



POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Systems and Methods for Big and Unstructured Data Project

Author(s): **Simona Malegori**

Nicole Perrotta

Michele Simeone

Alberto Pirillo

Simone Tognocchi

Group Number: **38**

Academic Year: 2022-2023

Contents

Contents	i
1 First Delivery	1
1.1 Problem description	1
1.2 Hypothesis	1
1.3 Data	2
1.3.1 ER Diagram	2
1.3.2 Data set description	4
1.3.3 Data Pre-Processing	5
1.4 Neo4j	7
1.4.1 Data Upload	7
1.4.2 Graph Diagram	9
1.4.3 Queries	12
1.4.4 Creation/Update Commands	22
2 Second Delivery	25
2.1 Introduction	25
2.2 Data	25
2.2.1 Data Pre-Processing	28
2.2.2 Data Completion	28
2.3 MongoDB	30
2.3.1 Data Upload	30
2.3.2 Document Example	30
2.3.3 Join Operation	34
2.3.4 Queries	34
2.3.5 Creation/Update Commands	45
3 Third Delivery	55

3.1	Introduction	55
3.2	Data	55
3.2.1	Data Pre-Processing	55
3.3	Spark	57
3.3.1	Data Schema	57
3.3.2	Data Upload	58
3.3.3	Queries	59
3.3.4	Creation/Update Commands	66
4	References	69

1 | First Delivery

1.1. Problem description

This project aims at building an Information System that manages a data set containing different type of scientific articles that can be used for: clustering with network and side information, studying influence in the citation network, finding the most influential papers and topics, modeling analysis.

The project is divided in the following steps.

At first it was made an ER Diagram that generalizes all the information gathered from different already existing data sets, then the most complete data set was chosen.

Afterwards the data set was pre-processed, transforming it from a JSON to a CSV format and then it was reduced in size.

After that it was uploaded on Neo4j and all the nodes, the relationships and the properties were edited to build the Graph Diagram.

At the end of the project 10 queries and 6 creation/update commands were initiated with a different level of complexity that was checked within the performance time.

1.2. Hypothesis

In order to model the database we made some assumptions:

- authors can have one or more papers associated
- authors can have zero, one or more affiliated organizations, assuming that there can be authors that didn't provide their organization;
- papers have at least one author;
- papers have at least one field of study;
- not all papers have keywords associated, assuming that they may not have been provided;

- papers can reference and be referenced by other papers;
- each paper has a venue, that is the place where it has been published/presented;
- a venue can host more than one paper;
- venues are of 4 types: Journal, Conference, Book and Patent;
- the volume n of a paper is the n -th published collection;
- the issue m of a paper is the m -th part of the volume in which it is published.

1.3. Data

1.3.1. ER Diagram

The ER Diagram of the chosen model is characterized by the following entities with the respective attributes:

- **Paper** is a scientific article that is associated with the following attributes: id , $title$, $date$ that corresponds to the publication date, doi that is the Digital Object Identifier, $volume$ that corresponds to the n-th published collection, $issue$ that corresponds to the m-th part of the volume, $language$, $issn$ that is an identification code associated with the title of the publication, $isbn$ that is a code that identifies printed or digital papers and it is used as inventory-tracking device, $n_citation$ that is the number of citations, $page_start$ and $page_end$ that are the starting and the ending point of the collection from which the paper was extracted, pdf_url that is the source from which to recover the paper, $abstract$ that is the summary of the paper, $publisher$ and $external_url$ that corresponds to the sitography of the paper;
- **Author** with the attributes: id , $name$, $surname$, $email$, $orcid$ that is a unique and persistent identification number and $organization$;
- **Keyword** that represents a word that allows to define immediately the topic within the paper, its attributes are: id and $name$;
- **FoS** that corresponds to the field of studies with the attributes: id , $name$, w that is the weight of the fields of study;
- **Venue** that is the collection from which the paper was extracted, with the attributes: id and $name$. The venue of the paper can be of different types, in fact there is a total and exclusive generalization of the entity Venue that can be a:

- **Journal** that has also the attribute *addressee* that is the type of audience of the paper;
- **Book** that has also the attributes *category* and *edition*;
- **Conference** that has also the attributes *type* that can be physical and online, and *location* that has cardinality one only if the type is physical and it represents the place in which the conference takes place;
- **Patent** that has also the attributes *type* and *expiration*.

In the ER Diagram there are the following relationships with the respective cardinalities:

- **Writing** between the entities Paper and Author. The relationship means that a Paper can be written by at least 1 to a maximum of N authors and that an Author can write **at least 1 to a maximum of N papers**.
- **Containing** between the entities Paper and Keyword. The relationship means that a Paper can contain from 0 to N keywords and that a Keyword is contained into at least 1 to a maximum of N papers.
- **Dealing** between the entities Paper and FoS. The relationship means that a Paper can deal with at least 1 to a maximum of N field of studies and that a FoS can be dealt from 0 to N papers.
- **In** between the entities Paper and Venue. The relationship means that a Paper is extracted from exactly 1 venue and that a venue can be the collection in which at least one paper is contained.
- **Referencing** that is a relationship on the same entity of the Paper, in fact it contains the roles referencer and referenced. The relationship means that a Paper can or not reference other Papers and can be referenced or not from other Papers.

After all, in the ER Diagram there is an external constraint on the attributes of the entity Conference. In fact, the attribute *location* must have a cardinality of (1,1) if the type is *physical*.

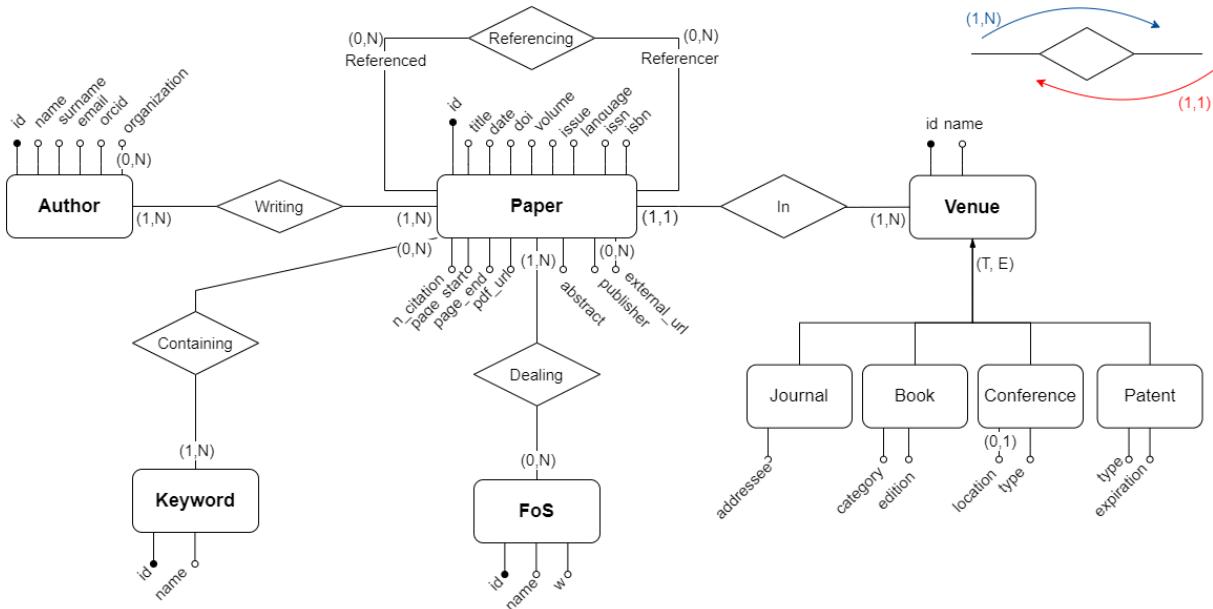


Figure 1.1: ER Diagram

errata corrigie:

- added cardinality also on the left side of the Referencing relationship
- changed cardinality of the left side of the Writing relationship from (0,N) to (1,N)

1.3.2. Data set description

The data set that was chosen for the goal of the project contains 8335 papers, 730 venues, 27576 authors and 25005 field of studies. It is a reduced version of the more complex model on which the ER Diagram is based. In fact, it doesn't contain the entities Keyword and the sub-entities Journal, Book, Conference and Patent that, instead, are transformed into an attribute of the entity Paper that is called doc_type. All the respective attributes of these entities are eliminated. Moreover, the attribute *date* of the Paper, that is generalized in the ER Diagram, becomes the year. Furthermore, in the reduced version of the data set, the relationship Referencing becomes a list of string that are the references of the Paper. Lastly, the chosen data set contains just the more significant attributes. Summarizing, the chosen data set is composed by the following entities and attributes:

- **Paper** with the attributes: id (integer type), title (string type), year (integer type), n_citation (integer type), page_start (integer type), page_end (integer type), doc_type (string type), publisher (string type), volume (integer type), issue (integer type), doi (string type), references (list of string type), abstract (string type);

- **Author** with the attributes: paper_id (integer type), author_name (string type), author_id (integer type), author_org (string type);
- **FoS** with the attributes: paper_id (integer type), fos_name (string type), fos_weight (float type);
- **Venue** with the attributes: paper_id (integer type), venue_name (string type), venue_id (integer type).

1.3.3. Data Pre-Processing

The original data set can be downloaded [here](#). Given that such data set was unnecessarily large for our purpose, we decided to reduce its size. We could have just cut it at a certain point, however we decided that it was better to work with consistent data, therefore we decided to carefully perform sub-sampling in an intelligent way. Such pre-processing was performed using multiple Python scripts and the Pandas library.

Here we provide a description of all the scripts used, together with the procedure to obtain the final data set starting from the initial one. All of these scripts can be found in the *scripts* folder of the *neo4j* section of the *GitHub repository* of the project.

The notebook **dataset_exploration.ipynb** contains a short description of every operation for explanatory reasons. This notebook processes only a small chunk of the data set. The same operations are performed on the whole data set in the script **dataset_preprocessing.py**.

Here is a summary of the operations performed:

- Removal of samples with Null and NaN values
- Removal of samples with an empty string in a field

The operations above are required to work with consistent data. Notice that we can afford to simply drop the samples that do not respect such conditions since we dispose of a very large data set.

The following operations are not required but were performed to reduce even further the size of the data set, with the objective of keeping only the "most important" samples.

We kept the samples:

- With a number of citations greater than a threshold
- With a reference count greater than a threshold

We converted the *indexed_abstract* field from an inverted index to a string of text, to make it easier to query once inserted into the database. The field was also renamed to *abstract*.

We also processed the data set in order to remove some special characters not supported by the import function of the database. An example of such characters is "\\". This operation is performed in the script **remove_special_characters.py**.

Given that many samples were removed from the data set, we also had to fix up the *references* field to keep only the valid ones. We consider a reference valid when it points to a sample of the data set that was kept. Otherwise, we say that such reference is invalid and we remove it from the data set. This operation is performed in the script **deprecated_references.py**.

Lastly, when importing the data into the database, we realized that it was more practical to split the data set into multiple data sets to speed up the process and to produce cleaner code. This functionality was added in the **dataset_preprocessing.py** script. We split the data set following the structure of ER diagram, ending up with one separate data set for each entity present in the original data set. Thus, we ended up with 4 data sets: **Paper**, **Author**, **Venue** and **FoS**.

To obtain the final data set starting from the downloaded one, run the scripts in this order:

1. **dataset_preprocessing.py**
2. **deprecated_references.py**
3. **remove_special_characters.py**

The input of the first script is the initial data set. Only the paper data set requires to be processed by the second script, then only the paper and the venue data sets require to be processed by the third script. At the end, you will obtain 4 data sets which are identical to the ones that we imported into the database.

1.4. Neo4j

1.4.1. Data Upload

To import the data into Neo4j, there is one last precaution needed: it is necessary to process the Paper data set to make the *references* field compatible with the import function of the database. To perform such action the script **preprocessing.py** can be used. The script is located in the *neo4j* folder of the *GitHub repository*. Now, the data sets can be correctly imported into the Neo4J database. In order to do that, you need to put the four data sets files in the program's import folder.

Thanks to the pre-processing that was performed, the four data sets are saved in the simple CSV format and it is possible to use the LOAD CSV command to easily load all the entities and relationships.

1. Clear the Database:

```
1 MATCH (x) DETACH DELETE x;
```

2. Load the Papers:

```
1 LOAD CSV WITH HEADERS FROM "file:///paper_dataset2.csv" AS csvLine
2 CREATE (p:Paper {id: toInteger(csvLine.id), title: csvLine.title,
    year: toInteger(csvLine.year), doi: csvLine.doi, volume: csvLine
    .volume, issue: csvLine.issue, abstract:csvLine.abstract,
    n_citation:toInteger(csvLine.n_citation), page_start:toInteger(
    csvLine.page_start), page_end:toInteger(csvLine.page_end),
    publisher:csvLine.publisher, doc_type:csvLine.doc_type})
```

3. Load the Authors:

```
1 LOAD CSV WITH HEADERS FROM "file:///author_dataset.csv" AS csvLine
2 CREATE (a:Author {id: toInteger(csvLine.author_id), name: csvLine.
    author_name, organization: csvLine.author_org})
```

4. Load the FoS:

```
1 LOAD CSV WITH HEADERS FROM "file:///fos_dataset_0.csv" AS csvLine
2 CREATE (f:Fos {weight: toFloat(csvLine.fos_weight), name: csvLine.
    fos_name})
```

5. Load the Venue (*errata corrige: CREATE instead of LOAD at line 2*):

```
1 LOAD CSV WITH HEADERS FROM "file:///venue_dataset.csv" AS csvLine
2 CREATE (v:Venue {id: toInteger(csvLine.venue_id), name: csvLine.
    venue_name})
```

6. Create the Writing relationship between papers and authors:

```

1 LOAD CSV WITH HEADERS FROM "file:///author_dataset.csv" AS csvLine
2 MATCH (p:Paper),(a:Author)
3 WHERE (toInteger(csvLine.paper_id)=p.id) AND (toInteger(csvLine.
    author_id) = a.id)
4 CREATE (p)-[w:writing]->(a)

```

7. Create the Referring relationship between papers and papers:

```

1 LOAD CSV WITH HEADERS FROM "file:///paper_dataset2.csv" AS csvLine
2 UNWIND split(csvLine.references, ':') as ref
3 MATCH (p1:Paper),(p2:Paper)
4 WHERE (toInteger(csvLine.id)=p1.id) AND (toInteger(ref)=p2.id)
5 CREATE (p1)-[r:referencing]->(p2)

```

8. Create the in relationship between papers and venues:

```

1 LOAD CSV WITH HEADERS FROM "file:///venue_dataset.csv" AS csvLine
2 MATCH (p:Paper),(v:Venue)
3 WHERE (toInteger(csvLine.paper_id)=p.id) AND (toInteger(csvLine.
    venue_id) = v.id)
4 CREATE (p)-[i:in_]->(v)

```

9. Create the dealing relationship between papers and FoS:

```

1 LOAD CSV WITH HEADERS FROM "file:///fos_dataset_0.csv" AS csvLine
2 MATCH (p:Paper),(f:Fos)
3 WHERE (toInteger(csvLine.paper_id)=p.id) AND (csvLine.fos_name = f.
    name) AND (toFloat(csvLine.fos_weight) = f.weight)
4 CREATE (p)-[d:dealing]->(f)

```

1.4.2. Graph Diagram

Entities:

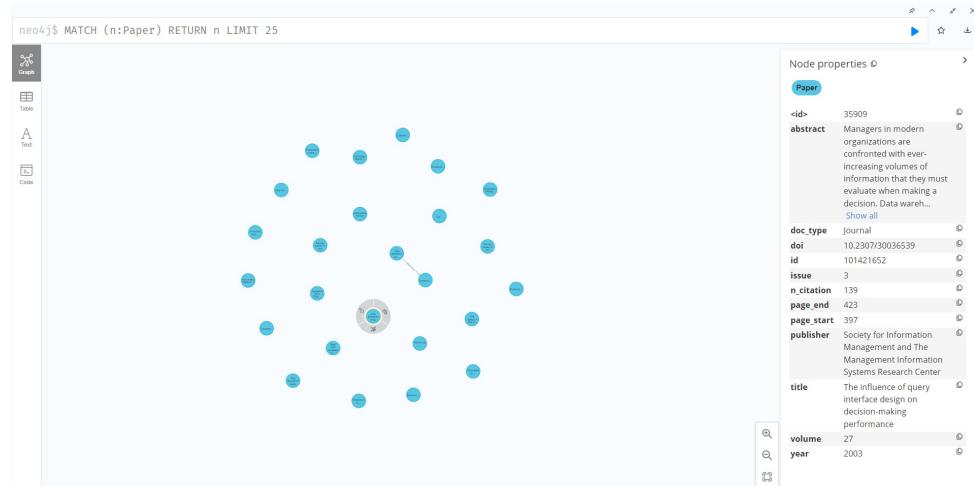


Figure 1.2: Paper

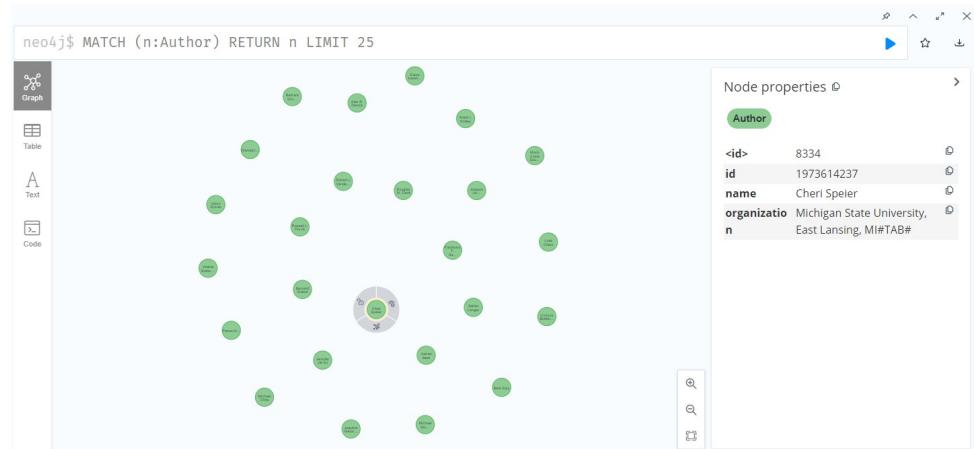


Figure 1.3: Author



Figure 1.4: FoS

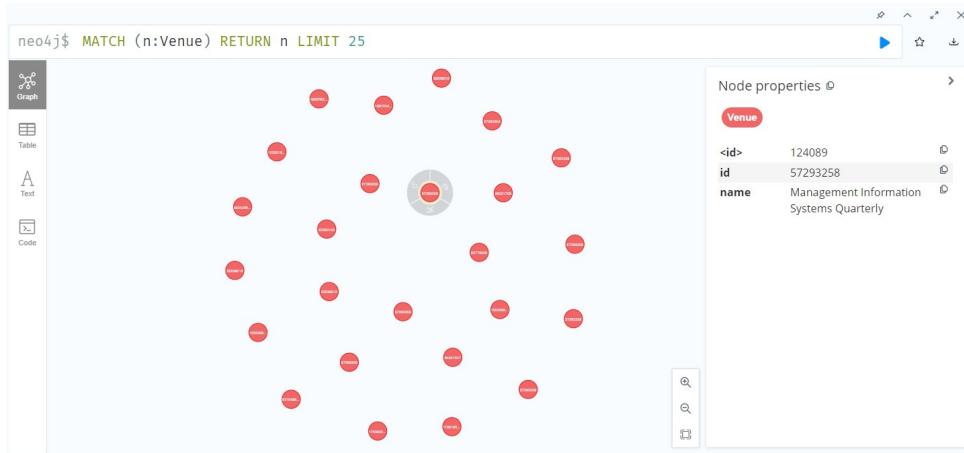


Figure 1.5: Venue

Relationships:

