

# **Συγκομιδή και Διαχείριση Δεδομένων Παγκόσμιου Ιστού για Σύστημα Συστάσεων Επιστημονικών Δημοσιευμάτων σε NoSQL Βάσεις Δεδομένων**

Αθηνά Λιακοπούλου Α.Ε.Μ.:2428

## **ΕΙΔΙΚΟ ΘΕΜΑ 2020-21 - 9ο εξάμηνο**

Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών  
Πανεπιστήμιο Θεσσαλίας, Βόλος  
atliakopoulou@e-ce.uth.gr

**Περίληψη** Σκοπό της εργασίας αυτής αποτελεί η επεξεργασία και αξιοποίηση ενός big dataset από τον παγκόσμιο ιστό. Συγκεκριμένα, η βάση δεδομένων που χρησιμοποιήθηκε αφορά σε συγγραφείς και έργα που έχουν δημοσιευτεί. Τα βασικά ερωτήματα που αναπτύσσονται παρακάτω είναι κάθε συγγραφέας τί αναφέρει έχει κάνει, και αντίστροφα κάθε έργο από ποιους συγγραφείς έχει χρησιμοποιηθεί ως αναφορά.

Για την ανάπτυξη αυτών των ερωτημάτων έχει χρησιμοποιηθεί ως NoSQL βάση δεδομένων η MongoDB που είναι κατάλληλη για την αποθήκευση και επεξεργασία big data. Τέλος η διαδικασία της επεξεργασίας των δεδομένων αναπτύχθηκε με δύο διαφορετικές μεθόδους: Αρχικά με τη χρήση java γλώσσας, που η υλοποίηση περιέχει ως user interface ένα web app που αναπτύχθηκε με τη χρήση του eclipse workspace. Επίσης τα ερωτήματα αναπτύχθηκαν και με python γλώσσα σε jupyter notebook με στόχο τη βελτιστοποίηση της απόδοσής τους.

Η επεξεργασία των δεδομένων και η ανάπτυξη των προαναφερθέντων ερωτημάτων θεωρώ ότι αποτελεί μια εισαγωγή για την ανάπτυξη ενός ολοκληρωμένου συστήματος συστάσεων στο συγκεκριμένο dataset.

**Λέξεις Κλειδιά:** **big data, MongoDB, web app, eclipse workspace, java, python**

## **1 Οδηγίες εγκατάστασης**

→ Τα παρακάτω βήματα θα μπορούσαν να αποφευχθούν αν το web app μπορούσε να εγκατασταθεί σε ένα virtual machine ώστε να λειτουργεί κανονικά ως site αλλά λόγω big data αυτό δεν μπορεί να γίνει στα δωρεάν διαθέσιμα VMs που προσφέρονται στα πλαίσια της σχολής οπότε πρέπει να γίνουν τα παρακάτω βήματα για την ανάπτυξη του user interface.

### **1.1 MongoDB**

Οδηγίες για την εγκατάσταση όλων των εργαλείων που χρησιμοποίησα για την ανάπτυξη και το γέμισμα της βάσης δεδομένων:

```

•download mongodb - https://www.mongodb.com/try/download/community
•download data - https://www.aminer.org/citation
•download mongodb compass - https://www.mongodb.com/try/download/compass
in C:\Program Files\MongoDB\Server\4.4\bin
download (extract)
mongodb-database-tools-windows-x86_64-100.2.0 (omngimport command)
•cd C:\Program Files\MongoDB\Server\4.4\bin\mongodb -database-tools-windows-x86_64-100.2.0\bin\mongodump --db MY_DB --collection users --drop --jsonArray
--batchSize 1 --file ./C:\Users\hp\Desktop\ΣΧΟΛΗ\ειδικό\dblp.v12.json

```

## 1.2 Java

- Για εγκατάσταση java στον υπολογιστή (C:\Program Files\Java\jdk-15.0.1)
- Java jdk download(<https://www.oracle.com/java/technologies/javase-downloads.html>)
- Run jdk-\_\_\_\_\_windows-x64 bin.exe
- Go to system properties of computer -> environment variables:
- System variables -> path ->edit -> New -> C:\Program Files\Java\jdk-15.0.1\bin System variables ->new system variable ->
 Variable name = JAVA\_HOME, Variable value = C:\Program Files\Java\jdk-15.0.1
- ok to all and close

