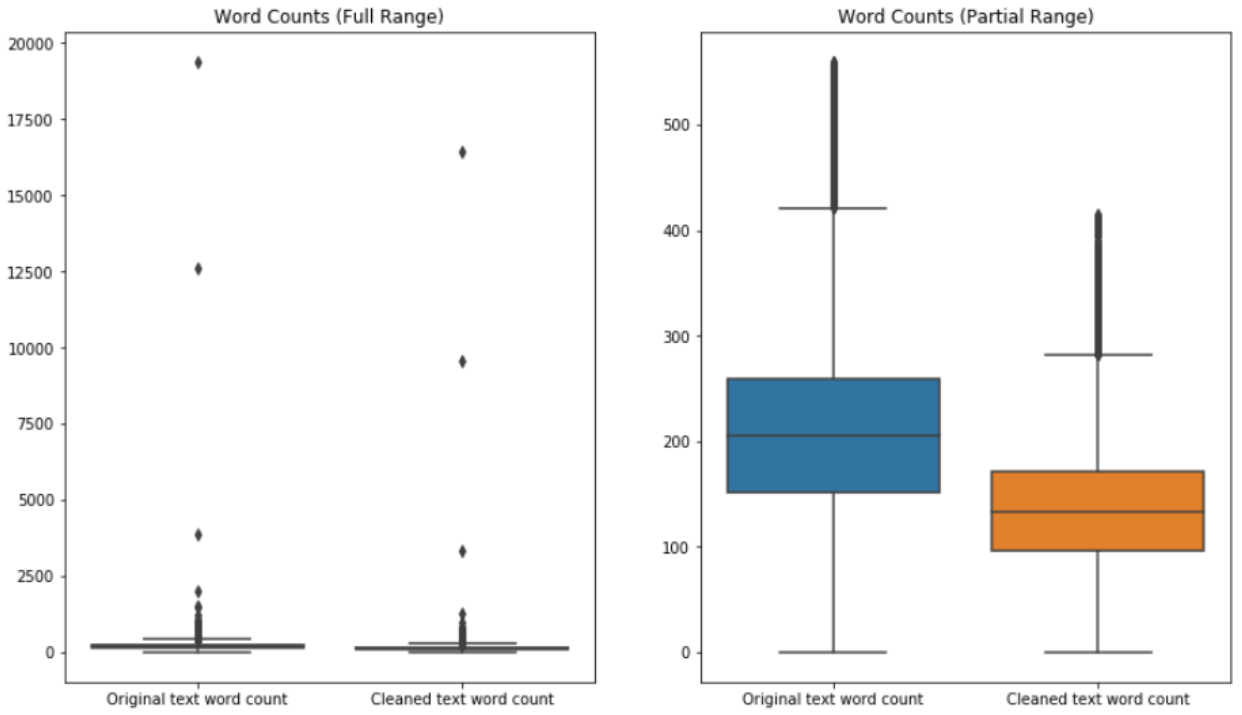
2. Exploratory Data Analysis

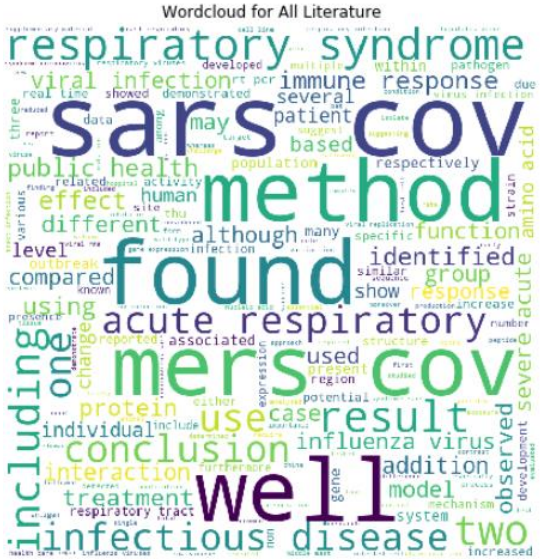
- To prepare for the **topic modeling**, the EDA part will be focused on 3 parts.
- Compare the word counts of original text with text after tokenization and cleaned.
- Visualize the word cloud of all database and literature only focused on COVID-19, to check whether there are differences.
- Choose vectorization methods by comparing the results of t-SNE of Word Frequency and TF-IDF.