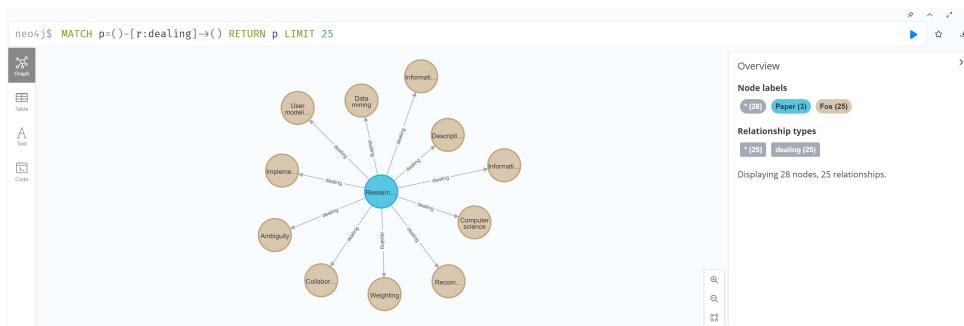


Figure 1.6: Paper-Dealing->FoS

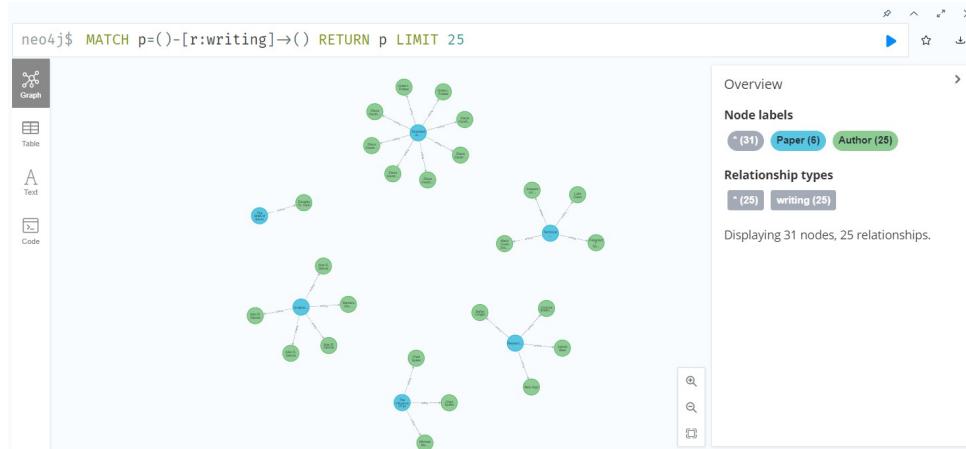


Figure 1.7: Paper-Writing->Author

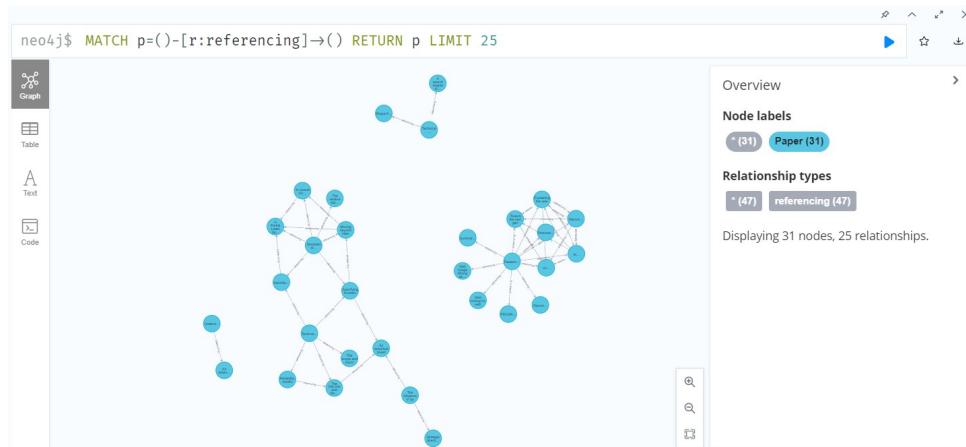


Figure 1.8: Paper-Referencing->Paper

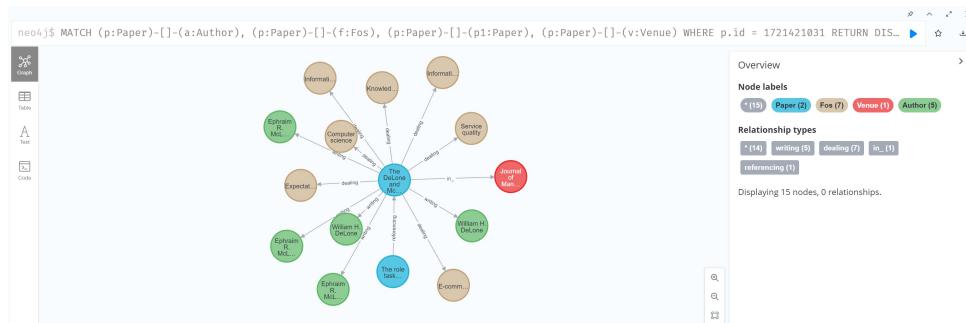


Figure 1.9: All the relationships

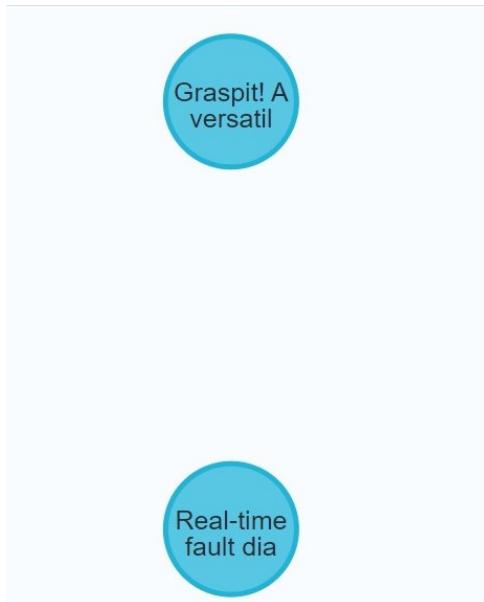
1.4.3. Queries

- Find papers of a determined venue, written after a certain date (Execution time 454ms):

```

1 MATCH (p: Paper)-[i:in_]->(v: Venue), (p: Paper)-[d:dealing]->(f:
  Fos)
2 WHERE (p.year > 2000) AND (v.name="IEEE Robotics & Automation
  Magazine")
3 RETURN p

```

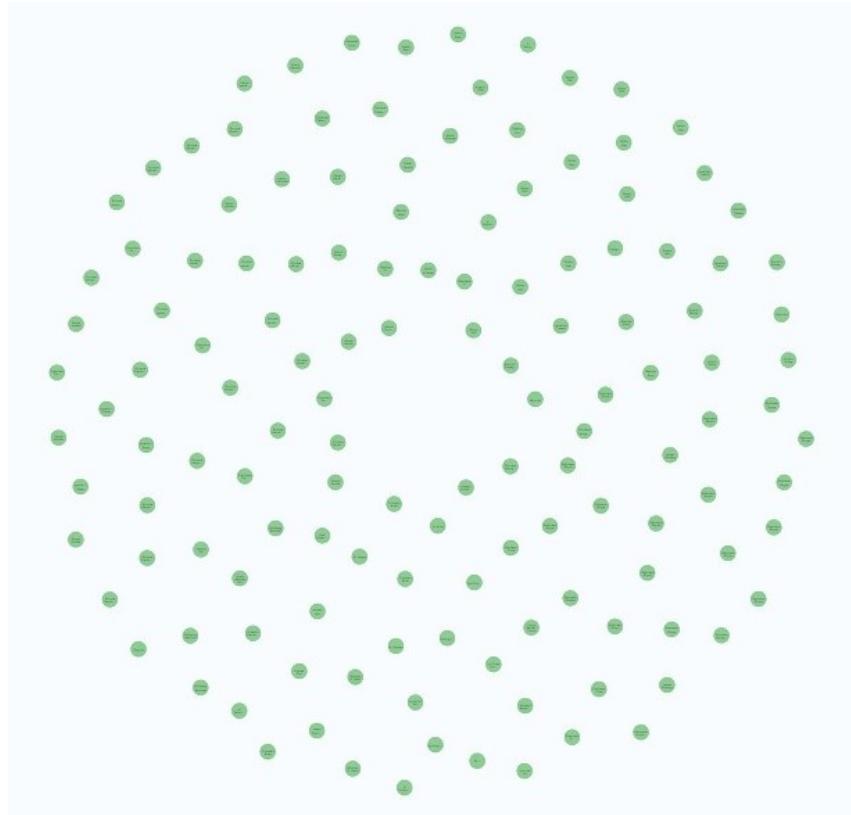


- Find **authors that wrote papers** in a set of venues, with a determined type (Execution time 2ms):

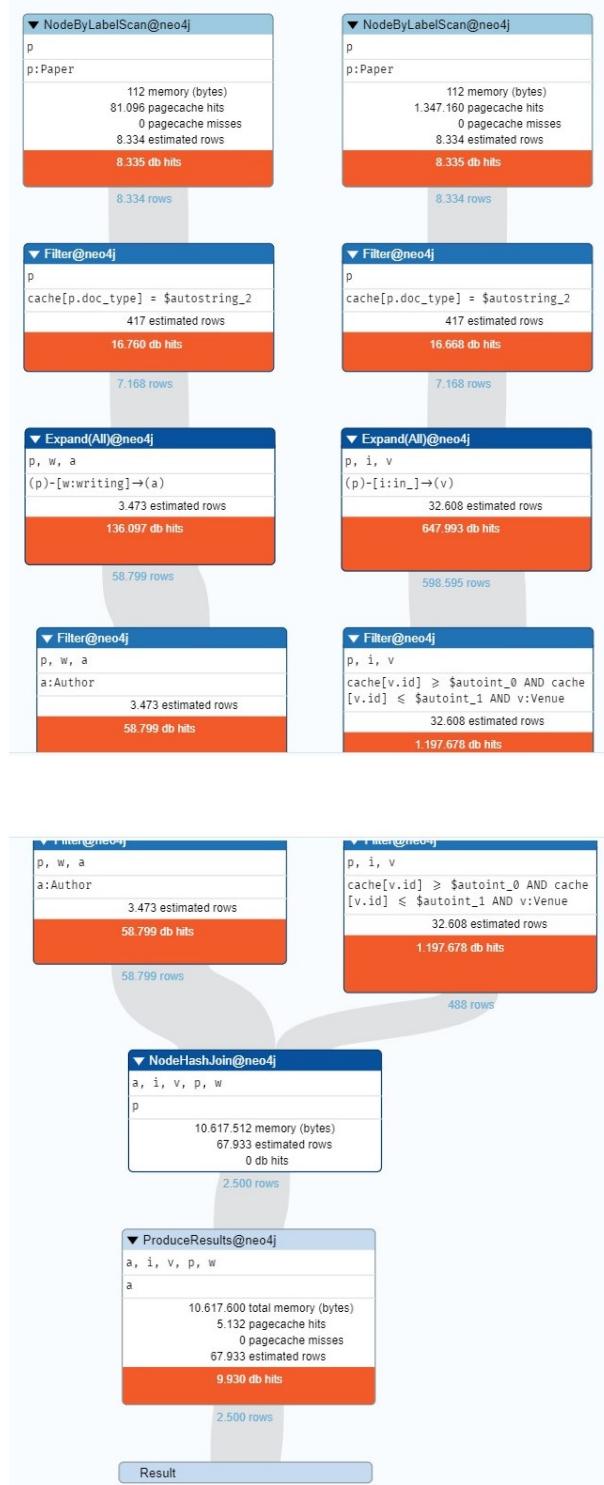
```

1 MATCH (p: Paper) -[w:writing]->(a:Author), (p: Paper) -[i: in_]->
  (v:Venue)
2 WHERE v.id >= 140000000 and v.id <= 140900000 AND p.doc_type='
  Journal'
3 RETURN a

```



Below we can see the output of the profile statement, that is used to track the query and the numbers of rows of each operation. It starts by scanning all the nodes with the papers, then it expands all the nodes with the 'writing' relationship on the left of the picture and the 'in' relationship on the right. Finally, it applies the filters on the venue id and on the paper's doc type and returns the results.



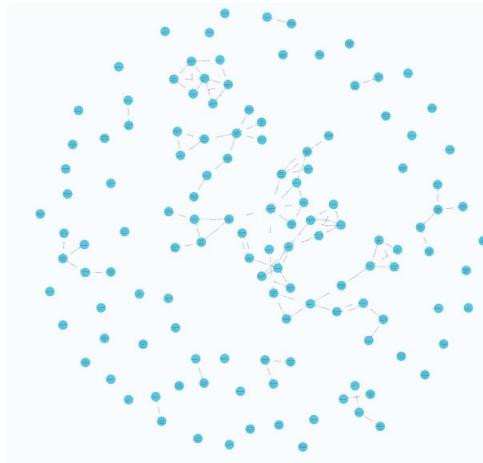
3. Find papers of a determined main argument, written after a certain date (Execution time 30ms):

```
1 MATCH (p: Paper)-[r:referencing]->(p2: Paper), (p2: Paper)-[d: dealing]->(f: Fos)
```

```

2 WHERE (f.weight >= 0.45) AND (p2.year >2010) AND (f.name="Artificial intelligence")
3 RETURN p

```



4. Count the authors that have written a famous paper of a determined argument (Execution time 38ms):

```

1 MATCH (p: Paper)-[w:writing]->(a:Author), (p: Paper)-[d: dealing]->(f: Fos)
2 WHERE f.name = "Machine learning" AND p.n_citation>5000
3 RETURN COUNT(a.id) AS num_aut

```

num_aut
1 2120

5. Count the papers divided per author written in a set of venue by a determined authors' organization (Execution time 103ms):

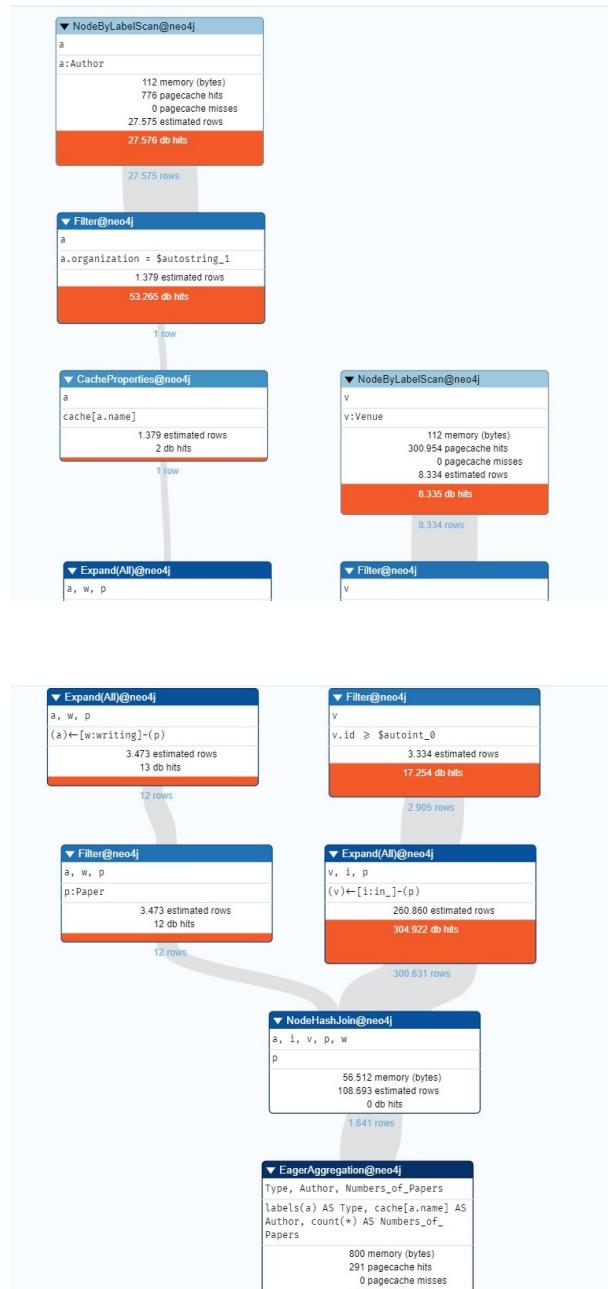
```

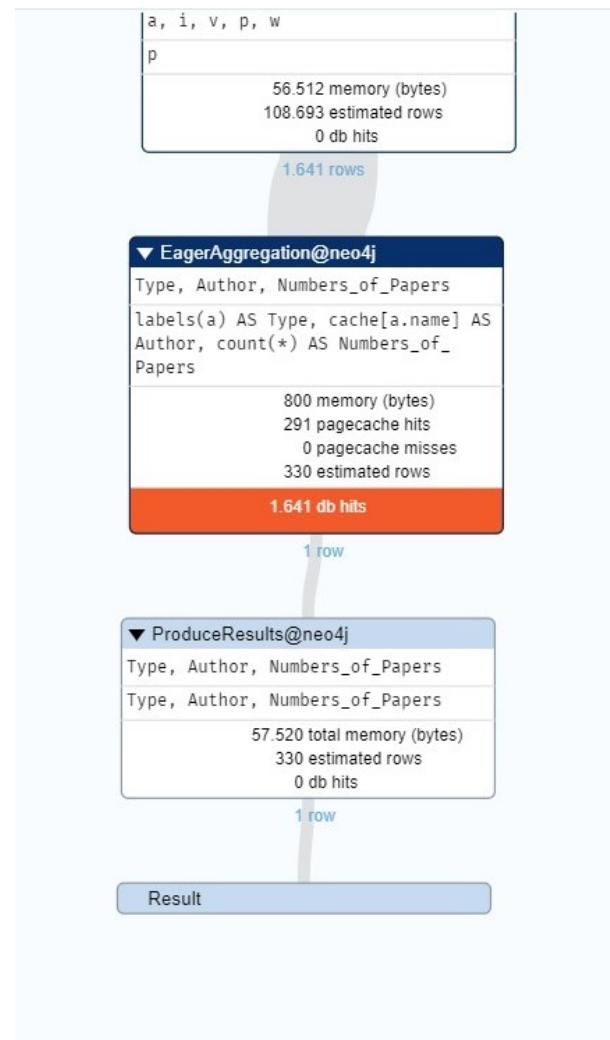
1 MATCH (p: Paper) -[w:writing]->(a:Author), (p: Paper) -[i: in_]->(v : Venue)
2 WHERE v.id >= 160000000 AND a.organization='Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA. hshum@microsoft.com#TAB#
'
3 RETURN labels(a) AS Type, a.name AS Author, COUNT(*) AS Numbers_of_Papers

```

Type	Author	Numbers_of_Papers
["Author"]	"Heung-Yeung Shum"	1641

The profile statement scans all the authors, then it expands the 'writing' relationship and applies the organization's filter. At the same time it scans all the venues and filters their id. At the end the aggregation is applied and all the resulting values are returned.