## 1.3 Eclipse

- Install eclipse (<https://www.eclipse.org/downloads/>)
- Run eclipse installer.exe choose eclipse ide for enterprise java developers
- Connect with filepath for java (C:\Program Files\Java\jdk-15.0.1)
- Install apache tomcat 9 (<https://tomcat.apache.org/download-90.cgi>) - 32-bit/64-bit Windows Service Installer (pgp, sha512)
- Run apache tomcat set up
- Type of install ->full
- On configuration don't change anything
- Go to services of computer
- Choose apache tomcat 9 ->startup type ->properties -> startup type change to -> manual ok and close (εγώ το έχω αυτόματα)
- Right click on apache tomcat9 ->start
- If you write on browser: <http://localhost:8080> then you should see "If you're seeing this, you've successfully installed Tomcat. Congratulations!"
- Open eclipse workspace
- Go to servers
- There should be none available
- Window-> show view -> other -> server ->servers -> Apache v9.0 server ->next -> in tomcat installation folder browse to where you downloaded apache (C:\Program

Files (x86) \Apache Software Foundation\Tomcat 9.0)

- Install maven plugin into eclipse : [https://hiplab.mc.vanderbilt.edu/projects/soempi/eclipse\\_m2e\\_install.html](https://hiplab.mc.vanderbilt.edu/projects/soempi/eclipse_m2e_install.html)
- Finish

#### 1.4 open project

- Download EidikoThema.rar from <https://github.com/AthinaLiakopoulou/EidikoThema>
  - Extract folder to eclipse workspace (C:\Users\hp\workspace)
  - In eclipse workspace file-> new ->import-> existing project into workspace -> browse to folder extracted.
- To run it as web app :
- Eidiko Thema-> WebContent -> right click on index.html ->run as ->run on server
  - Copy url of page that appeared and paste it in your browser

## 2 Μορφή Βάσης Δεδομένων

Η βάση δεδομένων είναι σε μορφή JSON αρχείου. Το σχήμα φαίνεται παρακάτω στις εικόνες. Στη συγκεκριμένη έκδοση υπάρχουν τα εξής χαρακτηριστικά: id, title, authors.name, authors.org, authors.id, venue.id, venue.raw, year, fos.name, fos.w, references, n\_citation, page\_start, page\_end, doc\_type, publisher, volume, issue, doi, indexed\_abstract. Id involved in V12 (paper id, author id, venue id) is represented as Long type (instead of string).

Field Name	Field Type	Description	Example
id	string	paper ID	53e9ab9eb7602d970354a97e
title	string	paper title	Data mining: concepts and techniques
authors.name	string	author name	Jiawei Han
author.org	string	author affiliation	Department of Computer Science, University of Illinois at Urbana-Champaign
author.id	string	author ID	53f42f36dabfaedce54dd0c
venue.id	string	paper venue ID	53e17f5b20f7dfbc07e8ac6e
venue.raw	string	paper venue name	Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial
year	int	published year	2000
keywords	list of strings	Keywords	["data mining", "structured data", "world wide web", "social network", "relational data"]
fos.name	string	paper fields of study	Web mining
fos.w	float	fields of study weight	0.659690857
references	list of strings	paper references	["4909282", "16018031", "16159250", "19838944", ...]
n_citation	int	citation number	40829
page_start	string	page start	11
page_end	string	page end	18

**Εικ. 1.** Database schema

n_citation	int	citation number	40829
page_start	string	page start	11
page_end	string	page end	18
doc_type	string	paper type: journal, book title...	book
lang	string	detected language	en
publisher	string	publisher	Elsevier
volume	string	volume	10
issue	string	issue	29
issn	string	issn	0020-7136
isbn	string	isbn	1-55660-489-8
doi	string	doi	10.4114/ia.v10i29.873
pdf	string	pdf URL	/static.aminer.org/upload/pdf/1254/ 370/239/53e9ab9eb7602d970354a97e.pdf
url	list	external links	["http://dx.doi.org/10.4114/ia.v10i29.873", "http://polar.lsi.uned.es/revista/index.php/ia/article/view/479"]
abstract	string	abstract	Our ability to generate...
indexed_abstract	dict	indexed abstract	{"IndexLength": 164, "Inverte dIndex": {"Our": [0], "ability": [1], "to": [2, 7, ...]}}

