ΠΡΟΓΡΑΜΜΑΤΙΣΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

ΠΑΝΑΓΙΩΤΗΣ ΤΟΛΟΥΔΗΣ

Προχωρημένη διαχείριση δεδομένων 2022-23
Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών
Πανεπιστήμιο Θεσσαλίας, Βόλος
ptoloudis@e-ce.uth.gr

1 Θέμα

Δημιουργία μιας βάσης δεδομένων για την ανακάλυψη φαρμακευτικών ουσιών.

2 Είδος NoSQL ΒΔ:

Document ή Column Database

3 Περιγραφή και χρησιμότητα της εφαρμογής:

Η δημιουργία μιας βάσης δεδομένων για την ανακάλυψη φαρμάκων θα περιλαμβάνει τη συλλογή και την οργάνωση δεδομένων που σχετίζονται με χημικές ενώσεις, πρωτεΐνες-στόχους, αποτελεσματικότητα φαρμάκων και άλλες σχετικές πληροφορίες. Στόχος της βάσης δεδομένων είναι να παρέχει ένα ολοκληρωμένο και εύκολα προσβάσιμο αποθετήριο πληροφοριών που μπορούν να χρησιμοποιηθούν από ερευνητές και επιστήμονες για την ανάπτυξη νέων φαρμάκων και θεραπειών.

Η δημιουργία μιας βάσης δεδομένων για την ανακάλυψη φαρμάκων απαιτεί προσεκτικό σχεδιασμό και εξέταση των διαφόρων τύπων δεδομένων, πηγών και απαιτήσεων που εμπλέκονται. Μια καλά σχεδιασμένη βάση δεδομένων μπορεί να παρέχει ένα ισχυρό εργαλείο για την επιτάχυνση της ανακάλυψης φαρμάκων και τη βελτίωση των αποτελεσμάτων των ασθενών.

4 Επιλογή Συστήματος NoSQLΒΔ:

Υπάρχουν αρκετές βάσεις δεδομένων NoSQL που θα μπορούσαν να είναι κατάλληλες για τη διαχείριση δεδομένων βιοπληροφορικής, ανάλογα με τις συγκεκριμένες απαιτήσεις του έργου. Ακολουθούν ορισμένες προτάσεις:

– MongoDB: Το MongoDB είναι μια δημοφιλής βάση δεδομένων NoSQL προσανατολισμένη σε έγγραφα που μπορεί να χειριστεί μεγάλες ποσότητες ημι-δομημένων και μη δομημένων δεδομένων. Παρέχει ευέλικτες σχεδιάσεις σχημάτων που μπορούν εύκολα να προσαρμοστούν στις μεταβαλλόμενες ανάγκες δεδομένων και υποστηρίζει εμπλουτισμένες δυνατότητες ερωτημάτων και συνάθροισης.

- Apache Cassandra: Το Apache Cassandra είναι μια εξαιρετικά κλιμακούμενη, κατανεμημένη βάση δεδομένων NoSQL που έχει σχεδιαστεί για υψηλή διαθεσιμότητα και ανοχή σφαλμάτων. Είναι κατάλληλο για το χειρισμό μεγάλων όγκων δεδομένων και παρέχει ρυθμιζόμενα επίπεδα συνέπειας που μπορούν να διαμορφωθούν με βάση τα συγκεκριμένα μοτίβα πρόσβασης δεδομένων.
- Neo4j: Το Neo4j είναι μια βάση δεδομένων γραφημάτων που έχει βελτιστοποιηθεί για την αποθήκευση και την αναζήτηση εξαιρετικά συνδεδεμένων δεδομένων, όπως βιολογικά δίκτυα ή δεδομένα γονιδιακής έκφρασης. Παρέχει ένα πλούσιο σύνολο αλγορίθμων γραφημάτων και εργαλείων οπτικοποίησης και υποστηρίζει συναλλαγές ACID για τη διασφάλιση της ακεραιότητας των δεδομένων.
- Couchbase: Το Couchbase είναι μια βάση δεδομένων NoSQL που συνδυάζει την ευελιξία των βάσεων δεδομένων προσανατολισμένων σε έγγραφα με την απόδοση των καταστημάτων κλειδιού-τιμής. Υποστηρίζει πολυδιάστατη κλιμάκωση, επιτρέποντας την οριζόντια κλιμάκωση τόσο των δεδομένων όσο και της επεξεργαστικής ισχύος.

Τελικά, η επιλογή της βάσης δεδομένων NoSQL θα εξαρτηθεί από παράγοντες όπως το μέγεθος και η πολυπλοκότητα των δεδομένων, η συγκεκριμένη περίπτωση χρήσης και οι διαθέσιμοι πόροι και εμπειρογνωμοσύνη.

