Exercise 3 - CORD-19 Summary

- Alon Moses 308177815
- Guy Attia 305743437

Data Preprocessing

At first, as you asked, we loaded all the data from the metadata file, and used the only newest 20k papers as follows:

1. Unzip the zip file
2. Load the original metadata csv file
3. Extract the latest 20K papers ID's from the metadata
4. Save the metadata of the 20K papers in a new csv file

Similarity Method

Compression Algorithm

The compression algorithm we used is the built in Python3 package 'lzma' (Lempel–Ziv–Markov).
The LZMA algorithm is an algorithm used to perform lossless data compression, and it uses a dictionary compression scheme.

Distance Function

To calculate the distance between two compressed papers we used the methodology we learned in class - Normalized Compression Distance (NCD).

Papers Similarity Solutions

We tested 2 different hypothesises:

1. Papers compression from the same journal will be relatively closer to each other.
2. Papers with joint bibliography references will be relatively closer to each other.

Journals Effect

Our hypothesis is that compressed papers from the same journal will be relatively closer to each other rather than compressed papers from different journals.
To assess our hypothesis we performed the following:

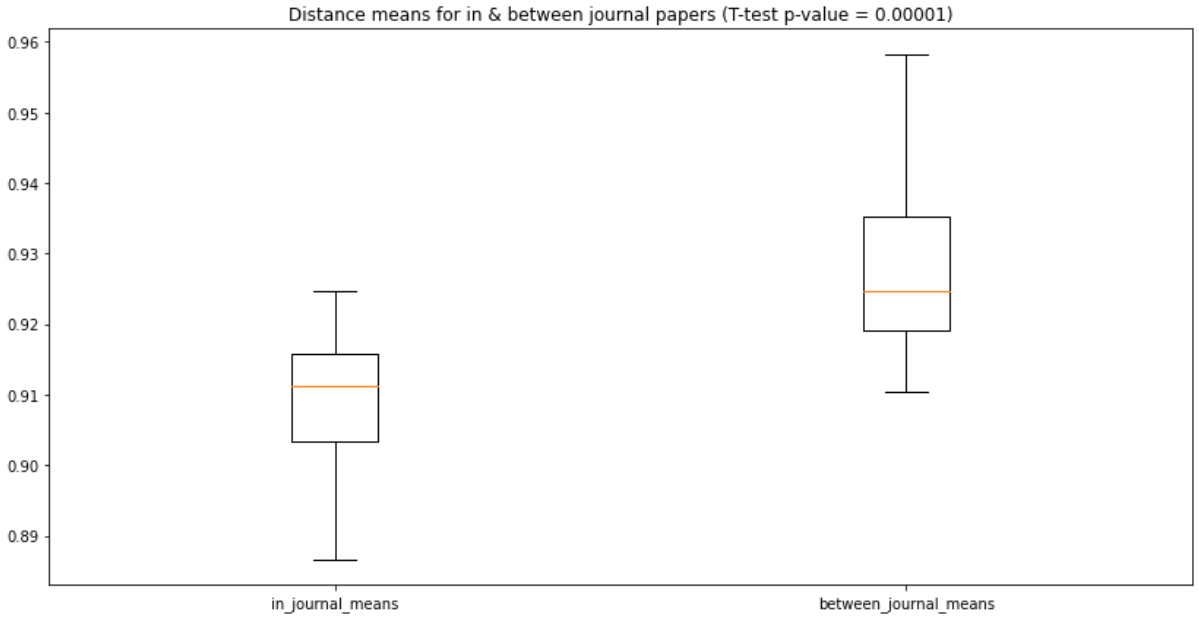
1. Filter papers from specific journals
2. Calculate the distances of the papers within the same journals (for each one)
3. Calculate the distances of the papers from different journals (for each combination)
4. Compare and analyze the distances results

Papers Filtering

Due to the long running times and the unnecessary need of comparing the entire papers combinations, we filtered the dataset by picking only papers from 10 specific journals containing around 27 papers each.

