Report

I have chosen to work on Project 2: The PubMed literature database.

DATA STRUCTURES

To answer the question I created different kinds of data structures. I did this by creating functions that loop over the xml files and then save the obtained data to pickle files. I applied these functions using multiprocessing.

Author – Paper dataframe

The first data structure I created was an author paper dataframe. Where one column contains all the authors and co-authors and the other column contains the respective PubMedID that they have written. (see figure 1)

I created this dataframe by combining all the pickles created from parsing the xml files into one Dask dataframe.

The required XML parser can be found in *pickle_maker.py.* The function that creates the Dask dataframe can be found in *question1.py*

	0	1
0	3624245	E D Green
1	3624245	J U Baenziger
2	3624246	E D Green
3	3624246	R M Brodbeck
4	3624246	J U Baenziger
5	3624247	D F Smith
6	3624247	P A Prieto

Figure 1: Part of author-paper dataframe

Author – Paper graph

Using the same pickle files I also created a bipartite graph which connects authors to papers they wrote and vice versa. Within the graph there are 72759 individual subgraphs. With the largest subgraph containing 95% of the full network nodes.

Most graph theory functions only work on connected graphs and I decided that looping over 72759 individual graphs just to obtain 5% more data was not worth it. So I only used the largest subgraph and did not look at the others.

This function and additional information can be found in *Assignment6.ipynb*

Paper/References/Authors/Keywords dataframe

This dataframe has been created in the same way as the Author-paper dataframe; by parsing the xml files and saving them to pickles. To then combine the pickles into one Dask dataframe. Its dataframe can be seen in the figure below.

	pubmedID	Authors	Refs	keywords
75	5969679	[F Forster]	[13918530, 14255452]	[Adult, Demography, Epidemiology, Female, Huma...
76	5969680	[G R McNab, E J Slator, C J Stewart]	[13364352, 13700260, 14401755]	[Bronchitis, Chronic Disease, Cough, England, ...
77	5969681	[J W Palmer]	[14171095]	[Adolescent, Adult, Aged, Communication, Human...
80	5969684	[P Dukor, F M Dietrich, M Rosenthal]	[5320432, 5858699, 5932903, 5330737, 4958328, ...	[Animals, Antibody Formation, Female, Hemolysi...
81	5969685	[R K Main, L J Cole]	[14088079, 14107045, 14113075, 14162535, 14162...	[Animals, DNA, Female, In Vitro Techniques, Ma...

Figure 2: Part of the P/R/A/K dataframe

The xml parser I used for this dataframe can be found in *pickle_maker_2.0.py*

Citation graph

In addition to a author – paper graph I also created a Citation graph. Where each paper is connected to a paper they reference. This graph has attributes; either the authors that have worked on the paper or the keywords of the paper. The code can be found in *citation_graph.py*

QUESTIONS

In this chapter I will answer two of the given questions and explain how I answered them. And for the questions that I didn't answer I explain how I thought about answering them.

There are two main reasons why I did not answer all the questions:

1. The questions are very difficult. After researching online I have found that people have written entire papers on answering even one of these kinds of questions. And answering them will sooner take two months than two days.
2. The Networkx library is very slow. So far the Networkx library is the only one installed on the school computers and it is the slowest of all libraries. And some of these questions would require a lot of computational power, because the graphs are absolutely massive.

The questions that I did answer are:

How large a group of co-authors does the average publication have?

I used the created dataframe with the IDs and the authors. And I divided the total authors with the amount of unique PubMedIDS to give me an answer of **4.2 average authors per paper.**

Is there a correlation between citations and the number of keywords that papers share? I.e. papers which share the same subject cite each other more often.

Yes there is a correlation. On average 21% of the keywords are shared between cited papers, while this is only 6% for the papers in general. This is a significant difference.

To obtain this answer I used the citation graph with the keywords as attributes. I looped through the edge list, and for both nodes in the list, I obtained a list of keywords. I then calculated the percentage of matching keywords and at the end averaged this for the whole network.

The method with additional comments can be found in *question5.py*

Own Question: What % of papers does not have a reference list

A lot of papers seem to be lacking a reference list so I wonder what % this is. I can answer it by dividing the amount of papers with a reference list by the total amount papers * 100.

6776743 / 31850929 * 100 = ~20%.

Meaning that only 20% of the papers include a reference list.

Thoughts on I would answer the other questions:

Do authors mostly publish using always the same group of authors?

To answer this question I wanted to use a modularity function. As this is the most common method used for finding communities (Newman, 2006). If there are few communities or very large communities it would mean that authors do not always use the group. If there are a lot of small communities it would mean that authors often use the same group of authors.

Finding modularity is difficult, especially for large scale graphs. I did find an article: *Community Detection in Large-Scale Bipartite Biological Networks* (Calderer & Kuijjer, 2021). Which lists multiple methods including non modularity methods. The one thing these methods have in common is that they will take a lot of work, both from me as well as in computational power. Therefore, I decided not to answer this question.

Do authors mainly reference papers with other authors with whom they've co-authored papers (including themselves)?

I found an article: *A small world of citations? The influence of collaboration networks on citation practices* (Wallace, Larivière, & Gingras, 2006) This article examines the proximity of authors of those they cite using degrees of separation.

I thought about answering this question with the citation graph + using the authors as attributes. But I could still not figure out how I could use this to answer the question exactly.

What is the distribution in time for citations of papers in general, and for papers with the highest number of citations? Do they differ?

The XML file does not give me information regarding how often and on what time a specific paper has ben cited as far as I could tell.

For the most-cited papers (define your own cutoff), is the correlation in shared keywords between them and the papers that cite them different from

This question should not be to difficult because I have already answered question 5. To find the most cited papers I could transform the citation graph into a directed one and calculate the indegree of the nodes.

However this would require me to change everything again and answering question 5 was difficult enough.

OBSERVATIONS

Chapter with interesting/useful information that I found.

After created the author – paper graph I checked for the nodes that have the most connections/degrees. This will either show the author with the most papers or the paper with the most authors. What I found was the results shown in figure 2. It shows Authors with more than 10 000 papers. This makes it seem like they are the busiest writers in the world. Which Is most likely not true as they are also all very common Chinese names, meaning that there are probably a lot of different authors publishing under that name. However the PubMed xml files contain no unique author id as far as I could tell so there is no way to remove/split duplicate names from the dataset.

```
[('Wei Wang', 15254),
 ('Wei Zhang', 13897),
 ('Wei Li', 11040),
 ('Jing Wang', 10499),
 ('Yan Li', 10418),
 ('Yan Wang', 9503),
 ('Jing Li', 9494),
 ('Lei Zhang', 9405),
 ('Yang Liu', 9296),
 ('Yan Zhang', 9282)]
```

Figure 3: Nodes with the most degrees.

The code for the observations can be found in assignment6.ipynb

REFERENCES

Calderer, G., & Kuijjer, M. L. (2021). Community Detection in Large-Scale Bipartite Biological Networks. *Applications and Methods in Genomic Networks* .

Newman, M. E. (2006). Modularity and community structure in networks. *PNAS*.

Wallace, M. L., Larivière, V., & Gingras, Y. (2006). A small world of citations? The influence of collaboration networks on citation practices. *Science and Engineering Indicators 2006*.