Trust Lab Coding Exercise

Kevin Yiu-Wah Cheung, kevinyiuwahcheung@gmail.com

Duration: 6 hours, 22 Jan 2021, 3:30pm to 9:30pm est

1 Introduction

Find a list of 1000 pages (including url and months) from common crawl archive that discuss or are relevant to COVID-19's economic impact.

2 General Preprocessing

Step 1:
I downloaded wet.gz becuase it stores extracted plaintext from the data stored in the WARC. Then, I made use of warc library to write the content of each record payload and stored content and url information into a Pandas dataframe structure.
Step 2:
Detect if the content is written in English or not. I spent much time on this session because some libraries took too long to run. I tried 1)`langdetect`, 2)`spacy-langdetect` and 3)`fasttext` for detecting if majority content is written in English. I ended up using `fasttext` because it return the result within couple minutes runtime.
Step 3:
I used sklearn CountVectorizer to do feature extraction.

3 Approaches

3.1 LDA

The first approach is LDA. I set the n_components to 7 and I was hoping that there will be some "economy", "finance", "business" and other keywords appears in the topic words and then I can filter them out and do regular expression to find article containing "covid" keywords. In reality, it does not show up.

3.2 NMF

I plan to implement it but I am running out of time. The idea of it is similar to PCA, NMF takes advantage of the fact that vectors are non-negative.

3.3 Regular expression on both content and title

My third idea is directly calling regular expression on both title and content to find out if they contain words belonging to "covid" and "economy" category.

4 Conclusion

I am not able to meet the requirement within 6 hours. I should have make use of cloud computing or MapReduce/PySpark to speed up the whole computational process. Overall, this exercise is challenging and I did learn a lot from it.