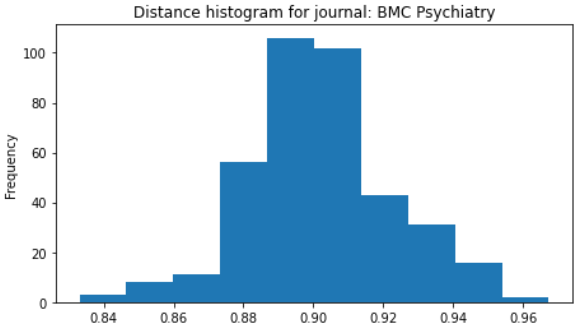
Results Analysis

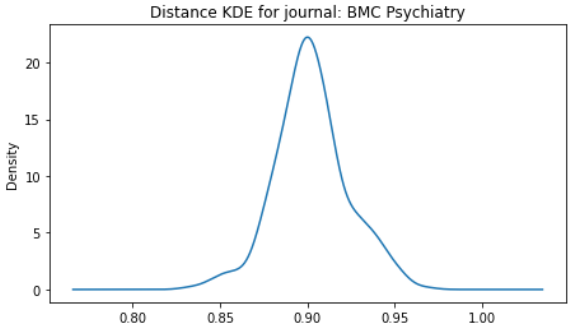
As we assumed, it's easy to see in the boxplot that the average distance between papers from the same journal is significant (p_value < 0.05) lower than distance between papers from different journals.



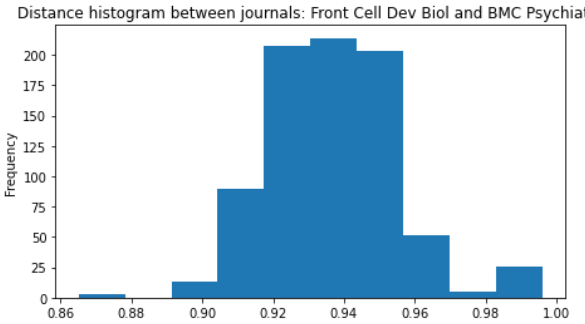
You can find in the attached Jupyter-notebook the detailed distribution plots.

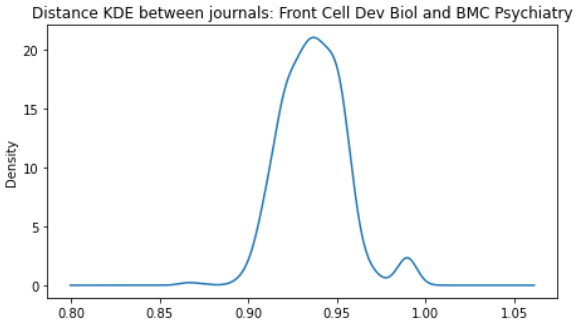
- "In_journal" example: The distance distribution between papers inside a specific journal ("BMC Psychiatry"):





- "Between_journals" example: The distance distribution between papers from two different journals ("BMC Psychiatry" and "Front Cell Dev Biol"):





Bibliographic References Effect

Our hypothesis is that compressed papers which reference joint bibliography will be relatively closer to each other compared to paper with no joint bibliography references at all.