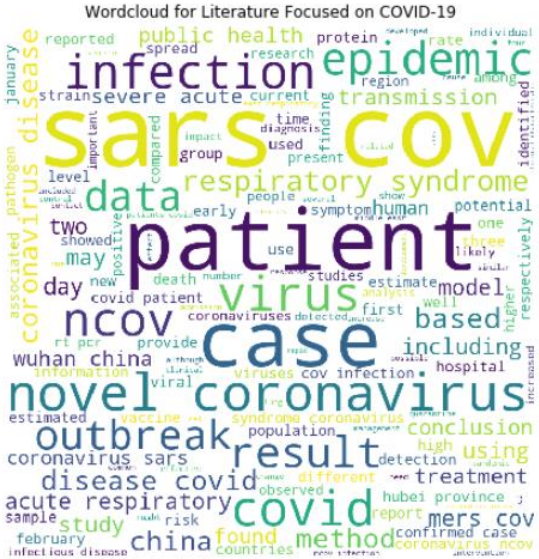
- **2.1 Comparison of word counts before and after removing stop words**



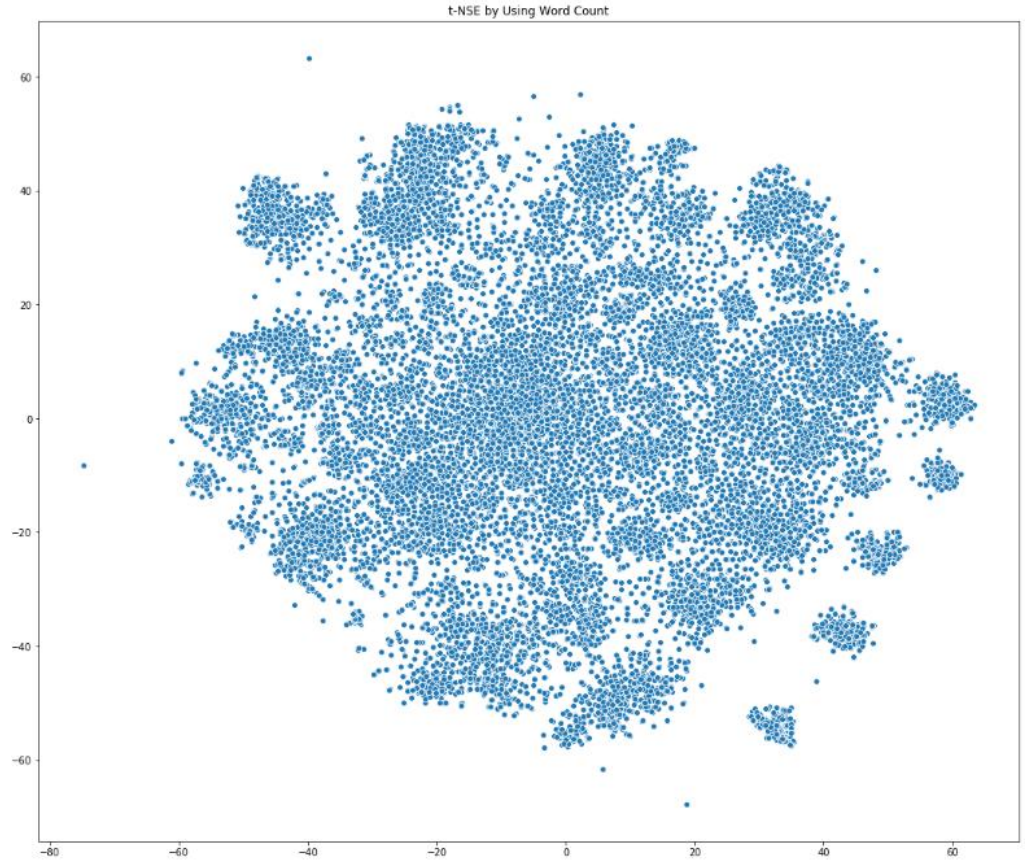
2. Exploratory Data Analysis

- **2.2 Word cloud of all literature and literature focused on COVID-19**





- **2.3 Check topic modeling potentials by t-SNE**



