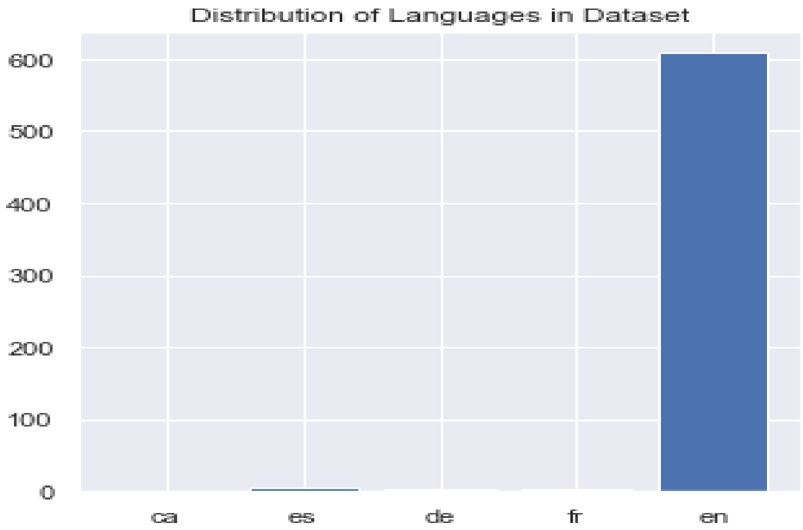
Covid-19 Literature Clustering

Description of project goals:

The rapid spread of COVID-19, given a large amount of literature and it is difficult for health professionals to keep up with new information on the virus. Clustering similar research articles together will simplify the search for related publications.

Data Used:

- The White House and a coalition of leading research groups have prepared the COVID-19 Open Research Dataset (CORD-19). CORD-19 is a resource of over 500,000 scholarly articles, including over 200,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses.
- The paper researches are written in 12 languages .
- We got a sample of 1000 paper research of all languages.
- We found 600 out of 1000 paper researches that were written in English language.



Next step, we start preprocessing on our data set:

start by removing stop words (common words that will act as noise in the clustering step). Second, we use the spacy library to convert the body text to lower cases, and remove punctuations. Third step, we handle Possible duplicates that appear because some articles were submitted a couple of times from different resources. Fourth, drop Null values to improve the clustering efforts. Fifth step, we handle multiple languages.

Sixth, we use (en core sci_lg) this is a model for processing biomedical, scientific or clinical text. Finally, applying the text-processing function on the **body_text**.

Features Engineering:
We add 3 features to our dataframe:

-word count in abstract

- word count in body

- number of unique words in body

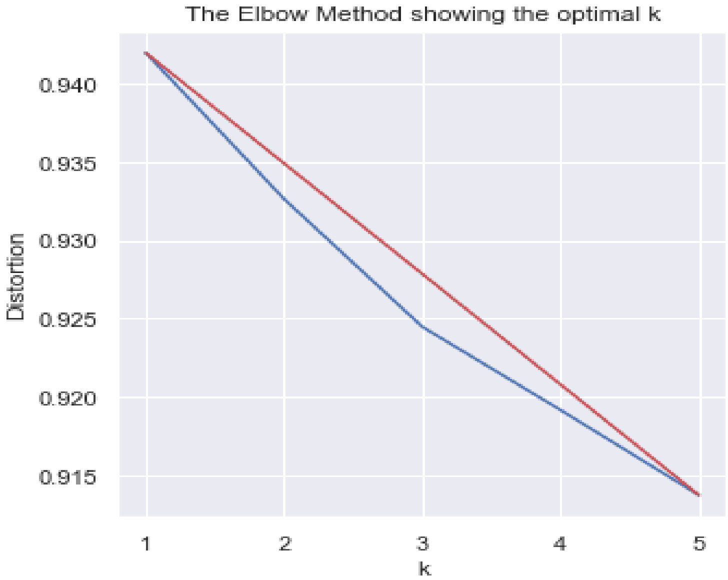
moving to Vectorization:

Use TF-IDF to convert our string data into a measure of how important each word is to the instance out of literature as a whole.

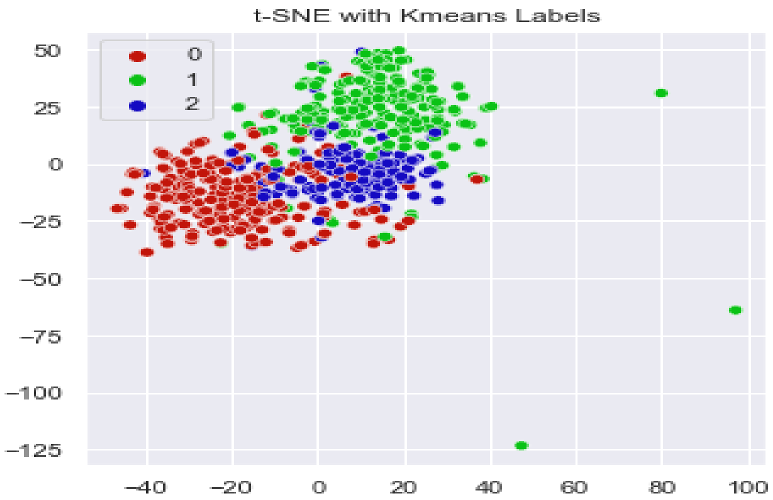
Tools Used:

- Numpy
- Pandas
- SKlearn
- Matplotlib

PCA Clustering:
Principal Component Analysis (PCA) to our vectorized data.
hopefully will remove some noise/outliers from the data
make the clustering problem easier for k-means
k=3



Work Flow:
by using clustering for labelling in combination with dimensionality reduction for visualization, the collection of literature can be represented by a scatter plot. On this plot, publications of highly similar topics will share a label and will be plotted near each other .



Classify:

Apply Stochastic Gradient Descent classifier and random forest.

Classifier	GDC	RF
F1	88.0	85.354

Conclusion :

In this project, we have attempted to cluster published literature on COVID-19 and reduce the dimensionality of the dataset for visualization purposes. However, we had some challenges during loading the data. First the size of the data was huge. We couldn't use a large amount of paper research as a sample, we tried a lot using colab ,however it didn’t work. We consumed our device's size, so we got a small sample and worked with it.