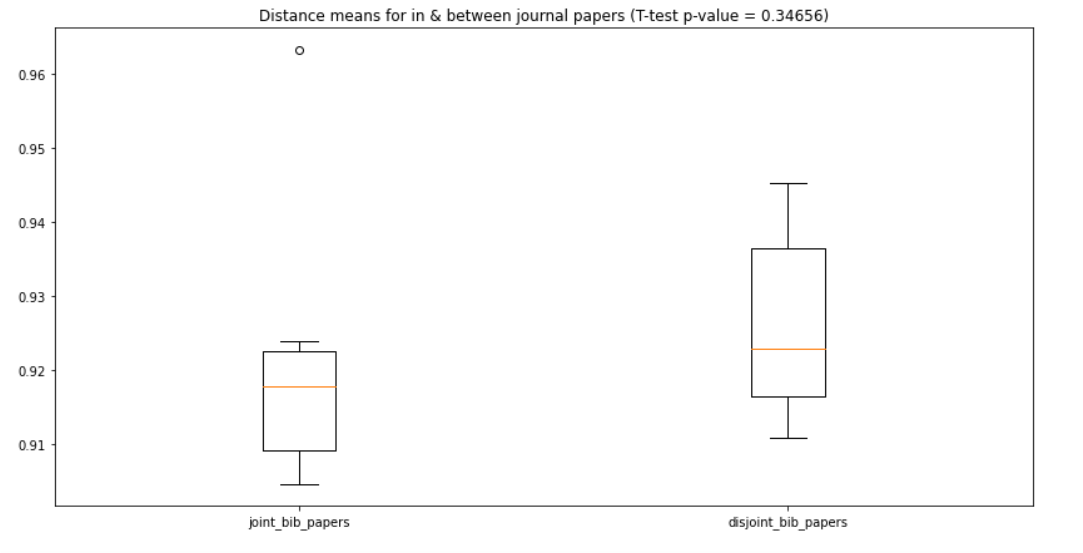
Papers Filtering

To prepare the data we followed this procedure-

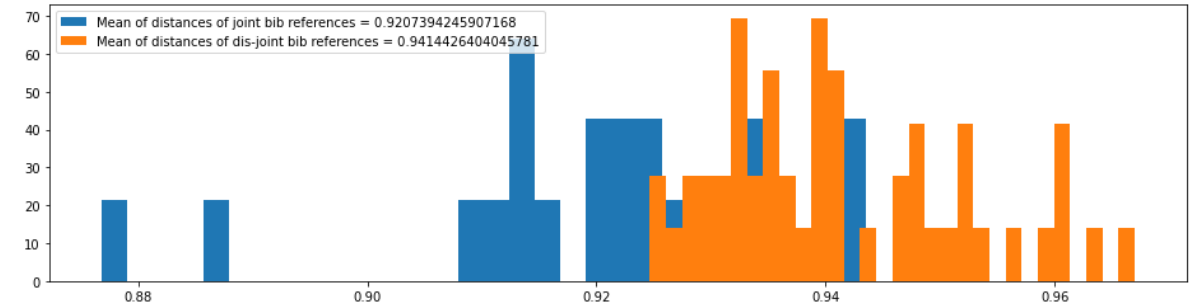
- Create a new bib_entries dictionary which contains for each of the files a list of corresponding files in the dataset. The files in this list will follow the below assumptions-
    1. The publish time distance between the files will not exceed 100 days.
    2. The files will have at least one common bibliography entry (reference to other paper, story, etc.)
- From the bib_reference dictionary create the following 2 types of distance calculations and contain them in dictionaries-
    1. Create a dictionary with sha's use as keys, pointing to a list of distances values between the paper in the key and the papers appearing in the bib_reference.json lists.
    2. For the same files used for calculating distance in section 1, calculate their distance with files which have no joint bib references at all.

Results Analysis

In this case, we see that the average distance between papers which reference to similar bibliography isn't significant (p_value >> 0.05).
But we still can observe some lower distance between papers which reference to similar bibliography compared to ones doesn't have any joint bib references



- Example of a histogram showing
    1. The distances between individual paper to other papers containing joint bibliography references (blue).
    2. The distances between individual paper to other papers containing joint bibliography references (orange).



** You can find in the attached Jupyter-notebook the detailed distribution plots.

It can be seen here and also in the other examples in the notebook, that the distances between a paper to other papers with joint bibliography reference are lower than the distances of the same paper to other papers without any joint bibliography reference.