Επέλεξα την MongoDB γιατί έχω ασχοληθεί και θα χρησιμοποιήσω το OnLine πρόγραμμα που έχει τα έξεις χαρακτηριστικά: STORAGE 512 MB, RAM Shared, vCPU Shared.

5 Δημιουργία δεδομένων με χρήση datagenerator ή ανεύρεση έτοιμων δεδομένων.

Έκανα αίτηση για να πάρω δεδομένα από RxNorm, BioGRID και το DrugBank. Η έγκριση για τα δυο τελευταία ήρθε μετά το Πάσχα, ενώ για το πρώτο είναι ελεύθερο αλλά δεν μπόρεσε να επιλέξω τα δεδομένα. Τελικά πείρα τα δεδομένα από το Kaggle.

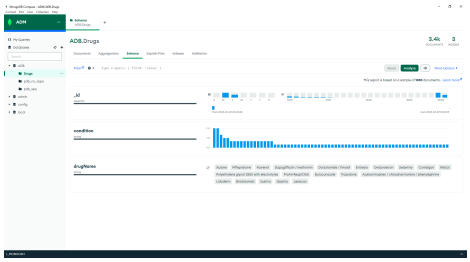
Τα link από τα dataset:
`https://www.kaggle.com/datasets/shahir/protein-data-set`,
`https://www.kaggle.com/datasets/jessicali9530/kuc-hackathon-winter-2018`.

Μετά αφαίρεσα πολλά δεδομένα γιατί είχαν πάρα πολλά δεδομένα. Τα τελικά είναι στο φάκελο data.

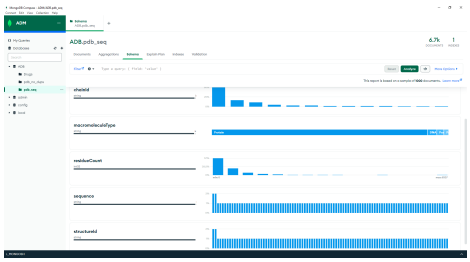
5.1 Η ανάλυση απο τα τελικά δεδομένα.

6 Τρέξιμο του κώδικα.

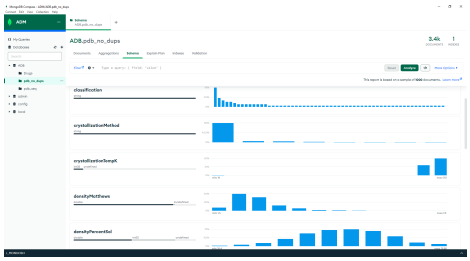
- Κατεβάζουμε και εγκαθιστούμε το node.js. `https://nodejs.org/en/download`.
- Στο τερματικό πάμε στο φάκελο του project.
- Αλλάζουμε ή δημιουργούμε στον φάκελο library το αρχείο με όνομα login.env. Το οποίο αποθηκεύει όλα τα login link για την MongoDB.
- Εγκαθιστούμε τις βιβλιοθήκες με την εντολή npm install package.json
- Κάνουμε compile με την εντολή tsc <όνομα>.ts
- Για το τρέξιμο γράφουμε την εντολή node <όνομα>.js ορίσματα.



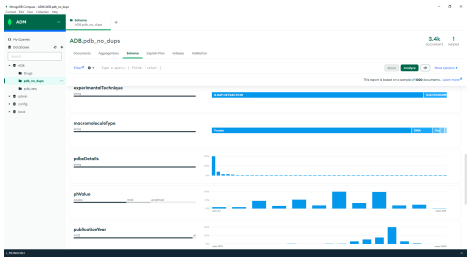
Εικ. 1. Για το Drugs



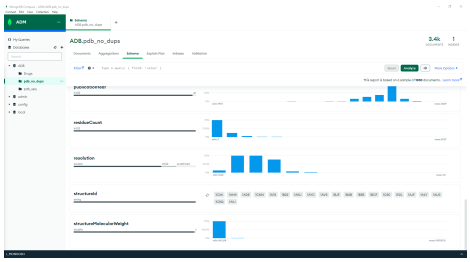
Εικ. 2. Για το pdb_seq



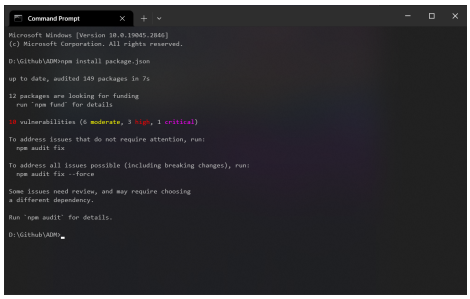
Εικ. 3. Για το pdb_no_dups 1



Εικ. 4. Για το pdb_no_dups 2



Εικ. 5. Για το pdb_no_dups 3



Εικ. 6. Instalasion

7 Φόρτωση των δεδομένων στη ΒΔ.

