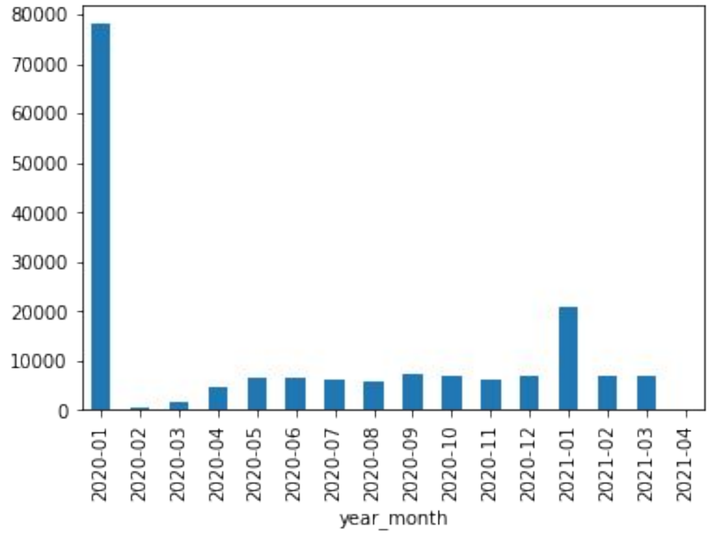
Scholarly data analysis

Pai Peng

Scholarly data overview

- There are 497906 articles in total. With 171262 of them published after the outbreak and are related to COVID-19 matters.

Over 70000 papers are published immediately after the outbreak in China during January. It's quite surprising how fast western world scholars in biomedical field react to such event.



From exploring the key words, there appears to be a wide range of interests in the scholar discussion about such event. Artificial intelligence, similarity to SARS, originating path and etc. It could be beneficial to catalogue them to see progress in fighting the disease from different aspects.

Everything about Modelling

- The target is to classify the scholar articles into categories, by the result one can see the progress of how each different aspect of the pandemic is analyzed/learned and the information can be helpful to several parties for macro decision making.
- Therefore, clustering methods are adopted. To be more specific, the job is done by unsupervised K-Mean algorithm with TF-IDF transformed and truncated SVD applied input data.

Why unsupervised method?

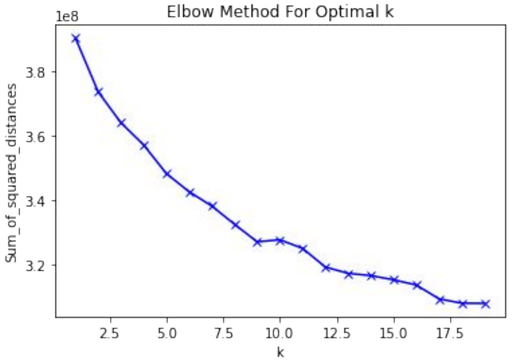
The data isn't labelled and in fact, one of the main results of the project is to label the data without actually knowing what potential labels might be. This falls into the category of clustering and most clustering methods are unsupervised.

Why truncated SVD?

For a large data like this, dimensionality reduction is urgently needed but a standard PCA would explode memory with such sparse matrix. After some research, I found that Truncated SVD does not cause the same problem and is commonly used in the industry to address similar matter.

About the model choice and how is it tuned?

K-Means algorithm comes in handy when classifying unlabelled data and it takes in a hyper-parameter "k" that indicates number of clusters. I used **"Elbow method"** to find the optimal k value. Evaluation is based on the sum of squared distance and it gets smaller with larger k value(i.e, more clusters). However, we don't want crazy many clusters, we want the right amount of clusters that yields relatively low squared error. The plot is showing arm-like curve and the turning point/elbow point on arm is determined as optimal point. In this case, optimal k is 9.



Results

- **Numerical labels do not carry much information therefore wordcloud is utilized to see the image of each group.**
- **The data is classified into 9 groups. Group 2 is the biggest class with 102748 papers labelled in. Group 8 has the least amount of observations that their wordcloud visualizations are messy/uninformative.**
- **Around 2000 articles are likely in Spanish, this is around 1% of all COVID-19 papers.**
- **Below are some important/insightful group classified:**

Group 2 and 0:

Articles mainly focus on understanding COVID-19 from an epidemiology angle. Including study into disease syndrome, effect on respiratory system.

Group 2 word cloud

Group 4:

Papers in this group seem to be discussing more about mental illness related to the disease.

Group 4 word cloud

Group 6:

A less popular group, with majority of discussion revolving around methods that help preventing COVID-19 and protect oneself .

Group 6 word cloud

Insights and advices

- From a government perspective, it's good to see how much effort scholars are contributing to tackle the COVID-19 pandemic in a basic level (i.e to cure the disease). However, with all the attention to disease itself and to already infected patients, not much of meaningful discussion are about the means to prevent it from spreading at the first place. Preventing is just as important as curing but the methods government are adapting now (i.e mask wearing, lockdown) do not seem to be all that effective. The governments need more guidelines so they might want to push academic world to look into the mechanism behind all the prevention methods and try to improve them or even build up new protocols that are more practical.

- From a healthcare system point of view, there had already been tremendously much medical force devoted to the caring of existing patients. Now that with us being more informed about the disease and less patients in intensive care unit, the system probably want to consider allocating more forces to deal with the aftermath of the pandemic. For example, mental issues seem to be a big problem that are studied quite often together with the COVID-19 matters. Caring for those who aren't infected but have mental illness caused by either fearing the disease or long-term lockdown would be top priority to prepare for when the pandemic enters next stage.

- From a scholar point of view, since properties of disease itself and the curing/vaccine for it are already well-studied, one might want to consider researching into COVID-19 from angles other than epidemiology or virology. As mentioned above, learning psychological problem caused by the pandemic can be a good direction. Also one can learn about why prevention is perfectly done in some countries but not working so well in others(i.e learning reasons behind refusal to wearing mask, obligation of lockdown order and etc).

LAST BUT NOT LEAST, I believe we are almost to the point where the virus is beaten. Wish us all good luck and thank you to all whom helping each other to get through difficult time with peace and love!

Thank you!