6. Calculate the weight keywords' average of papers that have a determined publisher and a reference to a famous paper (Execution time 27ms):

```

1 MATCH (p: Paper) -[d: dealing]->(f: Fos), (p: Paper) -[r:
  referencing]->(p2: Paper)
2 WHERE p2.n_citation>2000 AND p.publisher='Society for Information
  Management and The Management Information Systems Research
  Center'
3 RETURN p.title, avg(f.weight)

```

p.title	avg(f.weight)
"Shackled to the status quo: the inhibiting effects of incumbent system habit, switching costs, and inertia on new system acceptance"	0.449569698283237
"Technostress: technological antecedents and implications"	0.44386502606134975
"Business intelligence in blogs: understanding consumer interactions and communities"	0.41906192304436024
"Web and wireless site usability: understanding differences and modeling use"	0.42455533500106846
"Competing perspectives on the link between strategic information technology alignment and organizational agility: insights from a mediation model"	0.4178663593780648
"Reliability, mindfulness, and information systems"	0.42862394017677813

7. Count the numbers of bilateral reference between two papers that have the same venue and a large number of citation (Execution time 119ms):

```

1 MATCH (p: Paper)-[r:referencing]->(p2: Paper), (p2: Paper)-[r2:
    referencing]->(p: Paper), (p: Paper)-[w:writing]->(a: Author),(p
    : Paper)-[i:in_]->(v: Venue), (p2: Paper)-[i2:in_]->(v2: Venue)
2 WHERE p2.n_citation>500 AND p.n_citation>500 AND v.name=v2.name AND
      a.organization="Royal Institute of Technology"
3 RETURN type(r) AS Relation, COUNT(*)/2 AS
      Num_of_bilateral_referencingsame_venue, a.organization AS
      organization
4 LIMIT 1

```

Relation	Num_of_bilateral_referencing_same_venue	organization
"referencing"	28900	"Royal Institute of Technology"

8. Count the total citation of a paper that have references to two papers that deal of different field of study (Execution time 1849ms):

```

1 MATCH (p: Paper)-[r: referencing]->(p2: Paper), (p: Paper)-[r2:
    referencing]->(p3: Paper), (p2: Paper)-[d: dealing]->(f: Fos), (
    p3: Paper)-[d2: dealing]->(f2: Fos)
2 WHERE p.page_end-p.page_start>10 AND p2 <> p3 AND f.name='
    Artificial Intelligence' AND f.weight>0.0 and f2.name = 'Machine
    learning' AND f2.weight>0.0
3 RETURN p.title AS Title, SUM(p.n_citation) AS Sum_n_citation
4 LIMIT 5

```

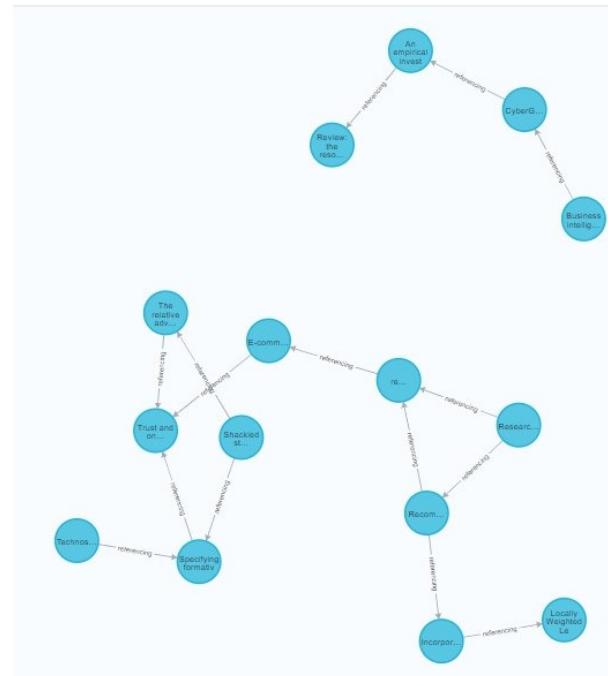
Title	Sum_n_citation
1 "The SUN Attribute Database: Beyond Categories for Deeper Scene Understanding"	890400
2 "Composition in Distributional Models of Semantics"	4053440
3 "Wikipedia-based semantic interpretation for natural language processing"	2170880
4 "Knowledge derived from wikipedia for computing semantic relatedness"	737760
5 "A survey of paraphrasing and textual entailment methods"	4324800

9. Find the shortest path between two different paper that have a reference in common where the first is less famous than the second one (Execution time 61ms):

```

1 MATCH s = shortestPath(
2   (p: Paper)-[r:referencing*]->(p2: Paper)
3 )
4 WHERE p2.n_citation > 10*p.n_citation AND p <> p2
5 RETURN s
6 LIMIT 5

```

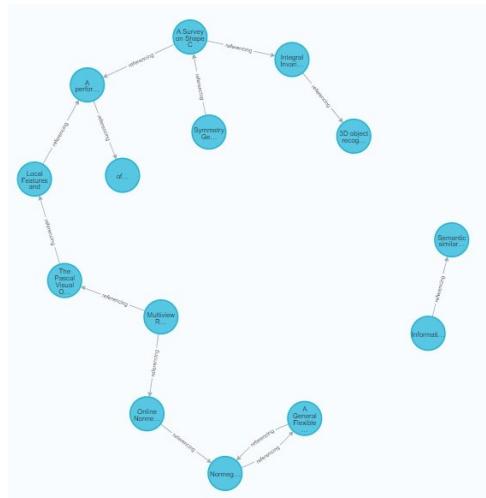


10. Find the shortest path between two different paper that have a reference in common and different FoS (Execution time 42ms):

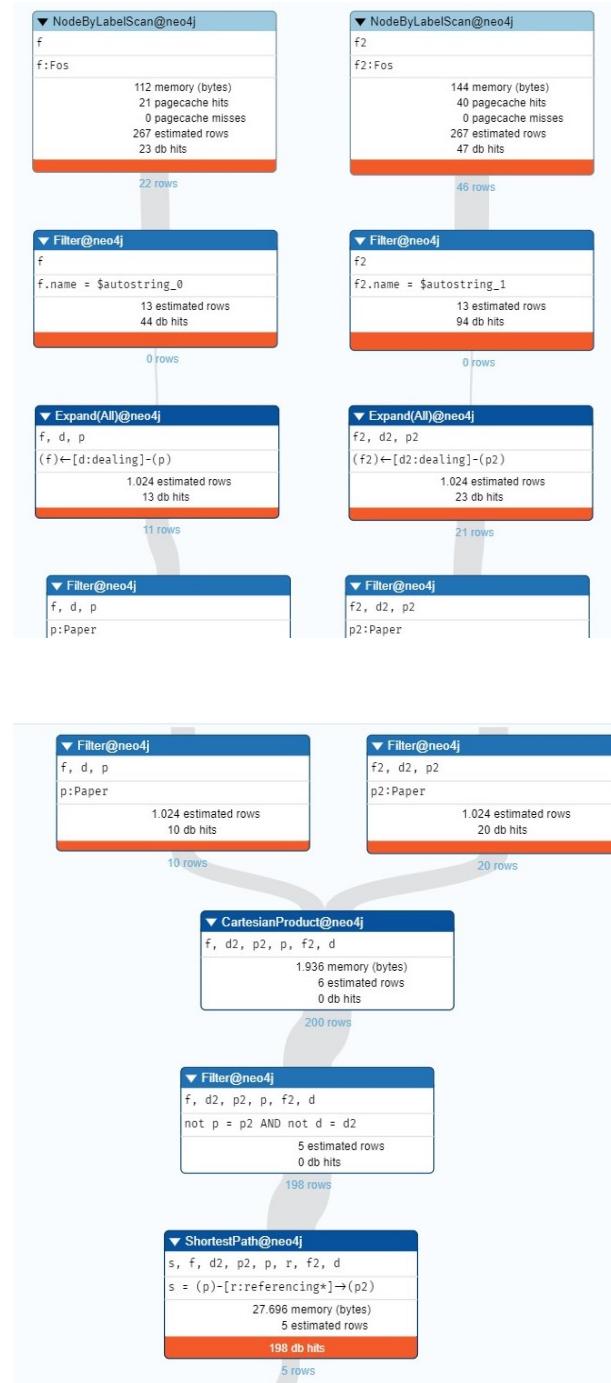
```

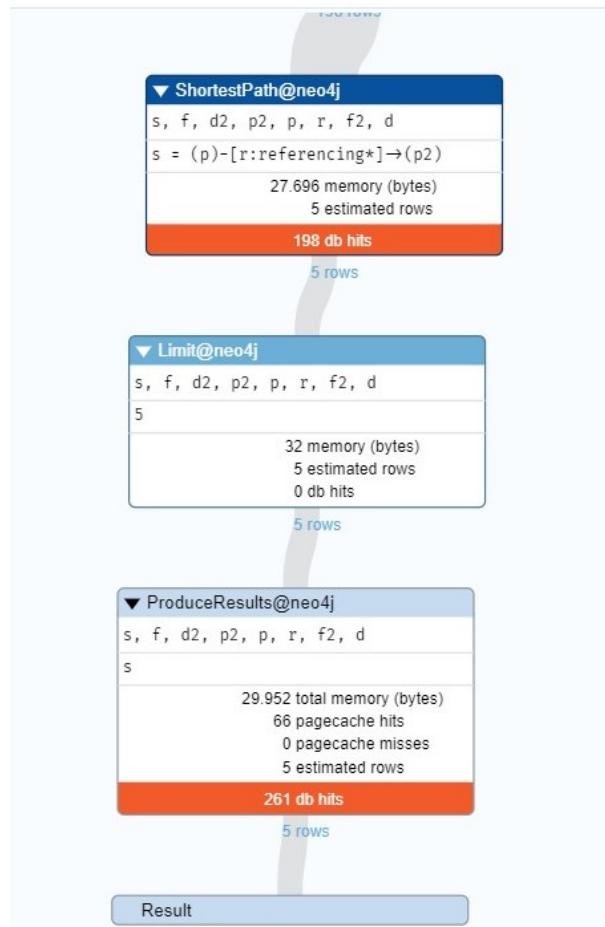
1 MATCH (p: Paper)-[d: dealing]->(f:Fos), (p2: Paper)-[d2:dealing]
      ->(f2:Fos), s = shortestPath( (p: Paper)-[r:referencing*]->(p2:
      Paper) )
2 WHERE f.name = "Artificial intelligence" AND f2.name = "Machine
      learning" AND p <> p2
3 RETURN s
4 LIMIT 5

```



In the following page we can see that the profile statement scans the FoS, applies the filters to them and expands the 'dealing' relationship. Then it checks that paper1 and paper2 are different and in the end it calls the function 'shortestPath' and returns the results.





1.4.4. Creation/Update Commands

1. Insertion of a new author in the database:

```
1 CREATE (a:Author {id: 3000000000, name:"Mario Rossi", organization:
  "Politecnico di Milano"})
```

2. Insertion of a new field of study in the database:

```
1 CREATE (f:Fos {weight:0.57640203, name:"Ambient intelligence" })
```

3. Insertion of a new venue in the database:

```
1 CREATE (v:Venue { id:3000000001, name:"Journal of Cloud Computing"
  })
```

4. Insertion of a new paper in the database, paper's relationships with authors, venue, fields of study and other papers are created:

```
1 MATCH (a:Author) WHERE a.id=3000000000
2 MATCH (f:Fos) WHERE f.name="Cloud computing" AND f.weight=0.6410412
```

```
3 MATCH (v:Venue) WHERE v.id=3000000001
4 MATCH (ref:Paper) WHERE ref.doi="10.1109/JPROC.2015.2483592"
5 MATCH (cit:Paper) WHERE cit.doi="10.1007/s10845-008-0158-5"
6 CREATE (p:Paper { id: 3000000000, title: "Application of
    deterministic, stochastic, and hybrid methods for cloud provider
    selection", year: 2022, doi: "10.1186/s13677-021-00275-1",
    volume: "11" , issue: "1" , abstract: "Cloud Computing
    popularization inspired ... requests.", n_citation: 5,
    page_start: 5, page_end: 10, publisher: "SpringerOpen", doc_type
    : "Journal" })
7 CREATE (p)-[w:writing]->(a), (p)-[r:referencing]->(ref), (cit)-[r:
    referencing]->(p), (p)-[i:in_]->(v), (p)-[d:dealing]->(f)
```

5. Update of the organization of an author:

```
1 MATCH (a:Author {name: 'Stefano Ceri'})
2 SET a.organization = "Politecnico di Milano"
```

6. Update of the number of citations of a paper:

```
1 MATCH (p:Paper {title: "Markov localization for mobile robots in
    dynamic environments"})
2 SET p.n_citation = 728
```


2 | Second Delivery

2.1. Introduction

This project aims at building an Information System that manages a data set containing different type of scientific articles that can be used for: clustering with network and side information, studying influence in the citation network, finding the most influential papers and topics, modeling analysis.

For the same problem described in Chapter 1 (First Delivery), we modeled a different system following the subsequent steps.

At first the data set was pre-processed to add the fields that were missing in the original data set. Such fields were taken from different Kaggle data sets. Then it was reduced in size and after that it was uploaded on MongoDB.

At the end of the project 11 queries and 6 creation/update commands were initiated with a different level of complexity that was checked within the performance time.

2.2. Data

This is the structure of a paper inside the database, it contains different sub-documents: authors, fos (metadata), venue and sections.

```

1 {
2   "id": "...",
3   "title": "...",
4   "abstract": "...",
5   "authors": [
6     {
7       "name": "...",
8       "org": "...",
9       "email": "...",
10      "bio": "..."
11    }
12  ],
13 }
```

```

11 "fos": [
12   "name": "...",
13   "weight": ...
14 ],
15 "year": ...,
16 "page_start": ...,
17 "page_end": ...,
18 "doc_type": "...",
19 "publisher": "...",
20 "volume": ...,
21 "issue": ...,
22 "doi": "...",
23 "n_citation": ...,
24 "venue": {
25   "raw": "...",
26   "id": "...",
27 },
28 "references": [ ... ],
29 "sections": [
30   {
31     "id": ...,
32     "title": "...",
33     "text": "...",
34     "subsections": [
35       {
36         "id": ...,
37         "title": "...",
38         "text": ...
39       }
40     ],
41     "figures": [
42       {
43         "url": "...",
44         "caption": ...
45       }
46     ]
47   }
48 ]
49 }

```

This is the description of a paper:

Paper is a scientific article that is associated with the following fields:

- *id*: Int;
- *title*: String;
- *abstract*: that is the summary of the paper: String;

- *authors*:
 - *name*: String;
 - *organization*: String;
 - *email*: that is a personal contact of the author: String;
 - *bio*: a short introduction written by the author (it is added from other sources): String;
- *FoS (Field of Study)*:
 - *name*: String;
 - *w*: that is the weight of the fields of study: Float;
- *Publication Details*:
 - *year*: that corresponds to the publication year: Int;
 - *page_start* and *page_end*: that are the starting and the ending point of the collection from which the paper was extracted: Int;
 - *doc_type*: that is the type of the paper, and can assume 3 values: *Journal, Conference, Patent*: String;
 - *publisher*: String;
 - *volume*: that corresponds to the n-th published collection: Int;
 - *issue*: that corresponds to the m-th part of the volume: Int;
 - *doi*: that is the Digital Object Identifier: String;
 - *n_citation*: that is the number of citations: Int;
- *Venue*: that is the collection from which the paper was extracted, with the attributes:
 - *id*: String;
 - *raw*: that is the name of the collection: String;
- *References*:
 - *id*: String;

- *Sections*: every section has an id, a title and a content (textual), it could have some subsections too (with the same structure of a section). Furthermore every section has one or more images. Every image has an url and a caption:
 - *id*: Int;
 - *title*: String;
 - *text*: String;
 - *subsections*:
 - * *id*: Int;
 - * *title*: Int;
 - * *text*: String;
 - *figures*:
 - * *url*: String;
 - * *caption*: String;

2.2.1. Data Pre-Processing

In this part of the project we used as a starting point the same data set used in the first delivery, which can be downloaded [here](#). However, the pre-processing this time was slightly different. Roughly the same operations of the first delivery were performed to clean and reduce the size of the data set. Those operations are performed with the script **dataset_preprocessing.py**, which can be found in the *scripts* folder of the *mongodb* section of the repository. It is worth to notice that such script does not split the original data set into multiple ones like in the first delivery. Instead, a single data set is kept.

2.2.2. Data Completion

We decided to avoid working with synthetic data (i.e. generate random meaningless data) and instead we decided to add to our dataset the missing fields by retrieving data from other data sets, trying to be as accurate as possible. The missing fields w.r.t. the description in section 2.2 were:

- *authors.email*
- *authors.bio*

- *sections*

We generated the email of every author and we inserted their bio, that were taken from Kaggle's Goodread-Authors data set. This is performed with the script **update_author.py**.

