COVID-19 Literature Clustering

OUTLINE:

Goal.

Data Set.

Preprocessing text

Feature Engineering.

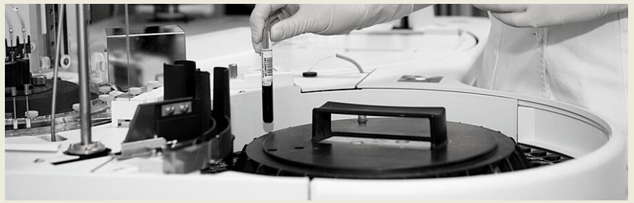
Vectorization.

PCA & Clustering.

Classify.

Challenges

GOAL:

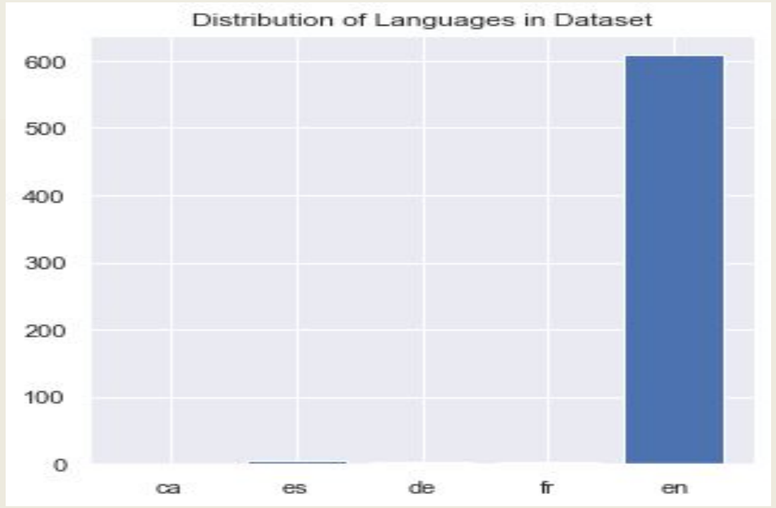


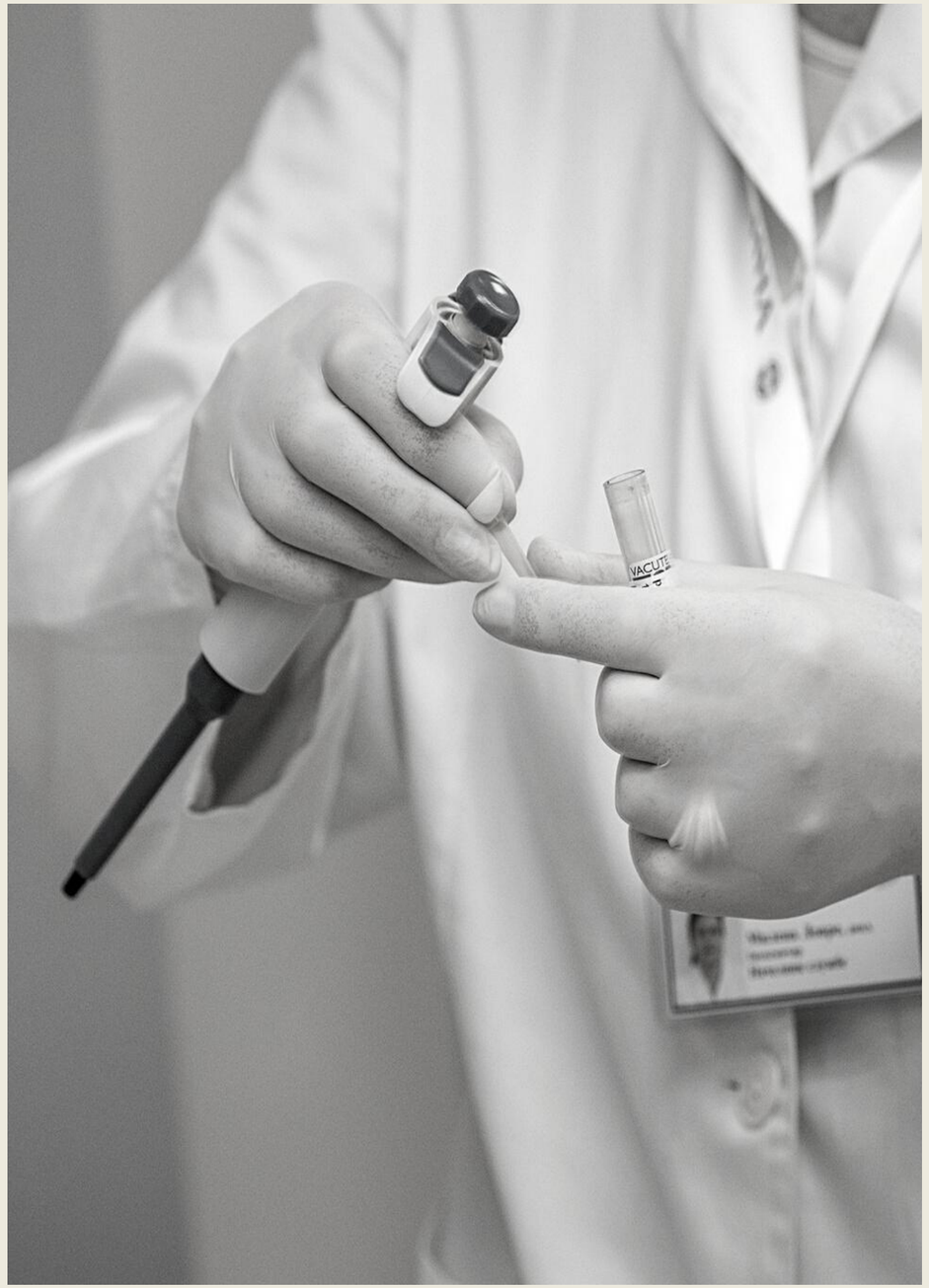
- Given the large number of literature and the rapid spread of COVID-19, it is difficult for health professionals to keep up with new information on the virus.

