Report

# Introduction

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

I think the assignment is very interesting but I started a bit late so I am not sure if I will be able to finish it before July 11th. However I do only need a 4 to pass this course, so maybe this unfinished report will be enough for that and then I will hand in a resit during/after the summer holydays. I also won’t need the 5 EC so a <4 would also be fine.

# Questions

In this chapter I will answer all the given questions and explain how I answered them.

## Question 1:

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

Graphical user interface

Description automatically generated with medium confidenceTo answer this question I did not use any graph theory. I just extracted all the authors mentioned in the xml files, and the PubMedID of which article they wrote. See figure:1

To do this I looped over all the xml files and saved the authors and IDS to a pickle file using multiprocessing. The exact code of how I did this can be found in *pickle\_maker.py*

After creating a 1065 pickle files, I had to combine them. I did this by loading in each pickle as a pandas dataframe, and then I load in each pandas dataframe as a dask dataframe. (pickle to dask support when?) I then combine all the dask dataframes into one.

And after I have one dataframe I simply devide the total length of the frame and thus the total authors with the amount of unique PubMedIDs.

This gives me the answer of 4.2 average authors per paper.