Furthermore we added the *Sections* part, using a Twitter dataset, that contains millions of tweets in different languages. For every paper we generated randomly a number of sections in the range from 1 to 3 and for each one from 0 to 2 subsections. For the title we used the text of one tweet and for the content the text of four tweets.

Then for the figures we used another data set (Train-GCC-training.tsv from the Google Conceptual Captions official website) in order to take the *image.url* and the *image.caption* fields. Every section contains a random number of figures between 1 and 3.

Those two operations above are executed in the notebook **sections_preprocessing.py**, that can be found in the *mongodb* section of the GitHub repository.

The final data set consists of 8333 documents.

To obtain the final data set starting from the downloaded one, run the scripts in this order:

1. **dataset_preprocessing.py**
2. **deprecated_references.py**
3. **bio_preprocessing.py**
4. **update_author.py**
5. **sections_preprocessing.py**

2.3. MongoDB

2.3.1. Data Upload

To import the data into a MongoDB collection, we rely on another Python script that exploits the PyMongo library, the official MongoDB driver for Python. We used the script `import_data.py`, that can be found in the `scripts` folder of the `mongodb` section of the GitHub repository. The script dumps the whole data set into a MongoDB collection, while preserving its complex and nested structure. It also performs some pre-processing on the `id` and on the `references` fields:

- The `id` field is renamed to `_id` in order to be used as an index inside of the database
- Both the `_id` field and the `references` field are converted to an ObjectId.

Once the correct connection parameters and the correct file path have been specified, it is enough to run the script with Python to import the data. We will use `papers` as the name of the collection.

2.3.2. Document Example

Below we can see the general structure of a document, with the help of MongoDB Compass.

```
_id: ObjectId('101421652000000000000000')
title: "The influence of query interface design on decision-making performance"
> authors: Array
> venue: Object
year: 2003
n_citation: 139
page_start: 397
page_end: 423
doc_type: "Journal"
publisher: "Society for Information Management and The Management Information Syst..."
volume: 27
issue: 3
> flos: Array
doi: "10.2307/30036539"
> references: Array
abstract: "Managers in modern organizations are confronted with ever-increasing v..."
> sections: Array
```

Figure 2.1: General structure

We will now expand the structure of the fields with a complex type.

The *authors* field is an array of sub-documents in which every element represents an author that worked on the paper.

```

    ✓ authors: Array
      ✓ 0: Object
        name: "Luke Olsen"
        id: "2142664686"
        org: "Department of Computer Science, University of Calgary, Calgary, AB, Ca..."
        email: "lukeolsen@mit.edu"
        bio: "Lucie Dufresne was born in 1951 in Trois-Rivières between two rivers: ..."
      ✓ 1: Object
        name: "Faramarz F. Samavati"
        id: "1821145341"
        org: "Department of Computer Science, University of Calgary, Calgary, AB, Ca..."
        email: "faramarzf.samavati@hotmail.com"
        bio: "Ramón González Férriz es editor y periodista. Actualmente, es columnis..."
      ✓ 2: Object
        name: "Mario Costa Sousa"
        id: "2105740368"
        org: "Department of Computer Science, University of Calgary, Calgary, AB, Ca..."
        email: "mariocostasousa@mit.edu"
        bio: "Maja Ilisch, geboren 1975 in Dortmund, studierte Öffentliches Biblioth..."
      ✓ 3: Object
        name: "Joaquim A. Jorge"
        id: "2120678171"
        org: "Departamento de Engenharia Informática, Instituto Superior Técnico, Li..."
        email: "joaquima.jorge@gmail.com"
        bio: "Author of TINCTURE (<a target=_blank href="http://www.tincturestory..."
```

Figure 2.2: Structure of the *authors* field

The *venue* field is a sub-document with two internal fields.

```

    ✓ venue: Object
      raw: "Computers & Graphics"
      id: "94821547"
```

Figure 2.3: Structure of the *venue* field

The *fos* field is another array of sub-documents. Each element corresponds to one of the topics covered by the paper.

```

    ↘ fos: Array
      ↘ 0: Object
        name: "Computer vision"
        w: 0.427553326
      ↘ 1: Object
        name: "Artificial intelligence"
        w: 0
      ↘ 2: Object
        name: "Sketch recognition"
        w: 0.6052215
      ↘ 3: Object
        name: "Geometric modeling"
        w: 0.48624804600000004
      ↘ 4: Object
        name: "Human-computer interaction"
        w: 0.4427773
  
```

Figure 2.4: Structure of the *fos* field

The *references* field is implemented just as an array of ObjectIds.

```

    ↘ references: Array
      0: ObjectId('175816726800000000000000')
      1: ObjectId('191857022600000000000000')
      2: ObjectId('209868554100000000000000')
      3: ObjectId('211598118400000000000000')
      4: ObjectId('212745062100000000000000')
      5: ObjectId('213274786700000000000000')
  
```

Figure 2.5: Structure of the *references* field

Lastly, the *sections* field is an array of sub-documents. Each element of such array corresponds to one section of the paper. Every section can contain another array of sub-documents in the *subsections* field and another array of sub-documents in the *figures* field.

```

✓ sections: Array
  ✓ 0: Object
    id: 1
    title: "İlk önce kendi gücünün bilgisine sahip olmalısın; ikincisi, meydan oku..."
    text: "Soyumuz soylansın, Boyumuz Boylansın.. Al Yıldızlı bayrağımız KUDÜSTE,...""
  ✓ subsections: Array
    ✓ figures: Array
      ✓ 0: Object
        url: "http://lh6.ggpht.com/-IvRtNLNcG8o/TpFyrudaT6I/AAAAAAAAM6o/_11MuAAKalo/..."
        caption: "a very typical bus station    pop artist attends the 3rd annual at que..."
      ✓ 1: Object
        url: "http://lh6.ggpht.com/-IvRtNLNcG8o/TpFyrudaT6I/AAAAAAAAM6o/_11MuAAKalo/..."
        caption: "a very typical bus station    illustration of a map , its flag and a c..."
    ✓ 1: Object
      id: 2
      title: "Gurbete düşmüş bir insan, ne denli varlık içinde bir yaşam sürüyor olsam..."
      text: "Türk milleti vatanını koruyan ordumuzun yanındadır. #BarışPinarıViyadü...""
  ✓ subsections: Array
    ✓ 0: Object
      id: 1
      title: "Hep öлerek çоgaldık... Biz Oğuzun erleri. #DavanınGüçü"
      text: "Bizim milliyetçiliğimiz ayırcı değil birleştirici, çatışmacı değil ba..."
    ✓ 1: Object
      id: 2
      title: ""YPG silah ve malzemeleri bırakıp çekilsin, bu gece harekatı durduralım."
      text: "RT @Enesovvic: Ya siz kimi kimin toprağından kovuyorsunuz? Burası biz..."
  ✓ figures: Array
    ✓ 0: Object
      url: "http://lh6.ggpht.com/-IvRtNLNcG8o/TpFyrudaT6I/AAAAAAAAM6o/_11MuAAKalo/..."
      caption: "a very typical bus station    rock artist performs on stage at awards ..."

```

Figure 2.6: Structure of the *sections* field

2.3.3. Join Operation

To get a temporary collection in which every document also contains an array with inside all the other documents it references, we can use the `$lookup` operator.

```

1 db.papers.aggregate([
2   {
3     $lookup:
4       {
5         from: "papers",
6         localField: "references",
7         foreignField: "_id",
8         as: "refs"
9       }
10    }
11  ])

```

The referenced documents will be contained in the `refs` field.

2.3.4. Queries

In this section queries with a different level of complexity are presented, with a brief description and a figure that shows their results. Notice that, for some queries, that return several documents, the result is only partially shown.

1. Find papers of 2006 with issue equal to 3 and volume greater or equal to 5. Then show just the list of authors with their name and organization, and also the venue of the paper. Limit the result to 2:

```

1 db.papers.find(
2   {"year": 2006, "volume": {"$gte": 5}, "issue": 3},
3   {"authors.name":1, "authors.org":1, "venue":1}
4 ).limit(2)

```

(nReturned: 2, executionTimeMillis: 57, totalDocsExamined: 382)

```

< { _id: ObjectId("168383441500000000000000"),
  authors:
    [ { name: 'A.D. Murugan',
        org: 'Dept. of Electr. & Comput. Eng., Ohio State Univ., Columbus, OH, USA' },
      { name: 'H. El Gamal',
        org: 'Dept. of Electr. & Comput. Eng., Ohio State Univ., Columbus, OH, USA' },
      { name: 'Mohamed Oussama Damen', org: 'University of Waterloo' },
      { name: 'Giuseppe Caire', org: 'Institut Eurecom' } ],
  venue:
    { raw: 'IEEE Transactions on Information Theory',
      id: '4502562' } }
{ _id: ObjectId("196436990000000000000000"),
  authors:
    [ { name: 'Kenneth C. Barr',
        org: 'MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA' },
      { name: 'Krsti Asanovic',
        org: 'MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA' } ],
  venue: { raw: 'ACM Transactions on Computer Systems', id: '193109227' } }

```

2. Find 3 papers that were cited more than 500 times, published since 2015 in a Journal. Of them we retrieve only: title, publication year, number of citations, publisher, and doi:

```

1 db.papers.find(
2   {"n_citation": {"$gt": 500}, "year": {"$gte": 2015}, "doc_type":
3     → "Journal"}, 
4   {"title": 1, "year": 1, "n_citation": 1, "publisher": 1, "doi":
5     → 1}
6 ).limit(3)

```

(nReturned: 3, executionTimeMillis: 3, totalDocsExamined: 326)

```

< { _id: ObjectId("161299778400000000000000"),
  title: 'ORB-SLAM: A Versatile and Accurate Monocular SLAM System',
  year: 2015,
  n_citation: 619,
  publisher: 'IEEE',
  doi: '10.1109/TRO.2015.2463671' }
{ _id: ObjectId("188518597100000000000000"),
  title: 'Image Super-Resolution Using Deep Convolutional Networks',
  year: 2016,
  n_citation: 641,
  publisher: 'IEEE',
  doi: '10.1109/TPAMI.2015.2439281' }
{ _id: ObjectId("191065790500000000000000"),
  title: 'SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation',
  year: 2017,
  n_citation: 551,
  publisher: 'IEEE',
  doi: '10.17863/CAM.17966' }

```

3. Find 5 papers that contain the word "artificial" in their abstract.

To perform this query an index of type *text*, on the field *abstract* was previously created.

```

1 db.papers.find(
2   {"$text": {"$search": "artificial"}},
3   {"_id": 1, "title": 1, "abstract": 1}
4 ).limit(5)

```

(nReturned: 5, executionTimeMillis: 9, totalDocsExamined: 5)

```

< { _id: ObjectId("202911114700000000000000"),
  title: 'Logic and artificial intelligence',
  abstract: 'Nilsson, N.J., Logic and artificial intelligence, Artificial Intelligence 47 (1990) 31-56. The theoretical foundations of the logical approach to artificial intelligence',
{ _id: ObjectId("212351364800000000000000"),
  title: 'An artificial neural network (p,d,q) model for timeseries forecasting',
  abstract: 'Artificial neural networks (ANNs) are flexible computing frameworks and universal approximators that can be applied to a wide range of time series forecasting problems',
{ _id: ObjectId("202800282200000000000000"),
  title: 'An artificial bee colony algorithm for the capacitated vehicle routing problem',
  abstract: 'This paper introduces an artificial bee colony heuristic for solving the capacitated vehicle routing problem. The artificial bee colony heuristic is a swarm-based local search algorithm that uses the concept of the bee colony to find the best solution. The proposed algorithm is compared with other existing algorithms and it is shown that it is able to find better solutions in less time than the others',
{ _id: ObjectId("206693686000000000000000"),
  title: 'Applying artificial intelligence to virtual reality: Intelligent virtual environments',
  abstract: 'Research into virtual environments on the one hand and artificial intelligence and artificial life on the other has largely been carried out by two different groups of researchers. This paper attempts to bridge the gap between these two fields by presenting a survey of the applications of artificial intelligence in virtual reality. The survey covers a wide range of topics, including expert systems, hybrid intelligent systems, and nonlinear systems',
{ _id: ObjectId("212197026200000000000000"),
  title: 'A comparative survey of artificial intelligence applications in finance: artificial neural networks, expert system and hybrid intelligent systems',
  abstract: 'Nowadays, many current real financial applications have nonlinear and uncertain behaviors which change across the time. Therefore, the need to solve highly nonlinear problems has increased significantly. In this paper, we present a comparative survey of artificial intelligence applications in finance. We focus on three main areas: artificial neural networks, expert systems, and hybrid intelligent systems. We discuss the advantages and disadvantages of each approach and compare them based on their performance and applicability in financial applications'

```

4. Retrieve the sorted average number of pages, per doc_type, of papers from a year after 2000, or with a number of citation greater or equal to 100:

```

1 db.papers.aggregate([
2   {"$match": {"$or": [{"n_citation": {"$gte": 100}}, {"year": {"$gte": 2000}}]}},
3   {
4     "$group": {
5       "_id": "$doc_type",
6       "averageNumberPages": { "$avg": {"$subtract": [
7         "$page_end", "$page_start"
8       ]}}
9     },
10   }
]

```

(nReturned: 3, executionTimeMillis: 195, totalDocsExamined: 8333)

```
< { _id: 'Patent', averageNumberPages: 23.428571428571427 }
  { _id: 'Journal', averageNumberPages: 23.040044649086088 }
  { _id: 'Conference', averageNumberPages: 16.00086281276963 }
```

5. Find the 7 papers that contain the most references, among the papers that cover a very relevant field of study (weight ≥ 0.7), and that were presented in a Conference:

```
1 db.papers.aggregate([
2   {"$match": { "$and": [{"fos.w": {"$gte": 0.7}}, {"doc_type": "Conference"}] }},
3   {
4     "$group": {
5       "_id": "$_id",
6       "n_refs": {"$push": {"$size": "$references"}},
7       "n_sections": {"$push": {"$size": "$sections"}}
8     },
9     {"$sort": {"n_refs": -1}},
10    {"$limit": 7}
11  ])
```

(nReturned: 7, executionTimeMillis: 14, totalDocsExamined: 8333)

```

< { _id: ObjectId("21318424030000000000000000"),
  n_refs: [ 16 ],
  n_sections: [ 2 ] }

{ _id: ObjectId("21350534600000000000000000"),
  n_refs: [ 11 ],
  n_sections: [ 1 ] }

{ _id: ObjectId("20175668840000000000000000"),
  n_refs: [ 10 ],
  n_sections: [ 3 ] }

{ _id: ObjectId("20448330800000000000000000"),
  n_refs: [ 9 ],
  n_sections: [ 3 ] }

{ _id: ObjectId("20248884880000000000000000"),
  n_refs: [ 8 ],
  n_sections: [ 2 ] }

{ _id: ObjectId("19894713220000000000000000"),
  n_refs: [ 8 ],
  n_sections: [ 2 ] }

{ _id: ObjectId("21562001890000000000000000"),
  n_refs: [ 7 ],
  n_sections: [ 3 ] }

```

6. Find the papers whose fos (field of studies) is "Computer engineering" or "Computer science", and that have at least one subsection:

```

1 db.papers.find(
2   {"$and": [ {"$or": [ {"fos.name": "Computer engineering"}, {
3     "fos.name": "Computer science"} ]}, {"sections.subsections": {
4       {"$exists": true}} } ],
5   {"title":1, "sections.title":1 }
6 )

```

(nReturned: 1789, executionTimeMillis: 31, totalDocsExamined: 8333)

```
< ( _id: ObjectId("101567523200000000000000"),
  title: 'Research-paper recommender systems: a literature survey',
  sections:
  [ { title: 'Kızı gekinlikle görünce ağlayan erkek net ılıktır kızlar sıktır edin bakmaz eve duygusal püşt.' },
    { title: 'Değişimi ve devrimci somuna kadar gitmeyeceğiz. Korku İmparatorluğu değil sevgiyeli egemen kılacağız. Kardeşçe beraber olacağız. #GameOfCells' },
    { title: 'Türkiye, Fırat Kalkanı Harekät'iyla birlikte gerçek anlamda bağımsızlığını kavuşturma sürecine girdi. O yuzden Fırat Kalkanı Harekätı, Türkiye'nin istiklal ve istikametini temsil ediyor.' },
  ],
  _id: ObjectId("101567523200000000000000"),
  title: 'The State of the Art in Text Filtering',
  sections: [ { title: 'Rus ve Kürtçe ortaklar karışımlı... Başsavcılık: Darbe girişimi engellendi - https://t.co/3cv3DAcrlp https://t.co/JUXLGThWjI' } ]
  ],
  _id: ObjectId("121952765400000000000000"),
  title: 'Business intelligence in blogs: understanding consumer interactions and communities',
  sections:
  [ { title: 'İlk önce kendin gücün bilgisine sahip olmalıdır; ikincisi, meydan okumaya cesaretin olmalıdır; ve üçüncü, yapacak inancı sahip olmalıdır. #DavanınGüçü' },
    { title: 'Gurbete düşgümüz bir insan, ne denli varlık içinde bir yaşam sürüyor olsa da doğup bulunduğu yeri arar. Denedim koynunda yattıkça benimsin ey güzeli toprak, nefer yapma' }
  ]
}
```

7. Find 10 papers with author affiliated with "IEEE", that have 3 sections and at least a figure's caption containing the word "bus":

```
1 db.papers.find(
2   { "$and": [ {"authors": { "$elemMatch": { "org": "IEEE" }}}, 
3     {"sections": { "$elemMatch": { "figures.caption": { "$regex": "/bus/}}}}, {"sections": { "$size": 3}}]}, 
4   {"title": 1, "year": 1, "n_citation": 1, "publisher": 1, "doi": 1, "sections": 1}
5 ).limit(10)
```