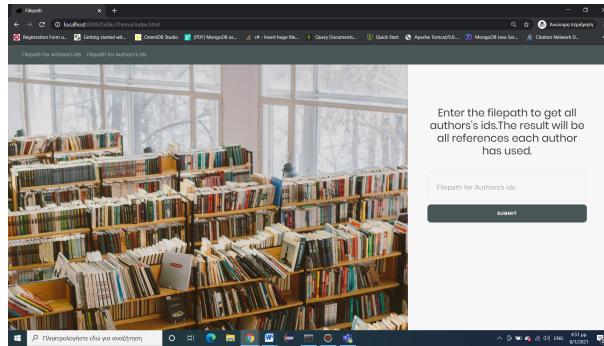
**Εικ. 2.** Database schema

### 3 Ανάπτυξη Java Web App

Στην εφαρμογή έχουν αναπτυχθεί 2 ερωτήσεις:

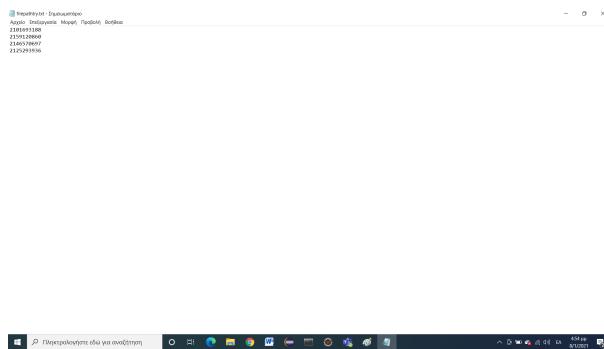
#### 3.1 1o ερώτημα

Ο χρήστης καλείται να δώσει το filepath ενός αρχείου που περιέχει ids συγγραφέων. Στην επόμενη σελίδα εμφανίζεται ένα block για κάθε συγγραφέα που δείχνει πόσα references έχει χρησιμοποιήσει συνολικά σε όλα τα άρθρα του και ποια (το κάθε άρθρο εμφανίζεται μια μοναδική φορά). Το ίδιο αποτέλεσμα γράφεται και σε ένα αρχείο που δημιουργείται αυτόματα (question1.txt).



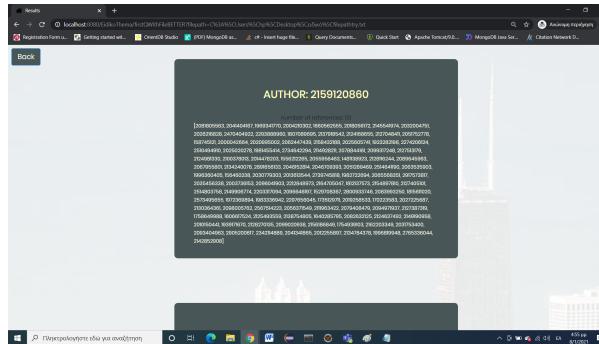
**Εικ. 3.** Starting page for 1rst quetion

Για παράδειγμα βάζω ως input το παρακάτω αρχείο (filepathtry.txt):



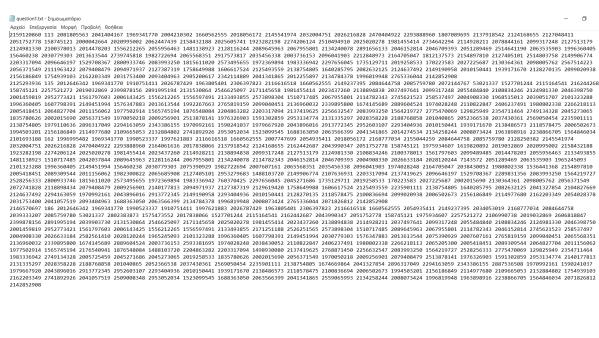
**Εικ. 4.** filepathtry.txt

Στο οποίο έχω πάρει 4 τυχαία authors' ids από τη mongodb βάση.  
Ως έξοδος προκύπτει:



