Text Mining Midterm Project Brief – Team 8

COVID-19 Reporting on China & EU Countries, and Research Papers

魏志宇、呂嘉霖、李子源、陳俞瑄

Link to Code

Goal: **Compare the differences between news reporting on COVID-19 between EU countries and China**, and look into some research papers so as to **gain a sounder understanding about the pandemic**. **Word Cloud** will be the media for interpretation. Also, we wish to present decent query result via **Information Retrieval**, and display top 10 relevant news/researches based on query texts via bar charts.

Text Source: Collected 15 pieces of news on EU/31 pieces of news on China via **CNN**, and also collected 14 research papers on COVID-19 via **Nature**, **NCBI**, **Springer**, **The Lancet**, and **ScienceDirect**. **Word Cloud** will be generated under these 3 cases.

Sequence of Processing Our Text: Load in our data → Perform **NER** via **spacy** to detect multi-tokens candidates → **Tokenization** via **nltk** → Add **multi-token** based on NER via **MWETokenizer** → Turn token into **lower-case** → **English stopwords removal** → **POS tagging** based on **Universal Tagset** → Perform **feature selection: POS = Noun/Adj** → Generate **Word Cloud** with **customized mask.**

Word Cloud Result:

| Case I: CNN news on EU | Case II: CNN news on China | Case III: Researches on COVID-19 |

Findings: Reports on EU focus more on UK news, means to protect COVID-19 (PPE, Personal Protective Equipment), and workers right; Reports on China focus more on Trump (politics), origin of COVID-19, and emphasis on "many" confirmed cases; Researches at this phase focus more on infectiousness, transmission means, and symptoms of COVID-19.

Information Retrieval Steps: Input **query** texts → Textual weighting (**TF-IDF weighting**) through texts/bagwords lists → Compute **Cosine Distance** between documents and query texts → Display **top 10 cosine similarity score news/researches** via **bar charts.**
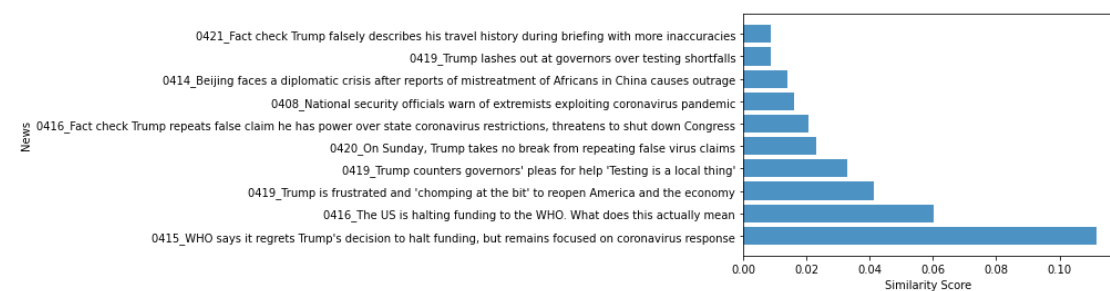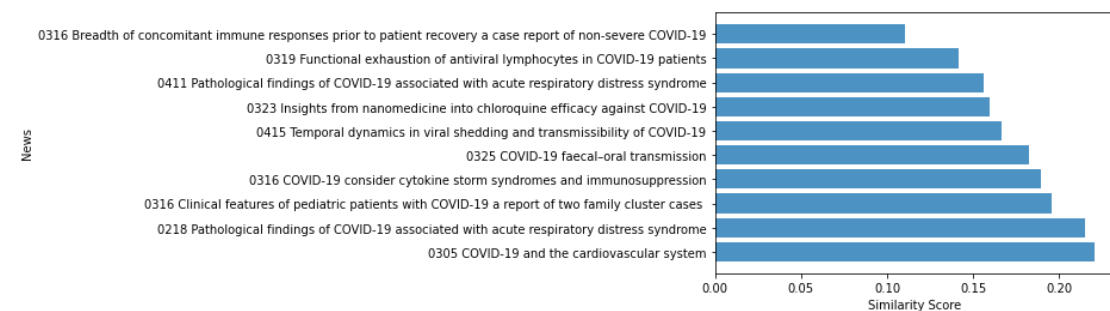
Case I: **CNN news on EU**

Case II: **CNN news on China**

Case III: **Researches on COVID-19**

Findings: Query results coincide some of our findings in word clouds.

Self-Evaluation & Future Development: For this project, we have not only practiced many techniques related to text mining, but also learned how to set our goal firmly, to find ways to achieve our goal, and to interpret results produced. We consider our achievement adequate for this phase, however, if time permits, we shall **collect more data** by a more efficient way (via **web crawler**), set up a **GUI** for user-friendly purpose (via **tkinter /ttk**), utilize **ML algorithms** to train our mode/detect sentiment, include **Rocchio Feedback** in IR, and improve query results. For the follow-up project, we plan to construct a GUI, from which user could input any sources they want, elevate accuracy and perform sentimental analysis based on ML algorithms, and improve query results in IR.

5 keywords: COVID-19, EU/China, PPE, Politics, Transmission