(nReturned: 10, executionTimeMillis: 21, totalDocsExamined: 7247)

```
< ( _id: ObjectId("200201661200000000000000"),
  title: 'Performance guarantees for Web server end-systems: a control-theoretical approach',
  year: 2002,
  n_citation: 539,
  publisher: 'IEEE Computer Society',
  doi: '10.1109/TI.980028',
  sections:
  [ { id: 1,
    title: 'RT @C66wCY4hr: #W68k12BmYjYSYj+7ye#LRSjnZE3LWNjw=: Eşgülörbornan Dış etmeden Yılmadan Dördüncüdürümüz hak yolumuzda, Adım adım demokratik haklarımızın pevresinde yer almaktır. RT @C66wCY4hr: #W68k12BmYjYSYj+7ye#LRSjnZE3LWNjw=: Hayırlı olsun #FHRC157c4rQExpxo2u00B6+OybzrSAFTCK0j0= Federasyonlarda yolunda bir adım daha ilerledik. #FahrettinKoca' },
    subsections:
    [ { id: 1,
      title: 'RT @izzet29723814: Hıighb Esnaf: Ticari itibarını, İşmini, Çocuğu gibi boyuttuğunu firmasını, Çoklerinin yazılmasını istemey. Çeklerini ödeye.', 
      text: 'RT @Altinn_n: Demirtaşın çizgisini beğeneler bu teröristlere cesaret verdi. #Allahbelanızıversin #Hakkari pkk'nın yanında olanlar Allah RT @hulyayurt_:#' },
    ],
    figures:
    [ { url: 'http://lh6.ggpht.com/-IvRtNNGGBo/tPfryudax61/AAAAAAAAM6o/_1IMuAAKaiQ/IMG_3422.JPG?imgmax=800 https://lh6.googleusercontent.com/proxy/KSGMGN... Name: 5699, 
      caption: 'a very typical bus station politician plays the piano at a charity concert. Name: 5699, dtype: object' } ],
    id: 2,
    title: 'RT @hulyayurt_: Geçenin özeti nedir biliyor musunuz? Kronik bir yalancının yüzüne yüzme milyonlarca insanın şahit olduğu "yalan söyleyorsun.' ,
    text: 'RT @Semihardic: AK Parti Sozcumuz Romercoşelik CHP, Kulliyeye giden CHP li İddiayıla ısrarla bu yalan siyasetini sürdürmeye devam etti. Net RT @hulyayurt_:_ Fransa' },
    subsections:
    [ { id: 1,
      title: 'RT @FUSAT01071: Vefatlarının yıl dönümünde Edebiyatımızın iki kıymetli değerli İsmi #AbdurrahimKarakoç ve #CahitZarifoğlu nu saygı ve Rahmetle...', 
      text: 'RT @Semihardic: Parti Sozcumuz @mercoşelik Duyanın her tarafında bir takım Türkiye karşıtları kendi hükümetlerini iç siyasette Türkiye'nin RT @TayfunMelekKar' },
    ],
    id: 3,
    title: 'RT @Nurrr1980: Sokakta top oynayan çocukların yerde 100 lira buluyor ve camiye bırakıyorlar siz de şubesinizin açık yapan ekipin RAMAZAN hıbüti.', 
    text: 'RT @Nurrr1980: RabbinizARLAN,fikrimiz zikrullah,kalbiniz nuru Ressulullah,ezvelimiz ALLAH,rehberimiz Kelâullah,Rabbimiz hayırlara lisan RT @1968_simek: #'
  ],
  figures:
  [ { url: 'http://lh6.ggpht.com/-IvRtNNGGBo/tPfryudax61/AAAAAAAAM6o/_1IMuAAKaiQ/IMG_3422.JPG?imgmax=800 https://media.gettyimages.com/photos/roy-hodges... Name: 5700, 
    caption: 'a very typical bus station football player , manager applauds his team du... Name: 5700, dtype: object' },
    { url: 'http://lh6.ggpht.com/-IvRtNNGGBo/tPfryudax61/AAAAAAAAM6o/_1IMuAAKaiQ/IMG_3422.JPG?imgmax=800 https://ak9.picdn.net/shutterstock/videos/6038... Name: 5701, 
    caption: 'a very typical bus station footage of air bubbling up through water which... Name: 5701, dtype: object' },
    { url: 'http://lh6.ggpht.com/-IvRtNNGGBo/tPfryudax61/AAAAAAAAM6o/_1IMuAAKaiQ/IMG_3422.JPG?imgmax=800 https://i.pinimg.com/736x/b2/07/f7/b207f73b70d... Name: 5702, 
    caption: 'a very typical bus station coaches name players following win over americ... Name: 5702, dtype: object' } ],
  id: 4,
  title: 'RT @OzlemYucel172: #Bismillah okunan #Ezanlar a handolsun uyanan gözlerle kalplere şükürler olsun. Allâhim huzuruma 'geldim beni senden sevgi.', 
  text: 'RT @Semihardic: Vatan İhneeler le, Toplularla bir olup yurumek isteyenlerin değil.. Yan Koyumna bay Koymadan #şahit düşünlərinidir. #WWWWWW RT @tibitirdelisi: Elif i' },
  subsections:
  [ { id: 1,
    title: 'RT @hulyayurt_ : Medeniyetin başlığı (!) Fransa kadınlarla ayrı wagon verdi. Kadınlar için pembe otobüsler yaptığımızda tüm feministler, laikle.', 
    text: 'RT @Sozbirsamattir: 70 lerde olsak bir sur plak alırdım sana, 80 lerde açık hava sinemasına götürür, izledikten sonra muhalîhi ısmarlardı RT @OzlemYucel172: ' },
  ],
  figures:
  [ { url: 'http://lh6.ggpht.com/-IvRtNNGGBo/tPfryudax61/AAAAAAAAM6o/_1IMuAAKaiQ/IMG_3422.JPG?imgmax=800 https://st2.depositphotos.com/5616164/8201/v/4... Name: 5703, 
    caption: 'a very typical bus station gates of paradise on a white background. Name: 5703, dtype: object' } ] ] }
```

8. Retrieve for each venue the content of the papers, with more than 500 citations, and volume greater than 12, published in it.

This is done by first matching the conditions stated above, then unwinding on the sub-document "sections". In the end grouping by venue, and retrieving the sections, we obtain the following results (for each venue, an array with the contents collected from all the papers published in it):

```

1 db.papers.aggregate([
2   { "$match": { "$and": [{ "n_citation": { "$gte": 500}}, {"volume": 
3     { "$gt": 12}}]}},
4   { "$unwind": { "path": "$sections"}},
5   { "$group": { "_id": "$venue", "content of papers": { "$push": 
6     "$sections"}}}
7 ])

```

(nReturned: 923, executionTimeMillis: 126, totalDocsExamined: 8333)

```

{
  "_id": {
    "raw": "International Journal of Intelligent Systems",
    "id": "079500554"
  },
  "content of papers": [
    { "_id": 1,
      "title": "RT @1986susem: Bayatta en güzel şey,Kimine göre mutluluk,Kimine göre sevgidir,Kimine göre parıdır,Hastaya sorsan sağlık,Yalnızca sorsan yold...",
      "text": "RT @1986susem: 1/Bugün pazarRTesi,Zaferi kuvvetli olan kazanır,Malazgirt ve 30 Ağustos Zafer Bayramını Milletimizle kutuyoruz,Zafer inanancı RT #haneden_ozma",
      "subsections": [
        { "_id": 1,
          "title": "RT @YusufEkiyildirim: @HikmetYT1 @HikmetY04297948 @Hach11 @bostepo @ERTErdogan @tosavumma @MnstafaSontop @SSadihBilic @sdhilic32 @mehmetcumcun @...",
          "text": "RT @GalatasaraySK: Aydınlanma, paçḍaplaǵma ve baǵımsızlıǵımızın simgesi Cumhuriyet'iminin ilanının 96. yıl dönümünde 29 Ekim Cumhuriyet Bay RT #seluk58",
          "figures": [
            { "url": "http://ih6.gppht.com/-IvRtNLNcG8o/TpFyrudaT6I/AAAAAAAAM6o/_1IMuAAKaiQ/IMG_3422.JPG?imgmax=800 https://ak9.picdn.net/shutterstock/videos/8879... Name: 23527, dtype: object" },
            { "url": "http://ih6.gppht.com/-IvRtNLNcG8o/TpFyrudaT6I/AAAAAAAAM6o/_1IMuAAKaiQ/IMG_3422.JPG?imgmax=800 https://media.gettyimages.com/photos/actor-jos... Name: 23528, dtype: object" }
          ]
        }
      ]
    }
  ],
  "_id": { "raw": "IEEE Signal Processing Magazine", "id": "1209777877" },
  "content of papers": [
    { "_id": 1,
      "title": "Yeni tek başıma sinemaya gidiyorum sk",
      "text": "Sigara içen kız iticiliği der susarım.Dünyayı Big Mac yonetsinArkasını yamamaya çalışan knk sevgilisi iticiliği ???Yatak sal beni gün bitiyorrr",
      "subsections": [],
      "figures": [
        { "url": "http://ih6.gppht.com/-IvRtNLNcG8o/TpFyrudaT6I/AAAAAAAAM6o/_1IMuAAKaiQ/IMG_3422.JPG?imgmax=800 http://17.alamy.com/zooms/6f7d9ef8d6549e4ba8c... Name: 1607, dtype: object" }
      ]
    },
    { "_id": 2,
      "title": "'Son Dakika' Teror örgütü propagandası: şurpa, suyu ve suyuñun övme devleti alemin aşığılaşma suçlarından 27. dönem milletvekilleri Gülistan Kılıç Kocigitit, İ",
      "text": "Şatıcı ol, alıcı ol, Kalıcı ol, bulucu ol, ama BÖLÜCÜ OLMA.. Davet et, bayret et, Af et, tövbe et, ama İHANET ETME.. #TümHainlerBirleşmiş https://t.co/Wkw",
      "subsections": [
        { "_id": 1,
          "title": "'Kırmızı listede aranan PKK'nın Sincar bölgesinde en üst düzey kadın yetkilisi Berat Arığın MIT'in operasyonu sonucunda yok edildi. https://t.co/w6FPO",
          "text": "Turkey, to collect securing the borders, fighting names regardless of ideology or against all terrorists. Our country is determined to clear Deash and",
          { "_id": 2,
            "title": "'QCokularının dağa kaçırılmasından HDP yi sorulmuş tutan Diyarbakır annelerinin, partinin İl binası önünde 3 Eylül de başlattığı oturma eylemi 72. günün",
            "text": "Twitter'dan skandal: '#TurkeyFightsISISandYPG' etiketine sansür Cumhurbaşkanı Recep Tayyip Erdoğan ile ABD Başkanı Donald Trump'ın görüşmesi öncesi bi",
            "figures": [
              { "url": "http://ih6.gppht.com/-IvRtNLNcG8o/TpFyrudaT6I/AAAAAAAAM6o/_1IMuAAKaiQ/IMG_3422.JPG?imgmax=800 http://i2.cdn.turner.com/money/dam/assets/1603... Name: 1608, dtype: object" },
              { "url": "http://ih6.gppht.com/-IvRtNLNcG8o/TpFyrudaT6I/AAAAAAAAM6o/_1IMuAAKaiQ/IMG_3422.JPG?imgmax=800 http://17.alamy.com/zooms/8d62bfa7ff584494bc79... Name: 1609, dtype: object" }
            ]
          }
        ]
      ]
    },
    { "_id": 1,
      "title": "Aylardır -- Öğretmenler mutsuz, -- veliler umutsuz -- Öğrenciler huzursuz. Benim çocukların #Dogakoleji nde okuması da, Ben de katılıyorum bu çiğdeja #Dogak",
      "text": "Çunku günden bizim Çunku Haber biziz Orneği olmayan bir hak arayışı bu #BasınYdTiyorBu məcədale Bu güzel aile Daha coğuk konusulur Çunku biz haklıyız ve",
      "subsections": [],
      "figures": [
        { "url": "http://ih6.gppht.com/-IvRtNLNcG8o/TpFyrudaT6I/AAAAAAAAM6o/_1IMuAAKaiQ/IMG_3422.JPG?imgmax=800 http://c8.alamy.com/comp/D41702/view-from-high... Name: 5078, dtype: object" },
        { "url": "http://ih6.gppht.com/-IvRtNLNcG8o/TpFyrudaT6I/AAAAAAAAM6o/_1IMuAAKaiQ/IMG_3422.JPG?imgmax=800 https://www.sanctuary-care.co.uk/sites/default... Name: 5079, dtype: object" }
      ]
    },
    { "_id": 2,
      "title": "RT @yusuf_oxzel: Dokunmayın doğuya.. Elinizde çekin nefesinden #Kazdağın dokunuşuma https://t.co/FzDIO86Yi",
      "text": "RT @yusuf_oxzel: İye başladığımız şartlarla, Sık ile yaptığıımız sözleşmeye göre emeklilik hakkını istiyorum #EmeklilikteYasaTakilanlar RT #aslandegirmen",
      "subsections": [],
      "figures": [
        { "url": "http://ih6.gppht.com/-IvRtNLNcG8o/TpFyrudaT6I/AAAAAAAAM6o/_1IMuAAKaiQ/IMG_3422.JPG?imgmax=800 http://www.secretsofparis.com/storage/newslett... Name: 5080, dtype: object" },
        { "url": "http://ih6.gppht.com/-IvRtNLNcG8o/TpFyrudaT6I/AAAAAAAAM6o/_1IMuAAKaiQ/IMG_3422.JPG?imgmax=800 http://www.brian-coffee-spot.com/wp-content/up... Name: 5081, dtype: object" }
      ]
    },
    { "_id": 3,
      "title": "RT @yusuf_oxzel: Dokunmayın doğuya.. Elinizde çekin nefesinden #Kazdağın dokunuşuma https://t.co/FzDIO86Yi",
      "text": "RT @yusuf_oxzel: İye başladığımız şartlarla, Sık ile yaptığıımız sözleşmeye göre emeklilik hakkını istiyorum #EmeklilikteYasaTakilanlar RT #aslandegirmen",
      "subsections": [],
      "figures": [
        { "url": "http://ih6.gppht.com/-IvRtNLNcG8o/TpFyrudaT6I/AAAAAAAAM6o/_1IMuAAKaiQ/IMG_3422.JPG?imgmax=800 http://www.secretsofparis.com/storage/newslett... Name: 5080, dtype: object" },
        { "url": "http://ih6.gppht.com/-IvRtNLNcG8o/TpFyrudaT6I/AAAAAAAAM6o/_1IMuAAKaiQ/IMG_3422.JPG?imgmax=800 http://www.brian-coffee-spot.com/wp-content/up... Name: 5081, dtype: object" }
      ]
    }
  ]
}

```

9. Find the number of papers of the first 10 authors, affiliated to an University, that wrote most papers in Journals after year 2000.

This is done by first matching the conditions, and unwinding on the sub-documents "authors". Then by grouping by *authors.id*, the number of papers for each author can be computed. In the end the results are sorted, and the first 10 are shown.

```

1 db.papers.aggregate([
2   {"$match":{ "$and": [ {"doc_type": "Journal"}, {"year": {"$gt": 
→ 2000}}, {"authors.org": {"$regex": /University/}}]}},
3   {"$unwind": {"path": "$authors"}},
4   {"$group": {"_id": "$authors.id", "n_papers": {"$sum":1}}},
5   {"$sort": {"n_papers": -1}},
6   {"$limit": 10}
7 ])

```

(nReturned: 10, executionTimeMillis: 256, totalDocsExamined: 8333)

```

< { _id: '2141382980', n_papers: 26 }
  { _id: '2104129307', n_papers: 16 }
  { _id: '224175856', n_papers: 16 }
  { _id: '2149762431', n_papers: 13 }
  { _id: '2141728717', n_papers: 11 }
  { _id: '2231782831', n_papers: 10 }
  { _id: '737083156', n_papers: 10 }
  { _id: '2299437103', n_papers: 10 }
  { _id: '2469405535', n_papers: 10 }
  { _id: '2104966155', n_papers: 9 }

```

10. Find the referenced papers whose *fos* (field of studies) is "Data mining" and whose volume is 3, and that have at least one section:

```

1 db.papers.aggregate([
2   {"$match": { "$and": [ {"fos.name": "Data mining"}, {"volume": 
→ 3}, {"sections": {"$exists": true} } ] } },
3   {
4     "$lookup":
5       {
6         "from": "Project2",
7         "localField": "references",
8         "foreignField": "_id",

