

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 εμπειρογνομosύνη.

Επέλεξα την MongoDB γιατί έχω ασχοληθεί και θα χρησιμοποιήσω το OnLine πρόγραμμα που έχει τα εξής 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:

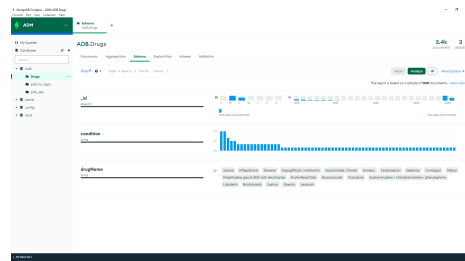
<https://www.kaggle.com/datasets/shahir/protein-data-set>,
<https://www.kaggle.com/datasets/jessicali9530/kuc-hackathon-winter-2018>.

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

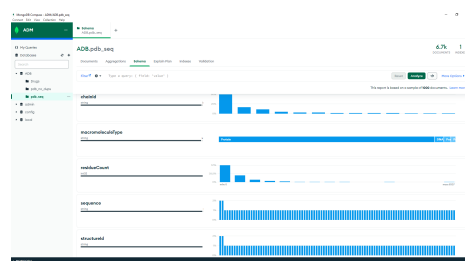
5.1 Η ανάλυση απο τα τελικά δεδομένα.

6 Τρέξιμο του κώδικα.

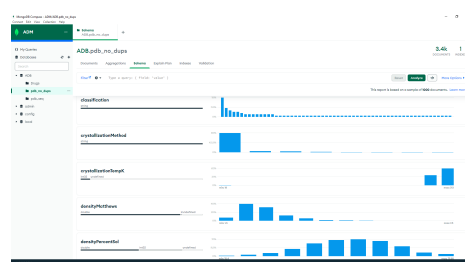
- Κατεβάζουμε και εγκαθιστούμε το node.js. <https://nodejs.org/en/download>.
- Στο τερματικό πάμε στο φάκελο του project.
- Αλλάζουμε ή δημιουργούμε στον φάκελο library το αρχείο με όνομα login.env. Το οποίο αποθηκεύει όλα τα login link για την MongoDB.
- Εγκαθιστούμε τις βιβλιοθήκες με την εντολή `npm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



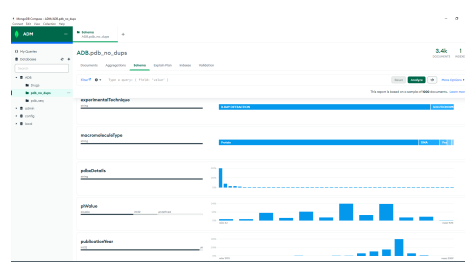
Εικ. 1. Για το Drugs



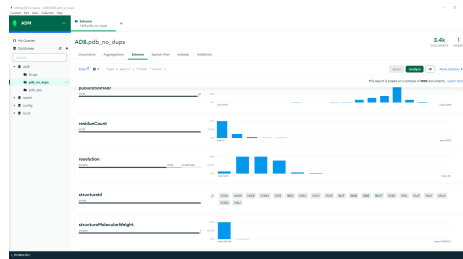
Εικ. 2. Για το pdb_seq



Εικ. 3. Για το pdb_no_dups 1



Εικ. 4. Για το pdb_no_dups 2



Εικ. 5. Για το pdb_no_dups 3

```

Microsoft Windows [Version 10.0.19045.2846]
(c) Microsoft Corporation. All rights reserved.

D:\Github\ADMPipe install package.json
up to date, audited 149 packages in 7s

12 packages are looking for funding
  run `npm fund` for details

found 0 vulnerabilities (0 moderate, 3 high, 1 critical)

To address issues that do not require attention, run:
  npm audit fix

To address all issues possible (including breaking changes), run:
  npm audit fix --force

Some issues need review, and may require choosing
a different dependency.

Run `npm audit` for details.
D:\Github\ADMPipe

```

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.

```

C:\WINDOWS\system32\cmd.exe
D:\Vithub\ADPnode>index.js
Usage: index [options] [-command]

Options:
  -h, --help            display help for command

Comments:
DrugCondition [options] Find the condition of a drug and return the names of the drugs that have the same condition
Protein [options]       Print the protein information of a protein
PhValue [options]       Get the ph and print the names of the proteins with that ph value
PhWidth [options]       Get the ph, width and print the names of the proteins with that ph value
Class_Name [options]    Get the class name and print the names of the proteins with that class name. If the class
                        name is DNA-RNA HYBRID, press "DNA-RNA HYBRID"
Sequence [options]      Get the name and print the sequence of the proteins
SequenceId [options]    Get the sequence and print the name of the proteins
ProteinId [options]     Get the name and print all the information and sequence of the proteins
Condition [options]     Print all Contradictions of a drug
Insert [options]         Insert a new
Remove [options]         Remove from a database
Update [options]         Update from a database
help [Command]          display help for command

D:\Vithub\ADP>

```

Рис. 7. Help

```

C:\WINDOWS\system32\cmd.exe
D:\Vithub\ADPnode>index.js
D:\Vithub\ADPnode>index.js insert -s Dev -s Drug
The case of the database is: Drug
What is the name of the Drug? ha
What is the name of the Drug? ha
What is the condition of the Drug? ha
Connected to MongoDB as dev user
The drug was inserted correctly
Disconnected from MongoDB

D:\Vithub\ADPnode>index.js insert -s Dev -s Drug
The case of the database is: Drug
What is the name of the Drug? ha
What is the name of the Drug? ha
What is the condition of the Drug? ha
What is the condition of the Drug? ha
Connected to MongoDB as dev user
Duplicate entry
Check the error
Disconnected from MongoDB

D:\Vithub\ADP>