Εικ. 5. 1st Question page

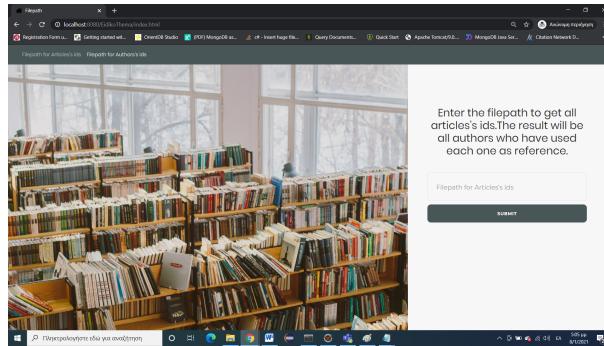
Οπως φαίνεται δημιουργείται ένα block για κάθε συγγραφέα που αναφέρεται πόσα references έχει χρησιμοποιήσει και ποια.  
Ταυτόχρονα δημιουργείται και το αρχείο που προανέφερα (question1.txt) και έχει την παρακάτω μορφή:



Εικ. 6. question1.txt

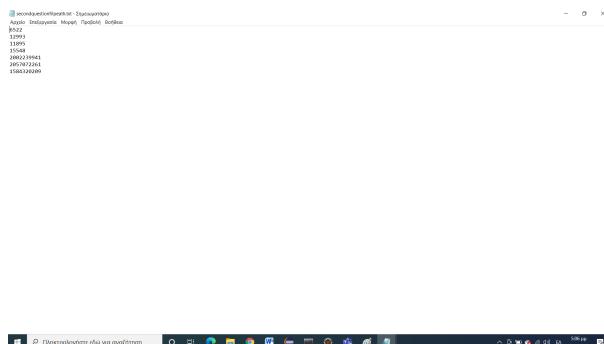
### 3.2 2o ερώτημα

Ο χρήστης καλείται να δώσει το filepath ενός αρχείου που περιέχει ids άρθρων. Στην επόμενη σελίδα εμφανίζεται ένα block για κάθε άρθρο που δείχνει πόσοι συγγραφείς το έχουν χρησιμοποιήσει ως reference και ποιοι (ο κάθε συγγραφέας εμφανίζεται μια μοναδική φορά). Το ίδιο αποτέλεσμα γράφεται και σε ένα αρχείο που δημιουργείται αυτόματα (queston2.txt).



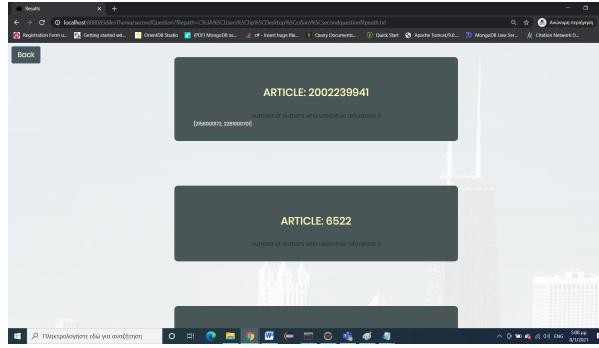
**Εικ. 7.** Starting page for 2nd question

Για παράδειγμα βάζω ως input το παρακάτω αρχείο (secondquestionfilpeath.txt):



**Εικ. 8.** secondquestionfilpeath.txt

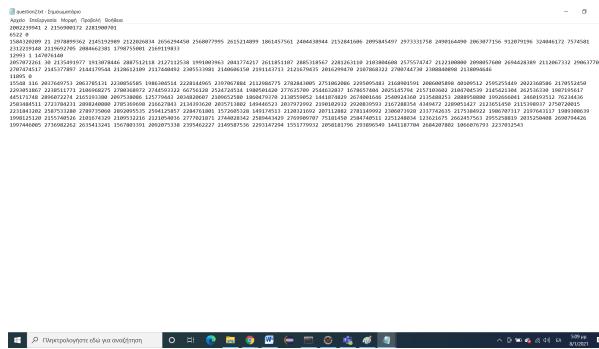
Στο οποίο έχω 7 τυχαία articles' ids από τη mongoDB βάση.  
Ως έξοδος προκύπτει:



**Εικ. 9.** 2nd question page

Όπως φαίνεται δημιουργείται ένα block για κάθε άρθρο που αναφέρεται πόσοι συγγραφείς το έχουν χρησιμοποιήσει ως references και ποια.

Ταυτόχρονα δημιουργείται και το αρχείο που προανέφερα (question2.txt) και έχει την παρακάτω μορφή:



**Εικ. 10.** question2.txt

## 4 Python Υλοποίηση

Την ίδια εφαρμογή υλοποίησα και σε γλώσσα python χωρίς να την αναπτύσσω ως web app.