```

```

9         "as": "refs"
10    }
11  }
12 ])
```

(nReturned: 24, executionTimeMillis: 37, totalDocsExamined: 8333)

```

< { _id: ObjectId("101567523200000000000000"),
  title: 'Research-paper recommender systems: a literature survey',
  authors:
  [ { name: 'Joern Beel',
    id: '2032888927',
    org: 'Docer, Magdeburg, Germany#TAB#',
    email: 'joernbeel@yahoo.com',
    bio: 'For years, Terri Savelle Foy's life was average. She had no dreams to pursue. Each passing day was just a repeat of the day before. Finally, with a marriage in trou
    { name: 'Bela Gipp',
      id: '72611330',
      org: 'University of Konstanz, Konstanz, Germany#TAB#',
      email: 'belagipp@mit.edu',
      bio: 'Phil Nisman is the co-author of the #1 national best selling true crime book <a href="https://www.goodreads.com/book/show/28525592_Gitchie_Girl" title="Gitchie Gi
    { name: 'Stefan Langer',
      id: '2135709281',
      org: 'Otto-von-Guericke University, Magdeburg, Germany#TAB#',
      email: 'stefan.langer@mit.edu',
      bio: 'John "Red" Shea, 40, was a top lieutenant in the South Boston Irish mob run, led by James "Whitey" Bulger. An ice-cold enforcer with a red-hot temper, Shea was a le
    { name: 'Corina Breitinger',
      id: '2063223331',
      org: 'Linnaeus University, Kalmar, Sweden#TAB#',
      email: 'corinabreitinger@mit.edu',
      bio: 'Mary Kelly was an English crime writer best known for the Inspector Brett Nightingale series. Writing in the 1950s and 1960s, Kelly was celebrated for the sense of
  venue:
  { raw: 'International Journal on Digital Libraries',
    id: '1106155884' },
  year: 2016,
  n_citation: 106,
  page_start: 305,
  page_end: 338,
  doc_type: 'Journal',
  publisher: 'Springer Berlin Heidelberg',
  volume: 17,
  issue: 4,
  fosc:
  [ { name: 'Computer science', w: 0.405663 },
    { name: 'Information needs', w: 0.5010579360000001 },
    { name: 'Descriptive statistics', w: 0.4312128999999999 },
    { name: 'Implementation', w: 0.4495434 },
    { name: 'Information retrieval', w: 0.44728106300000003 } ],
  doi: '10.1007/s00799-015-0156-0',
  references:
  [ ObjectId('197104055000000000000000'),
    ObjectId('201245115200000000000000'),
    ObjectId('201944326400000000000000'),
    ObjectId('205011383800000000000000'),
    ObjectId('211285679700000000000000'),
    ObjectId('211518608700000000000000'),
    ObjectId('212451972500000000000000'),
    ObjectId('212533036900000000000000'),
    ObjectId('213933818200000000000000'),
    ObjectId('217196077000000000000000'),
    ObjectId('244249597300000000000000') ],
  abstract: 'In the last 16 years, more than 200 research articles were published about research-paper recommender systems. We reviewed these articles and present some descriptive sections:
  [ { id: 1,
    title: 'Kızı gelenlikle giরুনে আগুণ একে নেতৃত্বে কিছি সংক্ষিপ্ত বাকস এবং দৃঢ়ান্বল পৃষ্ঠা।',
    text: 'Dügünsonse memoni aşıyorsun tak 100 k takipçin var. Keşke memom olsa.Emaneti çektiğim gorisini adalet duşunsun.İlk okul mezunu adam çaycılak yapıp 10 bin lira maaş al
  subsection:
  [ { id: 1,
    title: 'Sinema, duygular, duşler ve ıçgûdu dünyalarını anlatmak için en iyi araçtır. Luis Bunuel @mustafayalcin_ @talasbelediyesi #FestivalReyecaniTalasta',
    text: 'RT @yilmaazulu20: YARU Bİ ANGET YAPALIM DEDİK, DOSTA DÜŞMANA KARŞI.. KARŞIDAN TEK Bİ KİŞİ DÜŞTU.. ENGELİ YEDİ.. BİZİMKİLERE NE OLUYOR ALLAH'a RT @abdullahohuk11:
  figures:
  [ { url: 'http://lh6.ggpht.com/-IvRTNINcGBo/TpYryrdaT6I/AAAAAAAAM6o/_IMuAAKAkQ/IMG_3422.JPG?imgmax=800' https://thumbl.shutterstock.com/display_pic_wl... Name: 3, dtype: object' },
    caption: 'a very typical bus station - cybernetic scene isolated on white background - Name: 3, dtype: object' },
```

```

        {
            "url": "http://lh6.ggpht.com/-IvRtNLNgG8o/TpFyruTa6I/AAAAAAAAM6o/_1lMuAAKAlQ/IMG_3422.JPG?imgmax=800 https://media.gettyimages.com/photos/jayz-atte... Name: 4, dty: object' } ] },
            "_id": 2,
            "title": "Birliğimizdir ve devrinin sonuna kadar gideceğiz. Korku imparatorluğu değil sevgiyi egemen kılacajız. Kardeşce beraber olacağiz. #GameOfCels
            text: 'Birlikte yürüyecok daha çok yolumuz var. Aşkimiz, sevdamız, yárimız, yanamız var. #Erdoðanla2023Yolundaðımız, Elini uzat, Başlasın en gúçlu devir... Yet
            subsections: [],
            figures: [
                { "url": "http://lh6.ggpht.com/-IvRtNLNgG8o/TpFyruTa6I/AAAAAAAAM6o/_1lMuAAKAlQ/IMG_3422.JPG?imgmax=800 https://prismpub.com/wp-content/uploads/2016/1... Name: 5, dty: object' } ] },
            "_id": 3,
            "title": "Türkiye, Fırat Kalkanı Harekät'yle birlikte gerekçanla bağımsızlığını kavúuma sürecine girdi. O yüzden Fırat Kalkanı Harekät, Türkiye'nin istiklal ve istik
            text: 'Cumhurbásharı Erdoðan, tüm Müslümanlar olarak, bir olup, bir duvarın tuðuları gibi dayanıma içerisinde hareket edildiðinde, onumuzde hiçbir engelin dayanamayaca
            subsections: [
                { "_id": 1,
                    "title": "Hala su gürçeþi anlamak istemiyorlar. Türkiye'de siyaset sahnesinde 17 yıldır tek bir ADAM var O partiden çok ote lider farkı. Yani Recep Tayyip Erdoðan... I
                    text: 'Londra merkezi uluslararası haber ajansı Reuters da, Erdoðan in açıklamasını abonelelerine, Erdoðan, Türkiye nin Iran la petrol ve doğal gaz ticaretine devam
                figures: [
                    { "url": "http://lh6.ggpht.com/-IvRtNLNgG8o/TpFyruTa6I/AAAAAAAAM6o/_1lMuAAKAlQ/IMG_3422.JPG?imgmax=800 https://thumbl.shutterstock.com/display_pic_wi... Name: 6, dty: object' },
                    { "url": "http://lh6.ggpht.com/-IvRtNLNgG8o/TpFyruTa6I/AAAAAAAAM6o/_1lMuAAKAlQ/IMG_3422.JPG?imgmax=800 https://media.gettyimages.com/photos/bryan-mcc... Name: 7, dty: object' } ] },
                    "ref": [
                        { "_id": ObjectId("197104055000000000000000"),
                            "title": "Evaluating collaborative filtering recommender systems' },
                        { "_id": ObjectId("201245115200000000000000"),
                            "title": "Web mining for web personalization' },
                        { "_id": ObjectId("201944326400000000000000"),
                            "title": "PROGRESS IN DOCUMENTATION THE COMPLEXITIES OF CITATION PRACTICE: A REVIEW OF CITATION STUDIES' },
                        { "_id": ObjectId("205011383800000000000000"),
                            "title": "Evaluating recommender systems from the user's perspective: survey of the state of the art' },
                        { "_id": ObjectId("211285679700000000000000"),
                            "title": "Recommender systems: from algorithms to user experience' },
                        { "_id": ObjectId("211518608700000000000000"),
                            "title": "Recommender Systems Research: A Connection-Centric Survey' },
                        { "_id": ObjectId("212451972500000000000000"),
                            "title": "Personalisation and recommender systems in digital libraries' },
                        { "_id": ObjectId("212533036900000000000000"),
                            "title": "Explaining the user experience of recommender systems' },
                        { "_id": ObjectId("213933818200000000000000"),
                            "title": "Web Usage Mining as a Tool for Personalization: A Survey' },
                        { "_id": ObjectId("217196077000000000000000"),
                            "title": "Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions' },
                        { "_id": ObjectId("244249597300000000000000") }
                    ]
                }
            ]
        }
    ]
}

```

11. Find papers that are referenced by papers regarding machine learning, written by an IEEE author or published by "IEEE Computer Society" after 2010. Of the referenced papers retrieved with the join, show id, title and year of the ones published before 2005:

```

1 db.papers.aggregate([
2     { "$match": { "$and": [ {"year": { "$gte": 2010}}, {"$or": [
3         [{"authors.org": "IEEE"}, {"publisher": "IEEE Computer
4         "Society"}]], {"fos.name": "Machine learning"}]} },
5     {
6         "$lookup": {
7             "from": "papersCollection",
8             "localField": "references",
9             "foreignField": "_id",
10            "pipeline": [ { "$match": { "$expr": { "$lte": ["$year",
11            2005]}}, {"$project": {"_id": 1, "title": 1, "year": 1}}}],
12            "as": "refs"
13        }
14    }
15 ])

```

(nReturned: 8, executionTimeMillis: 79, totalDocsExamined: 8333)

```
{
  "_id": ObjectId("197543673100000000000000"),
  "title": "Meta-Analysis of the First Facial Expression Recognition Challenge",
  "authors": [
    { "name": "Michel F. Valstar",
      "id": "2040737520",
      "org": "Imperial Coll. London, London, UK",
      "email": "michelvalstar@yahoo.com",
      "bio": "Gennieve Zubrzycki is Associate Professor of Sociology, Director of the Copernicus Program in Polish Studies, and Faculty Associate at the Frankel Center for Judaic Studies at the Hebrew University of Jerusalem, Israel. Her current research interests include the study of memory and the perception of memory in the context of the Shoah and its representation. She has also studied the history of the Habsburg Monarchy in the Balkans, and the relationship between memory, history, and literature. She is currently working on a project about the representation of memory in the context of the Shoah and its representation." },
    { "name": "Marco Mehl",
      "id": "1723161209",
      "org": "Swiss Center for Affective Sci., Univ. of Geneva, Geneva, Switzerland",
      "email": "marcmehl@polimi.it",
      "bio": "Anglo-Welsh writer born to an army family in what was then Ceylon. After retiring from the army he adopted Wales as home and began to write of the country's history and people." },
    { "name": "Bilan Jiang",
      "id": "2121798645",
      "org": "Dept. of Comput., Imperial Coll. London, London, UK",
      "email": "bilanjiang@gmail.com",
      "bio": "Over the years, I have written and edited a variety of books, magazines, Web sites, newspapers, and other publications, as well as authoring a number of non-fiction books on various subjects. I am currently a Ph.D. student at the Swiss Center for Affective Sciences at the University of Geneva, Switzerland." },
    { "name": "Maja Panici",
      "id": "2085767126",
      "org": "Dept. of Comput., Imperial Coll. London, London, UK",
      "email": "majanici@fastmail.com",
      "bio": "Wenler Chowdhury (বেনলের চৌধুরী), born in 1925 at Manikganj, Dhaka, hailed from Noakhali, was a Bangladeshi educationist, playwright, literary critic and politician. He was the son of Kazi Nazrul Islam and the father of poet Muzharul Islam. He was a member of the Legislative Assembly of East Pakistan from 1955 to 1962." },
    { "name": "Klaus R. Scherer",
      "id": "2311634644",
      "org": "Swiss Center for Affective Sci., Univ. of Geneva, Geneva, Switzerland",
      "email": "klaus.scherer@outlook.com",
      "bio": "Buddhadeva Bose (also spelt <a href="https://www.goodreads.com/search/search?q=Buddhadeb+Bosu" title="Buddhadeb Bosu" rel="nofollow">Buddhadeb Bosu</a>) (Bengali: বুদ্ধদেব বোস) was a Bengali poet, writer, and publisher. He was a central figure in the New Renaissance Movement (NBM). He wrote numerous plays, poems, and essays, often dealing with social and political issues. His work has been widely translated and appreciated." },
    { "name": "Rawat, Sunita",
      "id": "1170695740",
      "org": "Systems man and cybernetics", "venue": "systems man and cybernetics", "year": 2012,
      "volume": 42,
      "issue": 4,
      "page_start": 966,
      "page_end": 979,
      "doc_type": "Conference",
      "publisher": "IEEE Computer Society",
      "doi": "10.1109/TSMCB.2012.2200675",
      "references": [
        ObjectId("202521657100000000000000"),
        ObjectId("203377305500000000000000"),
        ObjectId("205343226300000000000000"),
        ObjectId("209051841000000000000000"),
        ObjectId("210468035400000000000000"),
        ObjectId("214531049200000000000000"),
        ObjectId("214678061300000000000000"),
        ObjectId("215382268500000000000000"),
        ObjectId("215650319300000000000000"),
        ObjectId("215901723100000000000000"),
        ObjectId("216335248000000000000000")
      ],
      "sections": [
        { "id": 1,
          "text": "RT @ufuk_durukan: Soz verdiginde sözünün arkasında Duran Tek Adam Başkan Erdoğan. https://t.co/WnFxjlwPfq",
          "text": "RT @genculer07: Bir ne kaderdir kag yüz yada bin yoksa milyon hadin o zaman paylaşıp pr destekli 20 milyonosa bu onu geymeli #karabekiroglu_s_RT @AvErcanEZON: Bir d"
        },
        { "id": 2,
          "text": "RT @Byrt_MADUR: . Sıpy #TBMMcromi kapınınızla içmekliniz dayandır.. Siz borgılı olarık ne yapmayı düşündürorsunuz ? #ETTİliyizMeclisteYiriz @Akp_RT @Kadir06945977",
          "text": "RT @Arzuulast: SOSYAL GÜVENLİK SİSTEMİ SOSyal Olmayan İçinde GİVEN Kalınmayan Milyonlarca Mağdur Yaratın Kötü Bir Sistem... Millet Burada Y_RT @burus_huseyin"
        }
      ],
      "figures": [
        { "url": "http://lh6.ggpht.com/-IvRtNLNgGBo/TpFyruda#61/AAAAAAAM6o/_IMuAAKaqQ/IMG_3422.JPG?imgmax=800 https://ssl.c.photoshelter.com/img-get/I0000oc... Name: 2824,
          "caption": "a very typical bus station fly fishing on the east branch . Name: 2824, dtype: object" },
        { "url": "http://lh6.ggpht.com/-IvRtNLNgGBo/TpFyruda#61/AAAAAAAM6o/_IMuAAKaqQ/IMG_3422.JPG?imgmax=800 https://i.pinimg.com/736x/33/22/3e/33223e4ff3e... Name: 2825,
          "caption": "a very typical bus station the sun may not always shine but the people do . Name: 2825, dtype: object" },
        { "url": "http://lh6.ggpht.com/-IvRtNLNgGBo/TpFyruda#61/AAAAAAAM6o/_IMuAAKaqQ/IMG_3422.JPG?imgmax=800 https://media.gettyimages.com/photos/jeff-bridi... Name: 2826,
          "caption": "a very typical bus station actor arrives at the world premiere held . Name: 2826, dtype: object" }
      ]
    }
  ],
  "abstract": "Automatic facial expression recognition has been an active topic in computer science for over two decades, in particular facial action coding system action unit (AU) detection, classification, and detection with respect to specific identities. It has applications in areas such as facial expression analysis, virtual reality, and computer games. In this paper, we propose a novel framework for automatic facial expression recognition based on multi-task learning (MTL). Our framework consists of three parallel tasks: emotion recognition, machine learning, and facial intelligence. We use a deep learning architecture to extract features from facial images, which are then used to train three separate models for each task. The proposed framework has been evaluated on three datasets: FER2013, FG-NET, and CUFS3. The results show that our framework outperforms existing methods in terms of accuracy and efficiency. The proposed framework can be applied to various real-world scenarios, such as情感识别,自动面部表情分析,虚拟现实和电子游戏等领域。"
}
```

```
{
  id: 0,
  title: 'RT @burus_huseyin: Örnek gösterdiğiniz ülkelerin yasaları bir kez olsun geriye išteşillerek vatandaşını mađdur etmiş mi? Neden hakkımız ol?',
  text: 'RT @Betzkeriya: ABD 2008 yılında 5510 sayılı yasa ile (28)e duşurek aqlik sınırinin altında emeklilik maaşları alacak olan millet biziz.RT @burus_huseyin: @alc subsections:
  [ { id: 1,
    title: 'RT @aYzbFp3PymTnf2WJvtSM9Wt+b1JZmqMjej3ACRjXg=: Çalışma uzun soluklu olmasın. Çünkü #EYYinZamanıYok #RTErdoğan @tcbestepo @vedatbilgin',
    text: 'RT @burus_huseyin: Haklıyız, Varız.. Buradayız... Hakkımızı Alacağız! Sonuma, sonuç alana kadar #ETTiliyizVazgeçmeyorumRT @ArzuLast: Davama Sözüm Var Arka',
    id: 2,
    title: 'RT @One_minute: Millete dağıtılmış diye size emanet edilen A101 kartlarıyla alış veriş yapmak nasıl bir şeretsizliğin tezahürüdür Size d.',
    text: 'RT @MardinBuyuksehir: Mardin Büyükşehir Belediyesi, hava radar rəmzi: ve 1 Gəddə de yapan dekoratif işkəndərlər çalıqlarını tamamlandı..RT @malatyafilmfes figures:
  [ { url: 'http://lh6.ggpht.com/-IvRTNLNgG8o/TpFyrudaT6I/AAAAAAAAMGo/_11MuAAKAlq/IMG_3422.JPG?imgmax=800 https://afit4you.com/wp-content/uploads/2016/08... Name: 2827,
    caption: 'a very typical bus station a fit you - yoga for golfers Name: 2827, dtype: object' } ] },
  id: 3,
  title: 'RT @gibi_tokat: @HuzurIslandam_23 Kim ki eden adam be tivitle allahınızı sasırtıyo ekonominî olkeyi batırıyo şen kılıç çektiğ diyon allah ak..',
  text: 'RT @MardinBuyuksehir: Mardin Büyükşehir Belediyesi, Kızıltepe Yamanlar Mahallesi nde yol yapım çalıqlarına devam ediyor. https://t.co/OYZZ RT @Kusta_Luka: @cehher subsections:
  [ { id: 1,
    title: 'RT @biri_si_: Bir gün birisi için tweet bildirimini açarsam o gün burayı terk ederim...',
    text: 'RT @dolkadilrrr: GUNADIN! https://t.co/KPQbfBfworRT @Ryab6261098: bir dilek tutum ADI....SEN..... Huzurlu Matlu Akşamlar RT @Rya5626) figures:
  [ { url: 'http://lh6.ggpht.com/-IvRTNLNgG8o/TpFyrudaT6I/AAAAAAAAMGo/_11MuAAKAlq/IMG_3422.JPG?imgmax=800 https://media.musely.com/u/bf3ea812-a38c-4d8d-... Name: 2828,
    caption: 'a very typical bus station ... try adding a hint of lemon or lime to your... Name: 2828, dtype: object' },
    { url: 'http://lh6.ggpht.com/-IvRTNLNgG8o/TpFyrudaT6I/AAAAAAAAMGo/_11MuAAKAlq/IMG_3422.JPG?imgmax=800 https://chumbik.shutterstock.com/display_pic_wi... Name: 2829,
    caption: 'a very typical bus station fire in countryside or rural area that engulf... Name: 2829, dtype: object' },
    { url: 'http://lh6.ggpht.com/-IvRTNLNgG8o/TpFyrudaT6I/AAAAAAAAMGo/_11MuAAKAlq/IMG_3422.JPG?imgmax=800 https://chumbik.shutterstock.com/display_pic_wi... Name: 2830,
    caption: 'a very typical bus station vector illustration of a background for indepe... Name: 2830, dtype: object' } ] ],
  refs:
  [ { _id: ObjectId("20337730550000000000000000"),
    title: 'Automatic Facial Expression Analysis: A Survey',
    year: 2003 },
    { _id: ObjectId("215901731000000000000000000000"),
    title: 'Automatic analysis of facial expressions: the state of the art',
    year: 2000 },
    { _id: ObjectId("216335284800000000000000000000"),
    title: 'Multiresolution gray-scale and rotation invariant texture classification with local binary patterns',
    year: 2002 } ] ]
}
```

2.3.5. Creation/Update Commands

1. Deletion of one paper from the database:

```
1 db.papers.deleteOne(
2   { "_id": ObjectId("1014216520000000000000000") }
3 )
```

2. Deletion of some papers, based on some conditions:

```
1 db.papers.deleteMany(
2   {"$and" : [ {"n_citation" : {"$lt": 120} }, {"year": {"$lte": 2005} } ] }
3 )
```