```

Рис. 8. Insert

```
C:\WINDOWS\system32\cmd.exe
```

```
D:\Vilitha\ADNode>Index,Is insert -> Dev -n Drug
The name of the database is: Drug
What is the name of the Drug? he
What is the name of the Drug? he
What is the condition of the Drug? ha
The drug was inserted correctly
Connected to Hadoop as Dev user
The drug was inserted correctly
Disconnected from Hadoop

D:\Vilitha\ADNode>Index,Is remove -> Dev -n Drug -l 64dc7074293d8772ba756f79
Connected to Hadoop as Dev user
The name of the database is: Drug
new Object[] {Integer.valueOf(64dc7074293d8772ba756f79)}
The drug was inserted correctly
OK
D:\Vilitha\ADNode>
```

Ек. 9. Remove

Euk. 10. Drug Alternative

[illegible]

Euk. 11. Ph Width

```

C:\WINDOWS\system32\cmd.exe
@GistHub\ADPmode Index, % SequenceId %
  help Index SequenceId (options)

Get the sequence and print the name of the proteins

Options:
  -u, --user <user>
  -s, --Sequence <Sequence>
  -h, --help                display help for command

@GistHub\ADPmode Index, % SequenceId % User -> P4NQLCDSHVAIYVLCGGSRFFPTAT
connected to HangoDB as user user
The name of the protein with the sequence P4NQLCDSHVAIYVLCGGSRFFPTAT is:
A428
A429
A430
A431
A432
A433
A434
A435
A436
A437
A438
A439
A440
A441
A442
A443
A444
A445
A446
A447
A448
A449
A450
A451
A452
A453
A454
A455
A456
A457
A458
A459
A460
A461
A462
A463
A464
A465
A466
A467
A468
A469
A470
A471
A472
A473
A474
A475
A476
A477
A478
A479
A480
A481
A482
A483
A484
A485
A486
A487
A488
A489
A490
A491
A492
A493
A494
A495
A496
A497
A498
A499
A500
A501
A502
A503
A504
A505
A506
A507
A508
A509
A510
A511
A512
A513
A514
A515
A516
A517
A518
A519
A520
A521
A522
A523
A524
A525
A526
A527
A528
A529
A530
A531
A532
A533
A534
A535
A536
A537
A538
A539
A540
A541
A542
A543
A544
A545
A546
A547
A548
A549
A550
A551
A552
A553
A554
A555
A556
A557
A558
A559
A560
A561
A562
A563
A564
A565
A566
A567
A568
A569
A570
A571
A572
A573
A574
A575
A576
A577
A578
A579
A580
A581
A582
A583
A584
A585
A586
A587
A588
A589
A590
A591
A592
A593
A594
A595
A596
A597
A598
A599
A600
A601
A602
A603
A604
A605
A606
A607
A608
A609
A610
A611
A612
A613
A614
A615
A616
A617
A618
A619
A620
A621
A622
A623
A624
A625
A626
A627
A628
A629
A630
A631
A632
A633
A634
A635
A636
A637
A638
A639
A640
A641
A642
A643
A644
A645
A646
A647
A648
A649
A650
A651
A652
A653
A654
A655
A656
A657
A658
A659
A660
A661
A662
A663
A664
A665
A666
A667
A668
A669
A670
A671
A672
A673
A674
A675
A676
A677
A678
A679
A680
A681
A682
A683
A684
A685
A686
A687
A688
A689
A690
A691
A692
A693
A694
A695
A696
A697
A698
A699
A700
A701
A702
A703
A704
A705
A706
A707
A708
A709
A710
A711
A712
A713
A714
A715
A716
A717
A718
A719
A720
A721
A722
A723
A724
A725
A726
A727
A728
A729
A730
A731
A732
A733
A734
A735
A736
A737
A738
A739
A740
A741
A742
A743
A744
A745
A746
A747
A748
A749
A750
A751
A752
A753
A754
A755
A756
A757
A758
A759
A760
A761
A762
A763
A764
A765
A766
A767
A768
A769
A770
A771
A772
A773
A774
A775
A776
A777
A778
A779
A780
A781
A782
A783
A784
A785
A786
A787
A788
A789
A790
A791
A792
A793
A794
A795
A796
A797
A798
A799
A800
A801
A802
A803
A804
A805
A806
A807
A808
A809
A810
A811
A812
A813
A814
A815
A816
A817
A818
A819
A820
A821
A822
A823
A824
A825
A826
A827
A828
A829
A830
A831
A832
A833
A834
A835
A836
A837
A838
A839
A840
A841
A842
A843
A844
A845
A846
A847
A848
A849
A850
A851
A852
A853
A854
A855
A856
A857
A858
A859
A860
A861
A862
A863
A864
A865
A866
A867
A868
A869
A870
A871
A872
A873
A874
A875
A876
A877
A878
A879
A880
A881
A882
A883
A884
A885
A886
A887
A888
A889
A890
A891
A892
A893
A894
A895
A896
A897
A898
A899
A900
A901
A902
A903
A904
A905
A906
A907
A908
A909
A910
A911
A912
A913
A914
A915
A916
A917
A918
A919
A920
A921
A922
A923
A924
A925
A926
A927
A928
A929
A930
A931
A932
A933
A934
A935
A936
A937
A938
A939
A940
A941
A942
A943
A944
A945
A946
A947
A948
A949
A950
A951
A952
A953
A954
A955
A956
A957
A958
A959
A960
A961
A962
A963
A964
A965
A966
A967
A968
A969
A970
A971
A972
A973
A974
A975
A976
A977
A978
A979
A980
A981
A982
A983
A984
A985
A986
A987
A988
A989
A990
A991
A992
A993
A994
A995
A996
A997
A998
A999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
121
```

Еук. 12. SequenceId