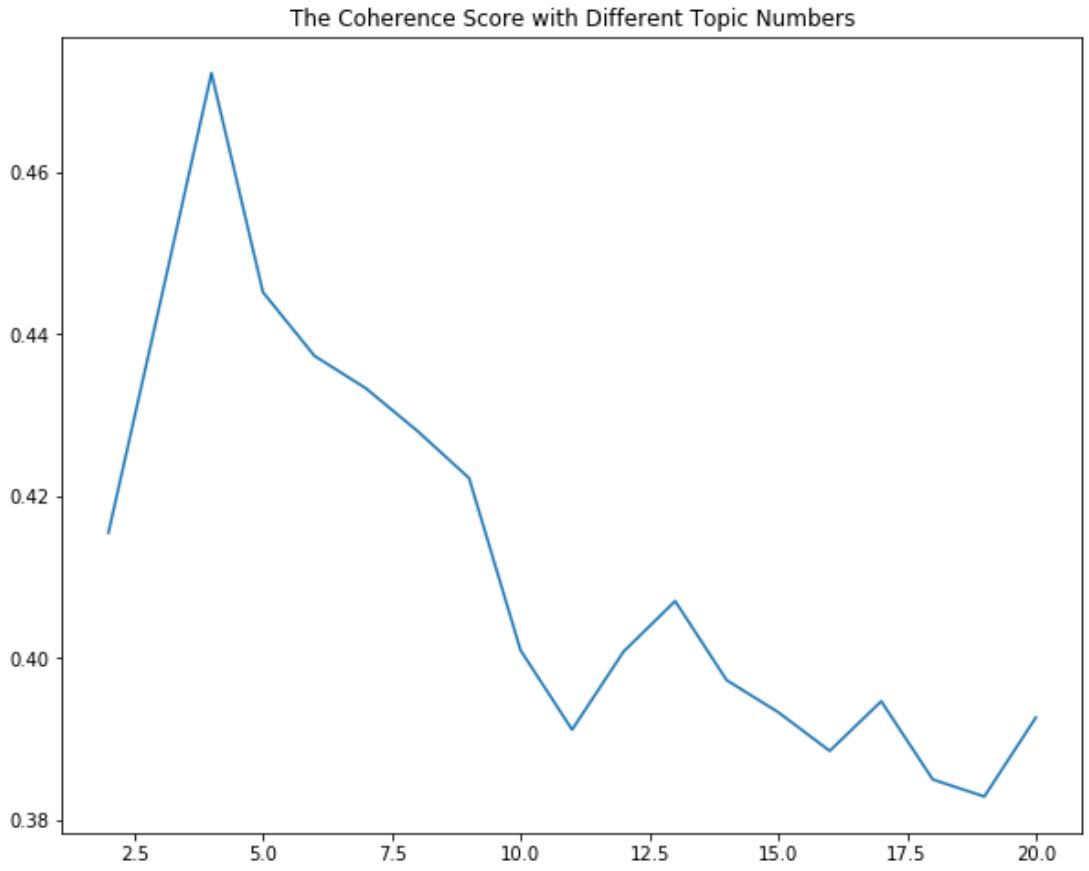
3. Model Selection and Training Process

•This report will use Latent Semantic Analysis (LSA) to model the topics of the literatures.

•The *reason* of using LSA is that it is the most commenly used topic modeling methods, also it is not computational and is suitable for the large dataset.