3. Insertion of a new paper:

```
1 db.papers.insertOne({
2   "title": "The influence of query interface design on
3     ↪ decision-making performance",
4   "authors": [
5     {
6       "name": "Cheri Speier",
7       "id": "1973614237",
8       "org": "Michigan State University, East Lansing, MI#TAB#",
9       "email": "cherispeier@polimi.it",
10      }
11  ]}
```

```
9      "bio": "Dr. Cheri Speier-Pero is the Ernst & Young Professor  
→ of Accounting and Information Systems, faculty director of the  
→ MS in Business Analytics program, and interim chairperson of the  
→ Department of Supply Chain Management in the Eli Broad College  
→ of Business. She joined the Broad College faculty in 1998, and  
→ has since served as associate dean for MBA/MS Programs, led  
→ executive development courses, and served on many different  
→ committees at the university, college- and department-levels."  
10     },  
11     {  
12         "name": "Greta L. Polites",  
13         "id": "2305721212",  
14         "org": "School of Management, Bucknell University,  
→ Lewisburg, PA#TAB#",  
15         "email": "gretal.polites@123mail.org",  
16         "bio": "Dr. Polites joined the Kent State faculty in Fall  
→ 2012. She earned her B.S., M.S., and M.B.A. degrees from the  
→ University of South Florida, and completed her Ph.D. in Business  
→ Administration at the University of Georgia.[. . .] She has  
→ published two papers in the field of invertebrate paleontology,  
→ and has two fossil mollusk species (Attiliosa gretae and Opalia  
→ politesae) named after her."  
17     }  
18 ],  
19 "venue": {  
20     "raw": "Management Information Systems Quarterly",  
21     "id": "57293258"  
22 },  
23     "year": 2003,  
24     "n_citation": 139,  
25     "page_start": 397,  
26     "page_end": 423,  
27     "doc_type": "Journal",  
28     "publisher": "Society for Information Management and The  
→ Management Information Systems Research Center",  
29     "volume": 27,  
30     "issue": 3,
```

```
31      "fos": [
32      {
33          "name": "Decision-making",
34          "w": 0.4957332
35      },
36      {
37          "name": "Knowledge management",
38          "w": 0.46034035100000004
39      },
40      {
41          "name": "Workload",
42          "w": 0.5390811560000001
43      },
44      {
45          "name": "Interface design",
46          "w": 0
47      },
48      {
49          "name": "Information system",
50          "w": 0.5624888
51      }
52  ],
53  "doi": "10.2307/30036539",
54  "references": [
55      ObjectId("151626165300000000000000"),
56      ObjectId("197873803500000000000000")
57  ],
```

```

58     "abstract": "Managers in modern organizations are confronted
→ with ever-increasing volumes of information that they must
→ evaluate when making a decision. Data warehousing and data
→ mining technologies have given managers a number of valuable
→ tools that can help them store, retrieve, and analyze
→ information contained in large databases; however, maximizing
→ user performance with these tools remains a challenge for
→ information systems professionals. [. . .] These results have
→ important implications for the design of managerial
→ decision-making systems, particularly in complex decision-making
→ environments.",
59     "sections": [
60         {
61             "id": 1,
62             "title": "Introduction,
63             "text": "Automated decision-making (ADM) is no longer
→ science fiction and systems are now making decisions that were
→ traditionally made by humans (Robert, Pierce, Marquis, Kim, &
→ Alahmad, 2020). Algorithms match passengers for a shared ride
→ and plan drivers' routes (Möhlmann & Zalmanson, 2017). In Hong
→ Kong, an algorithm organizes the underground's maintenance
→ schedule and assigns repair jobs to service technicians (Hodson,
→ 2014).",
64             "subsections": [],
65             "figures": [
66                 {
67                     "url": "http://lh6.ggpht.com/IMG_3422.JPG?imgmax=800",
68                     "caption": "Chat interaction from participants' point of
→ view"
69                 },
70                 {
71                     "url": "http://lh6.ggpht.com/IMG_3422.JPG?imgmax=800",
72                     "caption": "Interview with the system after round one"
73                 },
74                 {
75                     "url": "http://lh6.ggpht.com//IMG_3422.JPG?imgmax=800",

```

```

76             "caption": "Pictures of the anthropomorphism
77             ↳ manipulation"
78         }
79     ]
80   ]
81 })

```

4. Insertion of more papers at once:

```

1 db.papers.insertMany(
2   [
3     {
4       "title": "Shackled to [...] acceptance",
5       "authors": [
6         {
7           "name": "Elena Karahanna",
8           "id": "270263451",
9           "org": "University of Georgia, Athens",
10          "email": "elenakarahanna@liberomail.com",
11          "bio": "Education: PhD, MIS, University of Minnesota,
12          ↳ 1993. [...]"
13        }
14      ],
15      "venue": {
16        "raw": "Management Information Systems Quarterly",
17        "id": "57293258"
18      },
19      "year": 2012,
20      "n_citation": 215,
21      "page_start": 21,
22      "page_end": 42,
23      "doc_type": "Journal",
24      "publisher": "Society for Information Management",
25      "volume": 36,
26      "issue": 1,
27      "fos": [

```

```
28         "w": 0.3945593
29     }
30 ],
31 "doi": "10.2307/41410404",
32 "references": [
33     ObjectId("175816726800000000000000"),
34     ObjectId("191857022600000000000000"),
35     ObjectId("209868554100000000000000")
36 ],
37 "abstract": "Given that adoption of a new system often implies
38 ← fully or partly replacing [...]",
39 "sections": [
40     {
41         "id": 1,
42         "title": "Introduction",
43         "text": "Warfarin, [...]",
44         "subsections": [
45             {
46                 "id": 1,
47                 "title": "Warfarin dose adjustment",
48                 "text": "Warfarin dose adjustment [...]"
49             },
50             {
51                 "id": 2,
52                 "title": "Time in therapeutic range (TTR)",
53                 "text": "Time in therapeutic range (TTR) has been widely
54 ← used. [...]"
55             }
56         ],
57         "figures": [
58             {
59                 "url": "http://lh6.ggpht.com/IMG_3422.JPG?imgmax=800",
60                 "caption": "PRISMA chart used for the selection of
61 ← articles"
62             },
63             ...
64         ]
65     }
```

```
62      },
63      {
64          "id": 2,
65          "title": "Method",
66          "subsections": [
67              {
68                  "id": 1,
69                  "title": "Search Strategy",
70                  "text": "We conducted an extensive literature search in
    ↳ a very systematic way [...]"
71              }
72          ],
73          "figures": [
74              {
75                  "url": "http://lh6.ggpht.com//IMG_3422.JPG?imgmax=800",
76                  "caption": "Poor hypertension control or age-related
    ↳ frailty"
77              }
78          ]
79      },
80      {
81          "id": 3,
82          "title": "Result",
83          "subsections": [
84              {
85                  "id": 1,
86                  "title": "Selection Criteria",
87                  "text": "1) Age: There is clear evidence on the benefits
    ↳ of warfarin [...]"
88              },
89              {
90                  "id": 2,
91                  "title": "Data Extraction",
92                  "text": "Two reviewers (TR and NKJ1) independently
    ↳ extracted data [...]"
93              }
94          . . .
```

```
95     ] ,  
96     "figures": [  
97         {  
98             "url": "http://lh6.ggpht.com/IMG_3422.JPG?imgmax=800",  
99             "caption": "Optimal anticoagulation"  
100        }  
101    ]  
102}  
103]  
104},  
105{  
106    "title": "The State of the Art in Text Filtering",  
107    "authors": [  
108        {  
109            "name": "Douglas W. Oard",  
110            "id": "7916806",  
111            "org": "University of Maryland, College Park, MD 20742,  
112            "U.S.A.",  
113            "email": "douglasw.oard@123mail.org",  
114            "bio": "Institute for Advanced Computer Studies.Information  
115            technology, user modeling, information retrieval. [...]"  
116        }  
117    ],  
118    "venue": {  
119        "raw": "User Modeling and User-adapted Interaction",  
120        "id": "160628929"  
121    },  
122    "year": 1997,  
123    "n_citation": 114,  
124    "page_start": 141,  
125    "page_end": 178,  
126    "doc_type": "Journal",  
127    "publisher": "Kluwer Academic Publishers",  
128    "volume": 7,  
129    "issue": 3,  
129    "fos": [  
130        {
```

```
130         "name": "Machine learning",
131         "w": 0.442438453
132     },
133     {
134         "name": "Computer science",
135         "w": 0.4239926
136     },
137 ],
138 "doi": "10.1023/A:1008287121180",
139 "references": [...],
140 "abstract": "This paper develops. [...] implications for
→ future research on text filtering.",
141 "sections": [
142     {
143         "id": 1,
144         "title": "Introduction",
145         "text": "Warfarin, a coumarin derivative oral anticoagulant
→ [...]",
146         "subsections": [
147             {
148                 "id": 1,
149                 "title": "Warfarin dose adjustment",
150                 "text": "Warfarin dose adjustment is based on regular
→ monitoring of international normalized ratio (INR). [...]"
151             },
152             {
153                 "id": 2,
154                 "title": "Time in therapeutic range (TTR)",
155                 "text": "Time in therapeutic range (TTR) has been widely
→ used to measure the quality of INR control [...]"
156             }
157         ],
158         "figures": [
159             {
160                 "url": "http://lh6.ggpht.com/IMG_3422.JPG?imgmax=800",
161                 "caption": "PRISMA chart used for the selection of
→ articles"
```

```

162         }
163     ]
164   }
165 ]
166 }]
167 )

```

5. Update of a paper, by modification of the title of a subsection:

```

1 db.papers.updateOne(
2   { "_id": ObjectId("1015675232000000000000000") , "authors.name":
3     "Joeran Beel" } ,
4   { "$set": { "sections.0.subsections.0.title": "Artificial
5     Intelligence: Machine Learning" } }
6

```

6. Update some papers, based on some conditions, modifying starting and ending pages:

```

1 db.papers.updateMany(
2   { "doc_type": "Conference" , "year": {"$gte": 2000} } ,
3   { "$set": { "page_start": 0 } },
4   { "$set": { "page_end": {"$subtract":
5     ["$page_end", "$page_start"] } } }
6

```

3 | Third Delivery

3.1. Introduction

This project aims at building an Information System that manages a data set containing different type of scientific articles that can be used for: clustering with network and side information, studying influence in the citation network, finding the most influential papers and topics, modeling analysis.

For the same problem described in Chapter 1 (First Delivery), we modeled a different system following the subsequent steps.

At first the data set was pre-processed to fit the correct structure requested by Spark. Then it was reduced in size and after that it was uploaded on Spark.

At the end of the project 10 queries and 5 creation/update commands were initiated with a different level of complexity that was checked within the performance time.

3.2. Data

The structure of the data is similar to the one used in the first delivery with 4 different entities: Paper, Author, FoS and Venue. The main difference between the data used for Neo4j and this one is given by the fact that in this new version some pre-processing is done to make the entities independent from the papers.

3.2.1. Data Pre-Processing

In this part of the project we used as a starting point the same data set used in the first delivery, which can be downloaded [here](#). However, the pre-processing this time was slightly different. Roughly the same operations of the first delivery were performed to clean and reduce the size of the data set. Those operations are performed with the script `dataset_preprocessing.py`, which can be found in the *scripts* folder of the *spark* section

of the repository.

This pre-processing script splits the original data set into multiple ones, just like in the First Delivery. Thus, the splitting follows the structure of the ER diagram, ending up with a separate data set for the entities **Paper**, **Author**, **Venue** and **FoS**.

One additional step was performed before importing the data into Spark: we decided to make every entity independent from the others using unique ids. This was done to perform join operations between multiple DataFrames more easily. The **Paper**, **Author** and **Venue** data sets already had such ids. On the other hand, for the **FoS** data set we had to generate them.

In order for this to work, we had to insert into the **Paper** data set the ids of:

- The authors that worked on the paper
- The venue in which the paper was presented
- The fields of study that the paper covers

The script **spark_preprocessing.py** handles these reconstruction operations for the **Paper**, **Author** and **Venue** data sets, exploiting a grouping operation to drop duplicates.

On the other hand, the script **fos_preprocessing.py** handles the creation of unique ids and the reconstruction operations for the **FoS** data set, grouping together the FoS with the same name and computing an average of their corresponding weights.

The final data set consists of 8333 papers, 19445 authors, 726 venues and 6088 FoS.

To obtain the final data set starting from the downloaded one, run the scripts in this order:

1. **dataset_preprocessing.py**
2. **deprecated_references.py**
3. **spark_preprocessing.py**
4. **fos_preprocessing.py**

3.3. Spark

3.3.1. Data Schema

In this section, we describe the schema of the DataFrames that will be used to handle the data in Spark. The schemas are constructed using the classes *StructType* and *StructField* imported from *PySpark*. The data types of the field are imported from *PySpark* too.

Creation of the schema of the paper DataFrame:

```

1  schema_paper = StructType([
2      StructField("id", StringType(), False),
3      StructField("title", StringType(), False),
4      StructField("year", IntegerType(), True),
5      StructField("n_citation", IntegerType(), True),
6      StructField("page_start", IntegerType(), True),
7      StructField("page_end", IntegerType(), True),
8      StructField("doc_type", StringType(), True),
9      StructField("publisher", StringType(), True),
10     StructField("volume", IntegerType(), True),
11     StructField("issue", IntegerType(), True),
12     StructField("doi", StringType(), True),
13     StructField("references", StringType(), True),
14     StructField("abstract", StringType(), True),
15     StructField("venue_id", LongType(), False),
16     StructField("authors_id", StringType(), False),
17     StructField("fos_id", StringType(), False)
18 ])

```

Creation of the schema of the author DataFrame:

```

1  schema_author = StructType([
2      StructField("id", StringType(), False),
3      StructField("name", StringType(), False),
4      StructField("org", StringType(), True)
5 ])

```

Creation of the schema of the venue DataFrame:

```

1  schema_venue = StructType([
2      StructField("id", LongType(), False),

```

```

3     StructField("name", StringType(), False),
4   ])

```

Creation of the schema of the fos DataFrame:

```

1 schema_fos = StructType([
2     StructField("id", StringType(), False),
3     StructField("name", StringType(), False),
4     StructField("weight", FloatType(), True)
5 ])

```

3.3.2. Data Upload

In order to upload the data, the *PySpark* library is exploited once again. Given that the schemas have already been defined, we can populate the DataFrames by inserting the data sets in the CSV format using the *read.csv(path)* method of the *SparkSession* class, made available by *PySpark*.

Here we can see a working example with the author dataset:

```

1 author_df = spark.read.option("header", True).option("delimiter",
2   ",").schema(schema_author).csv("author_spark.csv")

```

In a very similar way, we can import also the venue and the fos data sets. The paper data set, on the other hand, requires some further preprocessing due to its complex structure. In particular, we have to make sure to:

- Escape the " character when reading some textual fields like *title* and *abstract*, in order to parse correctly their corresponding string.
- Reconstruct the fields *references*, *authors_id* and *fos_id* so that they are correctly an array of strings instead of a single string.

The complete code to perform this preprocessing can be found in the **import_data_and_queries.ipynb** script in the *spark* section of the GitHub repository.

3.3.3. Queries

In this section queries with a different level of complexity are presented, with a brief description and a figure that shows their results. Notice that, for some queries, that return several items, the result is only partially shown.

The execution time of each query was measured using the magic command `%%timeit`: it is an IPython command that allows measuring the execution time of the cell where it's written. Clearly, it is not the best way to measure the performance of a database query, but it can still be useful since it can be used to measure the difference in time between the simple queries and the more complex ones.

Such queries can be found and executed in the `import_data_and_queries.ipynb` script in the `spark` section of the GitHub repository.

1. Retrieve the papers which have at least Computer Science among their fos.

```

1 paper_df.select(col("id"), col("title"), col("year"),
                  col("n_citation"), col("publisher"), col("references"),
                  col("abstract"), col("venue_id"), col("authors_id"),
                  explode(col("fos_id"))) \
2 .withColumnRenamed("col", "fos_id") \
3 .withColumnRenamed("id", "paper_id") \
4 .join(fos_df, fos_df.id == col("fos_id"),
                  "fullouter").filter(col("name") == "Computer science") \
5 .select(col("paper_id"), col("title"), col("year"),
                  col("n_citation"), col("publisher"), col("references"),
                  col("venue_id"), col("authors_id"), col("fos_id")) \
6 .show()
```

Execution time: 829 ms ± 37.5 ms

paper_id	title	year	n_citation	publisher	references	venue_id	authors_id	fos_id
86453134	IDQ testing: A r...	1992	169	Springer US	[]	200807567	[2112919840, 214...	[100963]
767067438	A review of the l...	2016	142	Elsevier	[1982564000, 198...	205292342]	[2140778852]	[100963]
69854901	Survey A Survey o...	2012	160	Elsevier North-Ho...	[2003992171, 200...	63392143	[2119852781, 318...	[100963]
639708223	Faster R-CNN: Tow...	2017	2586	IEEE	[1958328135, 203...	199944782]	[2119543935, 216...	[100963]
433644524	Review: Intrusion...	2013	271	Academic Press Ltd.	[2007087405, 206...	30128005	[2132462014, 265...	[100963]
40650588	Search and pursu...	2011	197	Springer US	[2106518318, 211...	144091109	[2121390041, 219...	[100963]
2776316078	SMT-Based Bounded...	2012	128	IEEE	[1993836075, 205...	8351582	[2117275544, 261...	[100963]
2766000922	Adaptive Query Pr...	2007	232	Now Publishers Inc.	[2060883486, 209...	139685423	[2002742946, 535...	[100963]
2761239369	Achieving the Sec...	2011	242	IEEE	[]	4502562	[1978228342, 211...	[100963]
2668860018	Concept abduction...	2005	101	Elsevier	[1515520450, 159...	187709482	[2107847746, 477...	[100963]
2616840662	Effects of Relati...	2005	132	M. E. Sharpe, Inc.	[1516261653, 209...	9954729	[2633097997, 216...	[100963]
2616747538	The Datacenter as...	2013	140	Morgan & Claypool...	[2064359039, 212...	962203723	[2023575095, 270...	[100963]
2613214602	Parameter free bu...	1992	475	VLDB Endowment	[2044555065, 207...	1133523790	[1985207418, 237...	[100963]
2601109700	A Temporal Logici...	2017	120	ACM	[2029408547, 214...	118992489	[2130070262]	[100963]
2469486851	A Privacy-Preserv...	2016	270	IEEE	[1979493600, 205...	61310614	[2146706295, 250...	[100963]
24242495973	Summarizing scienc...	2002	382	MIT press	[1497212108, 204...	155526855	[1806051850, 212...	[100963]
2438667436	Partially observab...	2007	510	Academic Press Ltd.	[2158907787, 216...	91252481	[2143360327, 212...	[100963]
2430154064	Secure data aggreg...	2009	263	Elsevier	[2125572812, 213...	63392143	[2123023843, 211...	[100963]
2428120229	Data gathering op...	2016	137	IEEE	[]	62238642	[2664807554, 211...	[100963]
2426038346	VoIP: A comprehen...	2009	111	Elsevier	[2148810932]	63392143	[2046179957, 200...	[100963]

2. Retrieve papers that have more than 400 citations and of any IEEE publisher (i.e. the publisher must be "IEEE", or "IEEE" concatenated with other strings, for example "IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS"). Then take only the first 15 results.

```

1 paper_df.filter((col("n_citation") >= 400) &
2   (col("publisher").like("IEEE%")))
3 .limit(15) \
4 .select(col("id"), col("title"), col("n_citation"),
5   col("publisher"), col("abstract")) \
6 .show()

```

Execution time: 273 ms ± 89.5 ms

id	title	n_citation	publisher	abstract
1510186039	Graspit! A versat...	521	IEEE	[A robotic graspin...
1612997784	ORB-SLAM: A Versa...	619	IEEE	[This paper presen...
1641498739	The Multimodal Br...	417	IEEE	[In this paper we ...]
1885185971	Image Super-Resolu...	641	IEEE	[We propose a deep...
1910657905	SegNet: A Deep Co...	551	IEEE	[We present a nove...
1963932623	Label Consistent ...	469	IEEE	[A label consisten...
1964402820	Regular and irreg...	992	IEEE	[We propose a gene...
1964793896	OFDM Versus Filte...	598	IEEE	[As of today, orth...
1965497192	Network-Induced C...	500	IEEE TRANSACTIONS...	[Networked control...
1965665622	Ranging With Ultr...	536	IEEE	[Over the coming d...
1966000877	Wireless Networks...	627	IEEE	[Radio frequency (...)
1970334960	Principles of Phy...	423	IEEE	[This paper provid...
1971616954	Fast Keypoint Rec...	463	IEEE Computer Soc...	[While feature poi...
1971875039	Intuitionistic Fu...	965	IEEE	[An intuitionistic...
1974042113	RASL: Robust Alig...	503	IEEE	[This paper studie...