## 4.1 1ο ερώτημα

Ο χρήστης καλείται να τρέξει το .ipyynb αρχείο σε Jupyter notebook και να δώσει το filepath ακριβώς όπως και πριν και προκύπτουν πάλι τα ίδιας μορφής αρχεία.

1ο ερώτημα:

```
In [5]: #list question
import pymongo
from pymongo import MongoClient
#-----
filepath = input("Give filepath for first question: ")
#filepath = "1questionpython.txt"
authors_ids = []
#-----
client = MongoClient('localhost', PORT)
#-----lost variables for MongoDB
#-----DON'T USE ALMOST
PORT = 27017
# create an instance of MongoClient()
client = MongoClient(
    host = 'localhost' + str(PORT)
) # get the database names from the MongoClient()
db = client.local # get the database names()
print ("databases:", db.database_names())
# create database & collection instances
collection = db.users
collection.create_index("authors.id");
#-----connection.find({"authors.id": {"$in": authors_ids}})
results = []
for id in authors_ids:
    print(id)
    print(collection.getShardDistribution())
    for doc in collection.find({"authors.id": id}):
        print(doc)
        key_Set=doc.keys()
        if id in key_Set:
            references = doc["references"]
            AllAuthorsofId = doc["authors"]
            AllAuthorsofId_ids = []
            for a in AllAuthorsofId:
                AllAuthorsofId_ids.append(a["id"])
            for id2 in AllAuthorsofId_ids:
                if id2 in authors_ids:
                    if id2 in results.keys():
                        v1=[]
                        previous=1
                        for i in results.keys():
                            previous+=1
                            if i == id2:
                                previous+=1
                                v1.append(previous)
                                s.extend(v1)
                                s.extend([reference])
                                del(s[-1])
                                dict_keys(s)
                                results.update();
                    else:
                        d_id2= references
                        results.update();
            for one_ids in authors_ids:
                if one_ids not in results.keys():
                    ds=(one_ids,[])
                    results[one_ids]=ds
path = "1questionpython.txt";
if (os.path.exists(path)):
    f=open("1questionpython.txt","w")
    f.truncate()
else:
    f=open("1questionpython.txt","w+")
for key in results.keys():
    f.write(str(key))
    f.write("\n")
    f.write(str(len(results[key])))
    f.write("\n")
    for a in results[key]:
        f.write(str(a))
        f.write("\n")
    f.write("\n")
f.close()
Give filepath for first question: filepathtry.txt
```

**Euk. 11.** python code for 1st question

Output file: 1questionpython.txt



## Euk. 12. 1questionpython.txt

## 4.2 2ο ερώτημα

Αντίστοιχα ακολουθεί ο κώδικας που χρησιμοποιήθηκε για την ανάπτυξη του δεύτερου εωραίου ματιού:

**Euk. 13.** python code for 2nd question

Output file: 2questionpython.txt



### Euk. 14. 2questionpython.txt

## Αναφορές

1. MongoDB Manual  
link:  
<https://docs.mongodb.com/manual/tutorial/query-documents/>
2. MongoDB Java Driver Quick Start  
link:  
<https://mongodb.github.io/mongo-java-driver/3.4/driver/getting-started/quick-start/>
3. MongoDB Java Servlet Web Application Example Tutorial  
link:  
<https://www.journaldev.com/4011/mongodb-java-servlet-web-application-example-tutorial>
4. Citation Network Dataset  
link:  
<https://www.aminer.org/citation>
5. MongoDB find document examples  
link:  
<https://howtodoinjava.com/mongodb/mongodb-find-documents/>
6. Python MongoDB  
link:  
[https://www.w3schools.com/python/python\\_mongodb\\_getstarted.asp](https://www.w3schools.com/python/python_mongodb_getstarted.asp)
7. Python and MongoDB  
link:  
<https://www.mongodb.com/python>
8. Introduction to working with MongoDB and PyMongo.  
link:  
<https://pymongo.readthedocs.io/en/stable/tutorial.html>
9. Web app example  
link:  
<https://www.javaguides.net/2019/03/registration-form-using-jsp-servlet-jdbc-mysql-example.html>
10. Colorlib  
link:  
<https://colorlib.com/wp/free-bootstrap-registration-forms/>