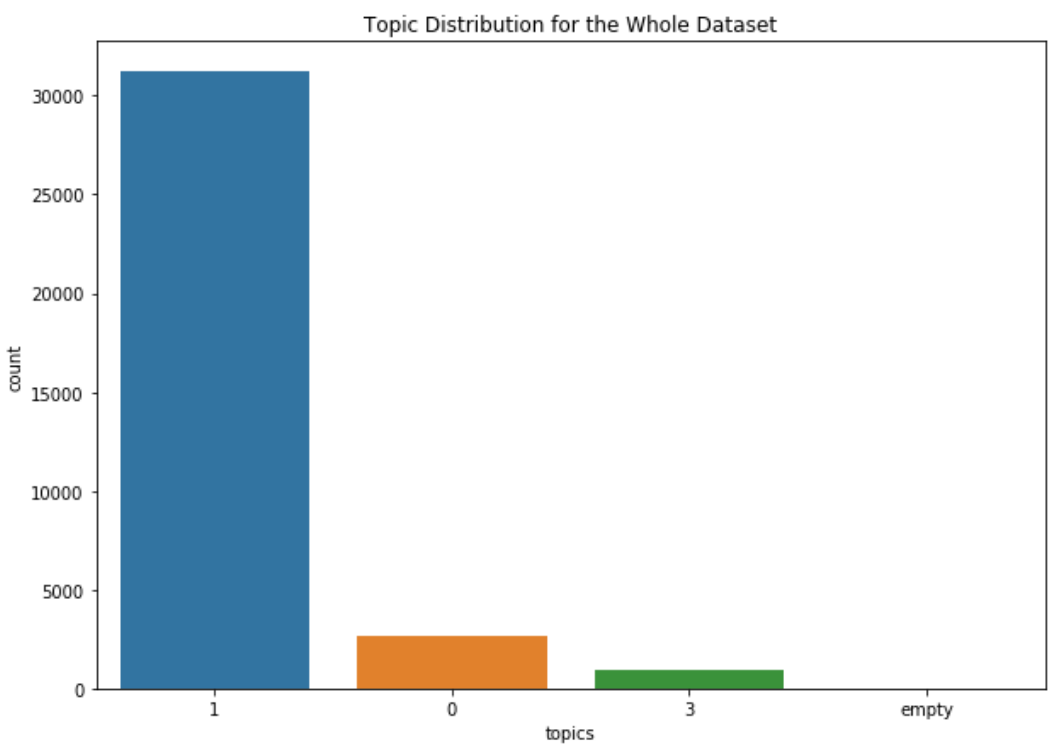
•The library will be used in gensim, the features will use Word Frequency.