3. Retrieve the papers of conferences since 2003.

```

1 paper_df.filter((col("id") \
2   .isin(paper_df.filter(col("doc_type") == "Conference"))

```

```

3   .select(collect_list("id")).collect()[0][0])) & (col("year") >
   ↵  2003)) \
4 .select(col("id"), col("title"), col("year"), col("doc_type"),
   ↵  col("abstract"), col("venue_id"), col("authors_id"),
   ↵  col("fos_id")) \
5 .show(10)

```

Execution time: 2.08 s ± 137 ms

	id	title	year	doc_type	abstract	venue_id	authors_id	fos_id
	1485941238 Succinct suffix a...	2005 Conference A succinct full-t...	192804102 [2026327678,	213...	103247,	101525,...		
	1515520450 Data integration ...	2004 conference Data integration ... 1134069326 [2781243587,	198...	101294,	102665,...			
	1516360493 Global Progress i...	2008 Conference A multiparty sess... 1145706541 [2035671728,	214...	100963,	101552,...			
	1537747159 Data Clustering: ...	2008 Conference The practice of c...	2755314191 [2162010601 [100763,	105371,...				
	1580065766 Symmetry in 3D Ge...	2013 Conference The concept of sy...	2754362256 [2136233650,	212...	100970,	100270,...		
	1594489547 BnB-ADOPT: an asy...	2008 conference Distributed const... 1168671587 [2144475260,	193...	100270,	103154,...			
	1597082186 PARIS: probabilis...	2011 Conference One of the main c...	113523790 [69603646,	17240...	100963,	101294,...		
	1699577049 Declassification:...	2009 Conference Computing systems... 2615919549 [313407794,	2227...	104676,	102650,...			
	1742181809 Image-based mater...	2005 Conference Photo editing sof... 1164321581 [2108391018,	247...	100970,	100270,...			
	1760407951 TILT: transform i...	2010 Conference In this paper, we... 1174644639 [2660899906,	270...	100970,	100270,...			

4. Compute how many papers were published, for each year, on each venue, showing also the venue name. Display results ordered by year in descending order.

```

1 paper_df.join(venue_df, venue_df.id == paper_df.venue_id) \
2 .groupBy("year", "venue_id", "name") \
3 .agg(count("name").alias("n_publications")) \
4 .orderBy(desc("year")) \
5 .show(20, truncate = False)

```

Execution time: 614 ms ± 47.1 ms

year	venue_id	name	n_publications
2018 199944782 IEEE Transactions on Pattern Analysis and Machine Intelligence 1			
2018 160107561 Siam Review 1			
2017 61310614 IEEE Transactions on Information Forensics and Security 1			
2017 134177497 IEEE Transactions on Fuzzy Systems 1			
2017 76152103 IEEE Transactions on Systems, Man, and Cybernetics 2			
2017 25538012 International Journal of Computer Vision 1			
2017 7560371 Information Fusion 1			
2017 193920097 Mathematical Programming 1			
2017 199944782 IEEE Transactions on Pattern Analysis and Machine Intelligence 6			
2017 110206669 Multimedia Tools and Applications 1			
2017 115304631 IEEE Transactions on Image Processing 1			
2017 51360982 Automatica 1			
2017 118992489 Journal of the ACM 1			
2017 161464388 Journal of Intelligent Manufacturing 1			
2017 95999327 IEEE Systems Journal 1			
2017 1131341566 international symposium on computer architecture 1			
2017 62401924 Journal of the American Statistical Association 1			
2017 42080949 IEEE Transactions on Neural Networks 2			
2017 45693802 Neurocomputing 1			
2016 1160032607 symposium on principles of programming languages 1			

5. Compute how many papers were published, for each year, by the IEEE and with a number of citations greater than 500.

```

1 paper_df.filter((col("publisher") == "IEEE") & (col("n_citation") >
2   500)) \
3   .groupBy("year", "doc_type") \
4   .agg(count("year").alias("n_years")) \
5   .orderBy(desc("year")) \
6   .show(20, truncate = False)

```

Execution time: 397 ms ± 79.7 ms

year	doc_type	n_years
2018	Journal	1
2017	Journal	2
2016	Journal	2
2015	Journal	5
2014	Journal	14
2013	Journal	12
2012	Conference	2
2012	Journal	11
2011	Journal	23
2011	Conference	1
2010	Journal	18
2010	Conference	1
2009	Conference	2
2009	Journal	16
2008	Journal	14
2008	Conference	2
2007	Journal	18
2007	Conference	3
2006	Journal	14
2005	Journal	14

6. Retrieve the publishers that have more than 100 papers stored in the database.

```

1 paper_df.groupBy("publisher") \
2   .agg(count("id").alias("n_papers")) \
3   .filter(col("n_papers")>=100) \
4   .show(truncate = False)

```

Execution time: 429 ms ± 167 ms

publisher	n_papers
Society for Information Management and The Management Information Systems Research Center	105
Elsevier	565
Elsevier Science Inc.	193
Elsevier Science Publishers Ltd.	108
ACM	1104
Kluwer Academic Publishers	277
Springer US	247
IEEE	2092
Springer-Verlag	163
IEEE Press	142
IEEE Computer Society	433
Elsevier Science Publishers B. V.	272
Pergamon	131

7. Retrieve the years in which the number of published paper of journals is greater than 150.

```

1 paper_df.filter(col("doc_type") == "Journal") \
2 .groupBy("year") \
3 .agg(count("id").alias("n_papers")) \
4 .orderBy(asc("year")) \
5 .filter(col("n_papers") > 150) \
6 .show(truncate = False)

```

Execution time: 380 ms ± 87.2 ms

+-----+ year n_papers
1998 158
2000 187
2001 181
2002 231
2003 309
2004 296
2005 421
2006 422
2007 464
2008 485
2009 505
2010 559
2011 574
2012 478
2013 466
2014 323
2015 188

8. Compute the total number of citations, for each publisher, of the papers that have a field of study with weight grater than 0.6.

```

1 paper_df.select(col("id"), col("title"), col("year"),
2   col("n_citation"), col("doi"), col("publisher"),
3   explode(paper_df.fos_id)) \
4 .withColumnRenamed("col", "fos") \
5 .filter(col("fos") \
6   .isin(fos_df.filter(col("weight") > 0.6) \
7   .select(collect_list("id")) \
8   .collect()[0][0])) \
9 .groupBy(col("publisher")) \
10 .sum("n_citation") \
11 .withColumnRenamed("sum(n_citation)", "total_citations") \
12 .show(20, truncate = False)

```

Execution time: 2.02 s ± 135 ms

publisher	total_citations
Cambridge University Press	374
捷頂科技有限公司	112
Wiley-Blackwell	478
John Wiley & Sons, Ltd.	345
Elsevier	2339
Kluwer Academic Publishers-Plenum Publishers	463
Decision Support Systems	136
Elsevier Science Inc.	678
MIT Press 238 Main St., Suite 500, Cambridge, MA 02142-1046USA journals-info@mit.edu	412
IEEE Computer Society Press	878
John Wiley & Sons, Inc.	248
John Wiley & Sons	155
Elsevier Science Publishers Ltd.	137
ACM	15492
Springer International Publishing	122
Academic Press, Inc.	642
North-Holland	943
Kluwer Academic Publishers	5407
Springer US	199
IEEE	11892

9. Retrieve the authors whose organization is the Stanford University and that wrote a number of papers greater than 1.

```

1 paper_df.select(col("id"), col("title"), col("year"),
2   col("n_citation"), col("page_start"), col("page_end"),
3   col("doc_type"), col("publisher"), col("volume"), col("issue"),
4   col("doi"), col("references"), col("abstract"), col("venue_id"),
5   explode(col("authors_id")), col("fos_id")) \
6 .withColumnRenamed("col", "authors_id") \
7 .withColumnRenamed("id", "paper_id") \
8 .join(author_df, author_df.id == col("authors_id"), "full") \
9 .filter(col("org") == "Stanford University") \

```

```

6   .groupBy("authors_id") \
7   .agg(count("paper_id").alias("n_papers")) \
8   .orderBy(asc("n_papers")) \
9   .filter(col("n_papers") > 1) \
10  .show(truncate = False)

```

Execution time: 837 ms ± 137 ms

authors_id	n_papers
2149433985	2
201828038	4
348630313	4

10. Retrieve the venues on which an organization has published 4 or more papers, that consist of more than 10 pages.

```

1 paper_df.filter((col("page_end") - col("page_start") > 10)) \
2   .select(col("id").alias("paper_id"), col("venue_id"), col("volume"),
3         → explode(paper_df.authors_id).alias("author_id")) \
3   .join(author_df, author_df.id ==
3         → col("author_id")).drop("id").withColumnRenamed("name",
3         → "author_name") \
4   .join(venue_df, venue_df.id ==
4         → paper_df.venue_id).drop("id").withColumnRenamed("name",
4         → "venue_name") \
5   .groupBy(col("venue_id"), col("venue_name"), col("org")) \
6   .agg(count("paper_id").alias("n_papers")) \
7   .filter((col("n_papers") >= 4) & (col("org").isNotNull())) \
8   .sort(desc("n_papers")) \
9   .show()

```

Execution time: 1.05 s ± 91 ms

venue_id	venue_name	org	n_papers
134177497	IEEE Transactions on Space Control & Instrumentation	Space Control & Instrumentation	9
134177497	IEEE Transactions on Department of Materials	Department of Materials	7
1127352206	programming languages at Microsoft Research	Microsoft Research	5
157921468	ACM Computing Surveys (CS)	Columbia Univ., NY	5
1152462849	acm special interest groups	University of Illinois Urbana-Champaign	5
168680287	IEEE Transactions on Technion - Israel Institute of Technology	Technion - Israel Institute of Technology	4
57293258	Management Information Systems at Fox School of Business	Fox School of Business	4
90119964	ACM Transactions on Information Systems (TOIS)	Univ. of Maryland	4
25538012	International Journal of Computer Science and Technology (IJCST)	Stanford University	4
63459445	IEEE Transactions on Dept. of Electrical Engineering	Dept. of Electrical Engineering	4
199944782	IEEE Transactions on National Laboratory of Pattern Recognition	Nat. Lab. of Pattern Recognition	4
414566	Pattern Recognition at Department of Computer Science	Department of Computer Science	4
93787993	IEEE Transactions on Department of Computer Science	Department of Computer Science	4
8351582	IEEE Transactions on Lane Department of Computer Science	Lane Department of Computer Science	4
80113298	Journal of the Association for Computing Machinery (JACM)	University of Washington	4
63392143	Computer Networks at Broadband and Wireless Communications	Broadband and Wireless Communications	4

3.3.4. Creation/Update Commands

Such commands can be found and executed in the `import_data_and_queries.ipynb` script in the `spark` section of the GitHub repository.

1. Drop the column year from the paper data frame.:

```
1 paper_df.drop(paper_df.year) \
2 .show()
```

Execution time: 195 ms ± 39.7 ms

id	title	n_citation	page_start	page_end	doc_type	publisher	volume	issue	doi	references	abstract
101421652	The influence of organizational culture on...	139.0	397	423	Journal	Society for Information Management	27	3	10.2307/30036539	[1516261653, 1978...]	Managers in modern organizations
1015675232	Research paper re-use	106.0	305	338	Journal	Springer Berlin Heidelberg	17		410.1007/s00799-01... [1971040550, 2012...]	In the last 16 years...	
10311529	Technical Section...	236.0	85	103	Journal	Pergamon Press, Inc.	33		1 10.1016/j.cag.200... [2075597533, 2118...]	User interfaces in...	
104754383	Shackled to the system? An empirical study of...	215.0	21	42	Journal	Society for Information Management	36	1	10.2307/41410404	[1758167269, 1918...]	[Given that adopti...
108157922	The state of the art in business intelligence: A...	114.0	141	178	Journal	Kluwer Academic Publishers	7	3 10.1023/A:1008287...	null	This paper develops...	
112035675	Understanding fit: A review and synthesis of...	414.0	167	193	Journal	Society for Information Management	25	2	10.2307/3259928	[1595507819]	Many previous pap...
115183765	Technotress: technology and its social conseque...	287.0	831	858	Journal	Society for Information Management	35	4	10.2307/41409963	[1510394286, 1721...]	With the proliferati...
117893765	Survey paper: A survey of business intelligence...	273.0	510	522	Journal	Elsevier Science Publishing Company	56	6 10.1016/j.comind... [2041748119, 2171...]	The research literature...		
127096639	Business intelligence systems: A critical review...	154.0	1109	1216	Journal	Society for Information Management	36	4	10.2307/3259930	[1565251030, 1565251031]	The surveying po...
125324109	Positioning and optimization of business intelligen...	215.0	245	255	Journal	Kluwer Academic Publishers	14	3 10.1023/A:1011209... [1538585633, 1572...]	Computational methods...		
1270866904	Positioning and p...[redacted]	665.0	337	356	Journal	Wiley Research Center	37	2 10.2308/MSQ-2001-... [1532738203, 1565...]	Design science re...		
127800959	Coherence-Enhancing Ontology Alignment	594.0	111	127	Journal	Kluwer Academic Publishers	31	2 10.1023/A:10980009...	[197269179]	The completion of...	
134908229	The adoption and ...[redacted]	154.0	289	323	Journal	Society for Information Management	27	2	10.2307/3003652	[1212035675, 15955...]	[This paper report...
136147141	Ontology alignment: A survey	162.0	158	192	Journal	Springer, Berlin, Heidelberg	15	6720 10.1007/978-3-642... [2015191210, 2047...]	In the area of se...		
13953467	Password Authentication	112.0	101	115	Journal	捷通科技有限公司	3		2 10.6633/IJNS.2006...	null	[Password authenti...
141544107	Interactive decision support systems: A review	115.0	293	320	Journal	Society for Information Management	33	2	10.2307/20650293	[1511558225, 1576...]	[This paper extend...
1422679120	Interval-valued intuitionistic fuzzy sets	101.0	183	191	Journal	Elsevier	248	1 10.1016/j.ejor.20... [2002574279, 2009...]	[Inter-dependency ...]		
1424218399	A review of definitional equivalence	125.0	47	61	Journal	Elsevier	145	1 10.1016/j.res.20...	[1980156780]	[Modeling and evalua...	
145670326	Survey paper: A survey of wireless network...	161.0	940	965	Journal	Elsevier North-Holland	56	2 10.1016/j.comnet... [2149863032, 2150...]	Wireless networki...		
1480498941	Automatic detection of...	154.0	63	77	Journal	Kluwer Academic Publishers	32	1 10.1023/A:1098145...	[2134864374]	[This paper demons...	

2. Drop the columns page_start and page_end from the paper data frame.:

```
1 paper_df.drop("page_start", "page_end") \
2 .show()
```

Execution time: 181 ms ± 38.6 ms

3. Insert an author in the author data frame.:

```
1 newRow = spark.createDataFrame([(1882405, "Elon Musk", "Tesla")],  
2     ["id", "name", "org"])  
3  
3 new_df = author_df.union(newRow)  
4 new_df.filter(new_df.name == "Elon Musk").show(truncate = False)
```

Execution time: 436 ms ± 29.2 ms

4. Update the title of a paper with a particular id.:

```
1 new_value = "New_Title"  
2 new_paper_df = paper_df.withColumn(  
3     "title",  
4     when(  
5         col("id") == 101421652,  
6         new_value  
7     )  
8 )  
9  
10 new_paper_df.filter(new_paper_df.id == 101421652).show()
```

Execution time: 250 ms ± 30.7 ms

5. Update the title of all the papers with more than 100 pages.:

```
1 new_value = "New_Title"  
2 new_paper_df = paper_df.withColumn(  
3     "title",  
4     when(  
5         col("page_end") - col("page_start") > 100,
```

```
6         new_value
7     )
8 )
9
10 new_paper_df.filter(new_paper_df.title == new_value).show()
```

Execution time: 188 ms ± 24.7 ms

4 | References

- AMiner data set V11
- Bio data set
- Tweets data set
- Images data set
- Draw.io
- Neo4j
- MongoDB
- Apache Spark
- Overleaf
- Python
- PyCharm
- Google Colaboratory