιάξτε μια απλή web app που να λειτουργεί ως μηχανή αναζήτησης πάνω στη βάση.

1 Το αρχείο data.txt.zip υπάρχει εδώ.
2 Το αρχείο docs.dump.zip υπάρχει εδώ. Θα πρέπει να αλλάξετε μέσα στο αρχείο την εντολή "ALTER TABLE public.docs OWNER TO gevan" ώστε να αναφέρεται στο όνομα του χρήστη με το οποίο συνδέεστε στην PostgreSQL.