Project Report

(prepared by Jayesh Deep Dubey)

Title: Information extraction from Clustered documents using NLP

Description:
Text data contains large amount of information in it but not all of that information might be useful to the user for his task.Manually going through all the text and extracting useful information can be a very daunting and sometimes impossible task, especially in today's time when we have so much of data available to us to deal with. The objective of this project is to extract useful and relevent information from large amount of text articles using **NLP** techniques. This will allow the user to gain quick and useful insights from large text without having to go through the entire text which is time consuming and not feasible in many situations.

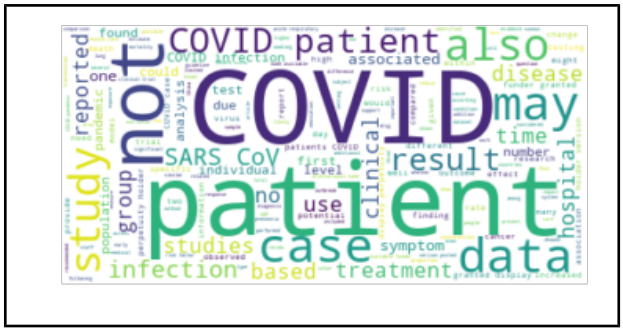
Dataset:
Covid-19 Open Research Dataset (**CORD-19**) has been used for this project. **CORD-19** is a resource of over 200,000 scholary articles, including over 90,000 with full text, about COVID-19, SARS-CoV-2 and related coronaviruses.(**https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge**)

Analysis Method:
Only 10,000 articles were considered for the study because of memory limitations of the system. After extracting the data, some basic preprocessing (decontraction, removing Nan values, stopwords removal etc.) was performed on it to clean the noise from the data. The text for some research articles was in languages other than english and therefore those articles were not considered in the analysis.The format of dataframe after performing above mentioned steps is as follows:

	paper_id	title	abstract	text	language
0	0001418189999fea7f7cbe3e82703d71c85a6fe5	Absence of surface expression of feline infect...	Feline infectious peritonitis virus (FIPV) pos...	Feline infectious peritonitis (FIP) is a fatal...	en
2	000379d7a7f37a2ccb978862b9f2016bd03259ea	ScienceDirect ScienceDirect Effect of Nanomate...	approach. The NM shape in the conformal circui...	Integration of functional electronic devices o...	en
3	00039b94e6cb7609ecbddee1755314bcfeb77faa	Plasma inflammatory cytokines and chemokines i...	Severe acute respiratory syndrome (SARS) is a ...	Severe acute respiratory syndrome (SARS) is a ...	en
8	00073cb65dd2596249230fab8b15a71c4a135895	Risk Parameters of Fulminant Acute Respiratory...	A clinical picture of patients with acute resp...	Since then, many clinical case reports have be...	en
10	0008c57de475138d903f2cca7003cf1e1ad93cf4	The effect of gramicidin inclusions on the loc...	We study the local effect of the antimicrobial...	The effect on the cell membrane of inclusions ...	en

The 'text' column (which consists of the entire text of research article) is considered for analysis. Since the amount of text is large so it is better to perform the analysis on the data in groups. However, instead of randomly dividing the data into groups a better approach is to use clustering algorithms for grouping as it will cluster similar text into one group and it will be easier for the user to analyse the data. Therefore the objective is to first cluster the research articles into groups using **K Means** clustering and then perform information extraction on the entire text in each cluster.

TF-IDF vectorization technique was used for converting text into vectors and then **PCA** was applied on it for reducing the dimensions while preserving 99% variance.After that K means clustering was applied. **Silhouette score** was used as a metric for determining the value of K to be used in K means clustering. The documents were clustered into 93 groups.**Wordcloud** was generated for each cluster so that user can get a brief idea about the text/content of the clusters and then decide the clusters on which he wants to perform information extraction. WordCloud for one of the clusters is shown below:



After documents are clustered, user can select the clusters by looking at wordcloud (i.e. selecting the clusters that are appropriate for his task) and provide these clusters as a python list to the program. Information extraction is then implemented on selected cluster. For information extraction task, **spaCy Matcher class** has been used. It matches a sequence of words based on the patterns given to it. **Spark** was used to speed up the computation. After the program is implemented it will extract the useful sentences from the text based on given pattern. These results are saved to a csv file which consists of following columns:

T**itle:** It is the title of the research article from which the **sentence** is extracted.
Sentence: It is the extracted sentence

Wordclouds for each cluster are saved to disk. Please look at the **results** folder which contains the extracted sentences csv file and wordclouds for some of the clusters.