Project Report

(prepared by Jayesh Deep Dubey)

Title: Using TextRank algorithm for summarizing clustered text

Description:
The goal of this project is to summarize large amount of text in order to gain fast and useful insights from the text without reading the entire text which is time consuming and not feasible in many situations.TextRank algorithm has been used for summarizing the text. TextRank is an extractive and unsupervised text summariation technique that is derived from PageRank algorithm and is used for generating a concise and meaningful summary of the text from multiple text resources.It assigns a pagerank score to each sentence in the text and top ranked sentences become the final summary.

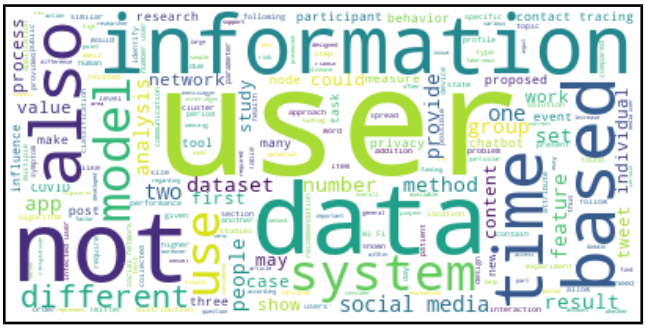
Dataset:
Covid-19 Open Research Dataset (**CORD-19**) has been used for this project. **CORD-19** is a resource of over 200,000 scholary articles, including over 90,000 with full text, about COVID-19, SARS-CoV-2 and related coronaviruses.(**https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge**)

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. The objective is to first cluster the research articles into groups using K Means clustering and then perform text summarization using TextRank algorithm on each cluster. The final results of TextRank algorithm also shows the ‘title’ of research paper for each ranked sentence so that user can identify the research paper from the sentence. This is useful in cases where user finds the sentence useful for his task and wants to read the research article that the particular sentence belongs to.

TF-IDF vectorization technique was used for converting text into vectors and then PCA was applied on it for reducing the dimensions while preserving 95% 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 98 groups.**Wordcloud** was generated for each cluster so that user can get a brief idea about the text of the cluster.



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. **TextRank** algorithm is then implemented on selected cluster. For text summarization part, **GloVe** word embeddings was used instead of **TF-IDF** to convert clustered text into vectors because TF-IDF results in very large number of features. Since text summarization is time consuming so Spark was used to speed up the process through distributed computing. A snapshot of the final results of TextRank algorithm is as follows:

```
+-----+--------------------+--------------------+--------------------+
|   id|               title|            Sentence|            pagerank|
+-----+--------------------+--------------------+--------------------+
| 4904|The effect of gra...|3(d) ) we observe...|  1.875997247779063|
| 9200|Development of po...|Empty vector-tran...|0.21133057003666011|
|10664|Quantitative Prot...|Scale bars, 20 m....| 1.1237006576966584|
| 3456|Identification of...|Hemocytin Basic t...|  2.496453085945066|
| 7608|La, PTB, and PAB ...|This poly(A) sequ...| 2.3745473664508516|
| 6400|Regulated Export ...|Abbreviations use...| 0.1869074312870813|
| 9208|Development of po...|The OD 490 value ...| 0.5533710572692503|
|10984|Nucleolar-cytopla...|EGFP by itself, b...| 0.9569977054934737|
| 3272|GLYCOMICS AND GLY...|The importance of...|0.15765250145255702|
|  752|Characterization ...|Possibly, hydroph...| 0.5721868399591703|
| 7184|Background K 2P C...|An aliquot of HBS...|0.15874565659306106|
| 4352|Nucleotide Sequen...|In three places t...| 0.8071918587968431|
| 1328|Protein Microarra...|The results showe...| 0.7532727763952681|
|  464|Review Viral and ...|Because enveloped...|  3.127072977034685|
|10264|Isolation of the ...|Tris, NaCl, HCl, ...|0.21717583099554164|
| 1040|Expression and me...|Fixed cells were ...| 1.0004523770234823|
| 1336|Protein Microarra...|             2006 ).| 0.8478794516867104|
|  912|TGEV Corona Virus...|Immunoprecipitati...| 2.5694667447760624|
| 5384|Shaping the flavi...|Due to the centra...| 0.1585854663052905|
|10920|Comparison of enz...|The newly establi...| 2.4442138597807626|
+-----+--------------------+--------------------+--------------------+
only showing top 20 rows
```

Here '**title**' is the title of the research article that contains the ranked sentence, '**sentence**' is the ranked sentence and '**pagerank**' is the score assigned to sentence based on its importance. Please look at the **results** folder which contains the text summary and wordclouds for some of the clusters.