- Can clustering similar research articles together simplify the search for related publications?

Data Set:

- The White House and a coalition of leading research groups have prepared the COVID-19 Open Research Dataset (CORD-19).
- The data contains 256684
- paper search that written in 12 languages .
- sample 1000 paper research of all languages
- 600 English papers only





Pre-Processing Text.

- Remove stop words (common words that will act as noise in the clustering step).
- Using the spacy library to convert the body text to lower cases.
- Remove Punctuations.
- Handle Possible Duplicate.
- Drop Null values to improve the clustering efforts.
- Handling multiple languages
- For the parser, we will use (en core sci_lg) This is a model for processing biomedical, scientific or clinical text.
- Applying the text-processing function on the **body_text**.

Feature Engineering:

We add 3 features to our dataframe:

1-word count in abstract

2- word count in body

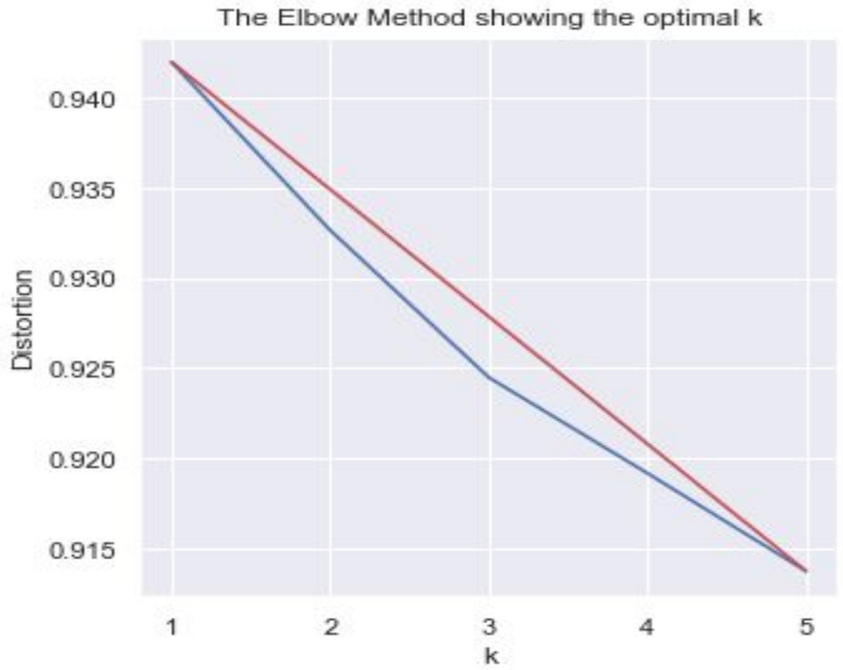
3- number of unique words in body

Vectorization

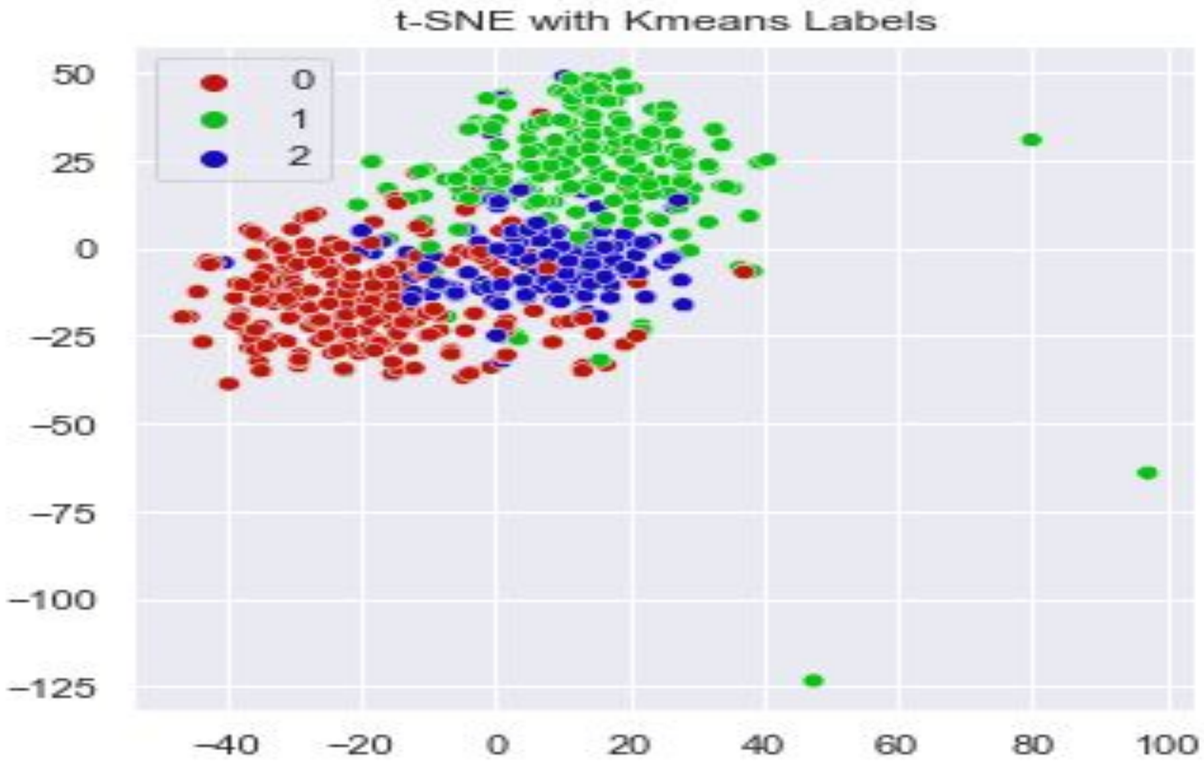
TF-IDF to convert our string data into a measure of how important each word is to the instance out of literature as a whole.

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



PCA Clustering:



Classify

apply Stochastic Gradient Descent classifier and random forest

Classifier	GDC	RF
F1	88.0	85.354

challenges

- size of data.
- consuming memory capacity and time

Presented by

NAME: Rana alqahtani

NAME: Ohoud Albabtain

NAME: Khulud Alshamrani