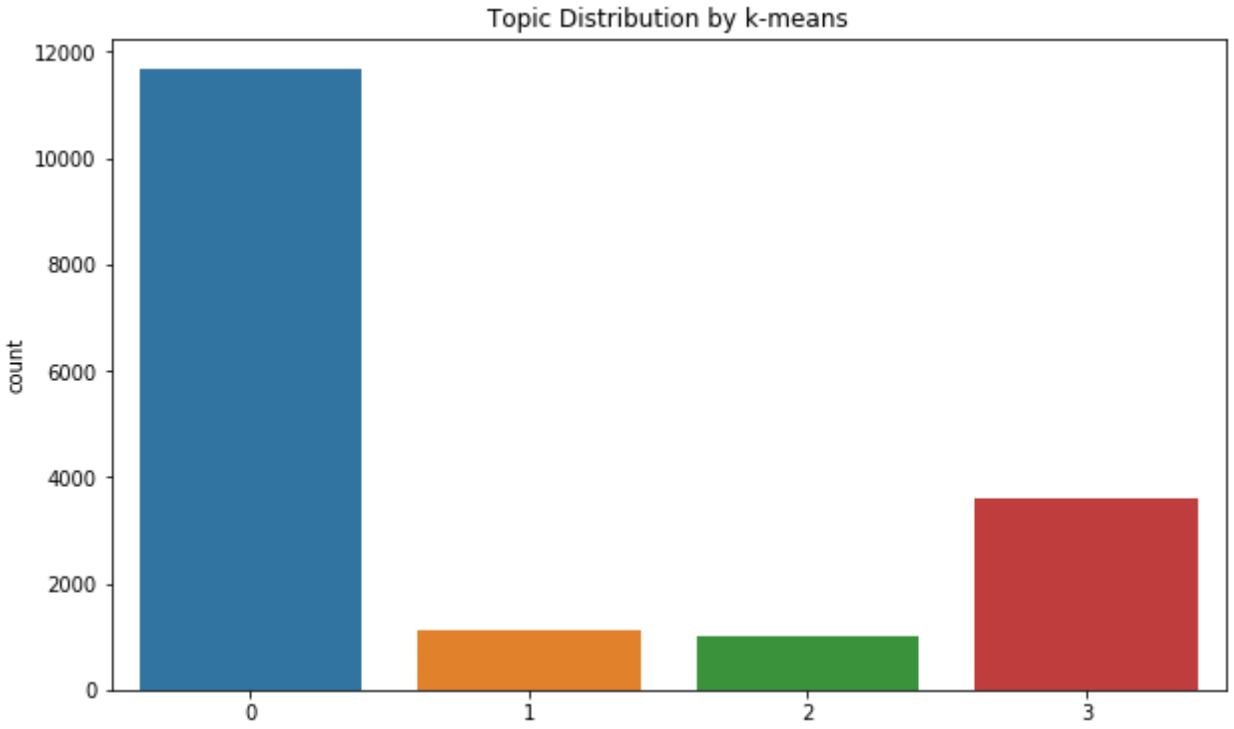
- **Model tuning process**



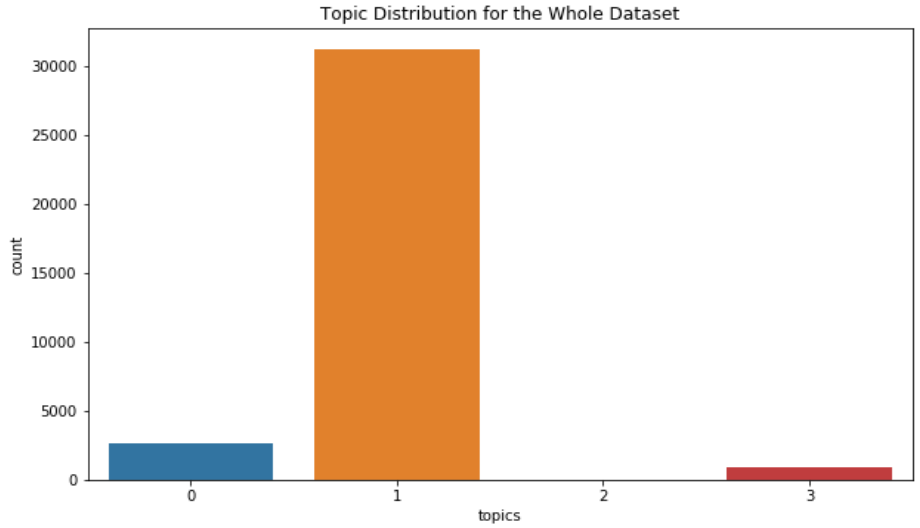
3.3 Verify the results by mini-batch k-means

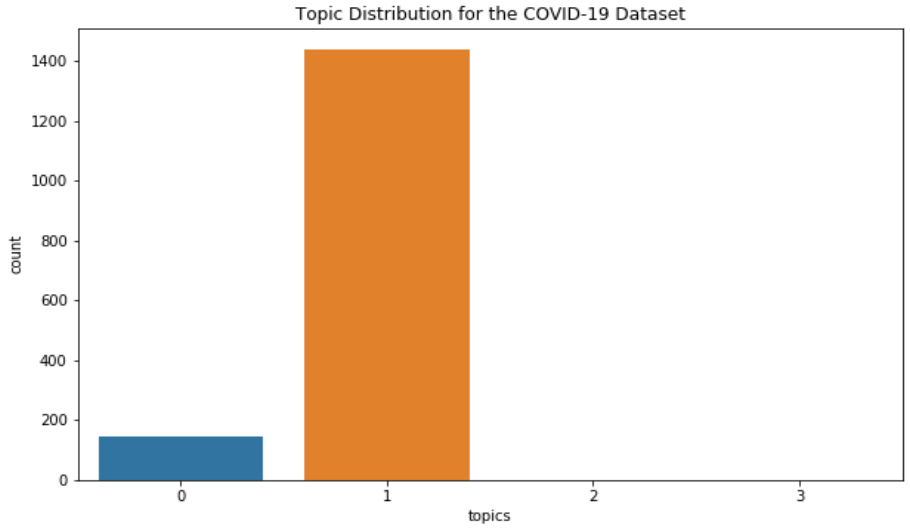
- **To verify whether the unsupervised learning is reasonable, I also used k-means to cluster the data.**





4. Deriving Insights for Current Situation





- **The comparison of the two figures illustrates that the literatures related to COVID-19 has almost the same trend as the whole dataset on topic 0 and topic 1. But there are not any papers related to topic 3 for COVID-19.**
- **This finding is reasonable because topic 3 related to the micro-level research. The COVID-19 appears only in a short time with outbreaks all over the world. Thus, the earliest research must focus on the macro characteristics, infection models and vaccines, which are the mostly needed so far. But cell or molecule-level research will cost longer time to gain useful insights, thus there are any papers collected by the dataset.**
- **Thus, the insight gain from the analysis is that although lots of papers about COVID-19 has been published, there aren't any knowledge focusing on micro-level has bee shared to research community.**