Πρέπει να τρέξετε στον φάκελο load τα δυο αρχεία. Το load.ts είναι για το upload, ενώ για το set_unique.ts είναι να θέσουμε τα unique.

8 API.

Το manual με για το API είναι στο manual.pdf.

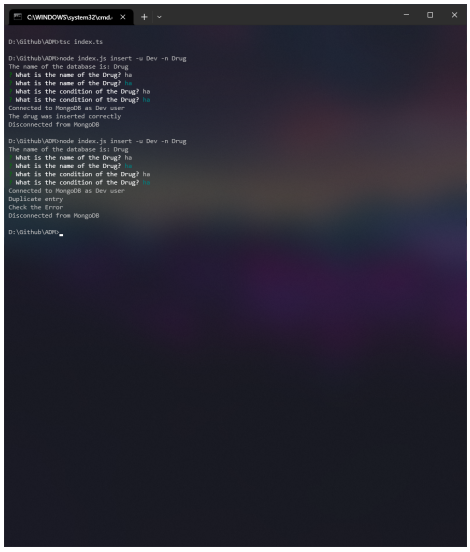
9 Για τα δικά μου ερωτήματα.

Τρέχω την εντολή node index.js θα εμφανιστούν οι υλοποιημένες εντολές.

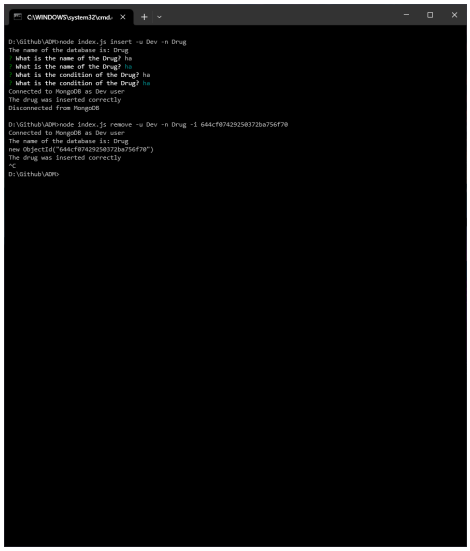
10 Μελλοντικές επεκτάσεις και προσθήκες.

- Δημιουργία γραφικό παραθύρου.
- Δημιουργία web application.
- Προσθήκη περισσότερων δεδομένων και ερωτημάτων.

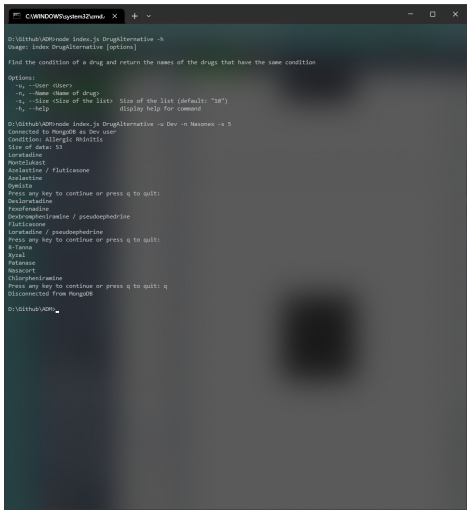
Εικ. 7. Help



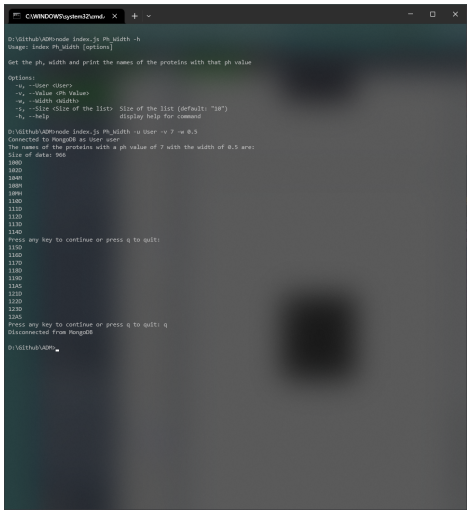
Εικ. 8. Insert



Εικ. 9. Remove



Εικ. 10. Drug Alternative



Εικ. 11. Ph Width

Εικ. 12. SequenceId