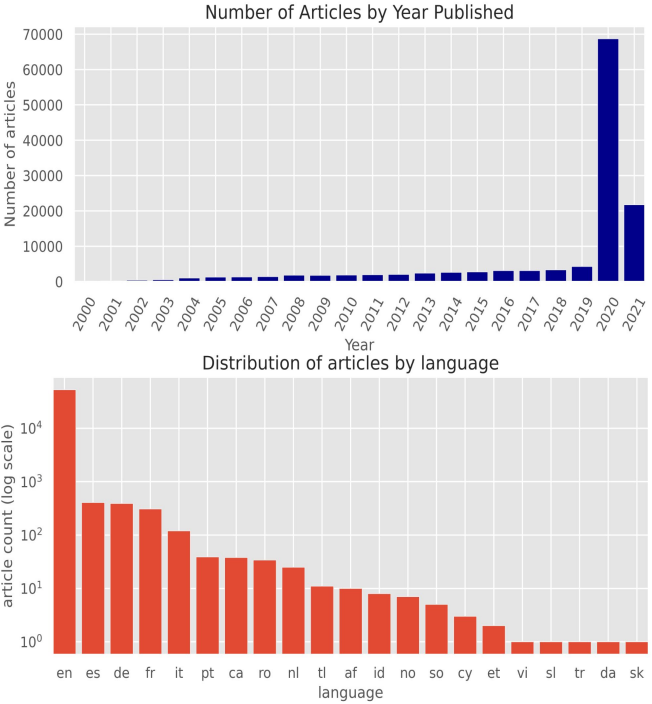
Final Project - Text-Mining COVID-19 Research Dataset

MIE1624 - Introduction to Data Science & Analytics

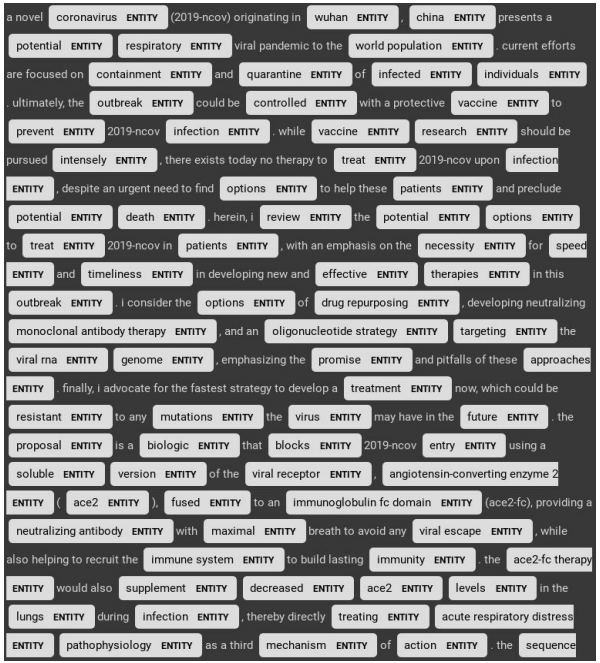
Sugumar Prabhakaran (id: 994126815)

15 Apr 2021

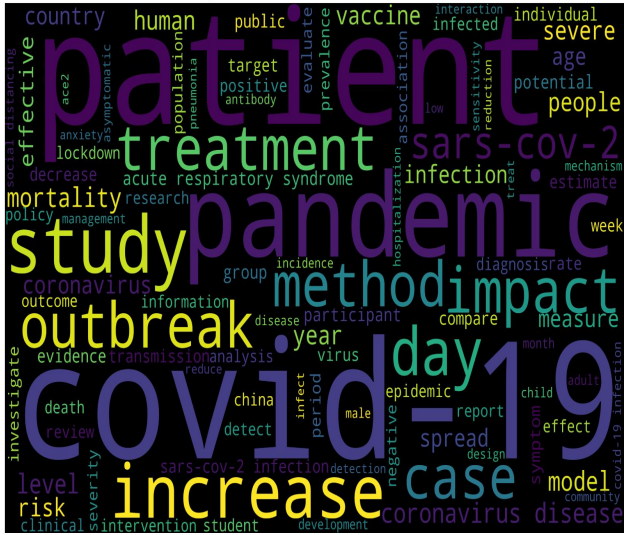
Data Cleaning and EDA



scispaCy Identification of Biomedical Terms



WordCloud of Top 100 words from Abstracts



Data Cleaning steps:

1. Drop duplicates, null rows
2. Drop articles before Jan 2020
3. Drop non-english articles

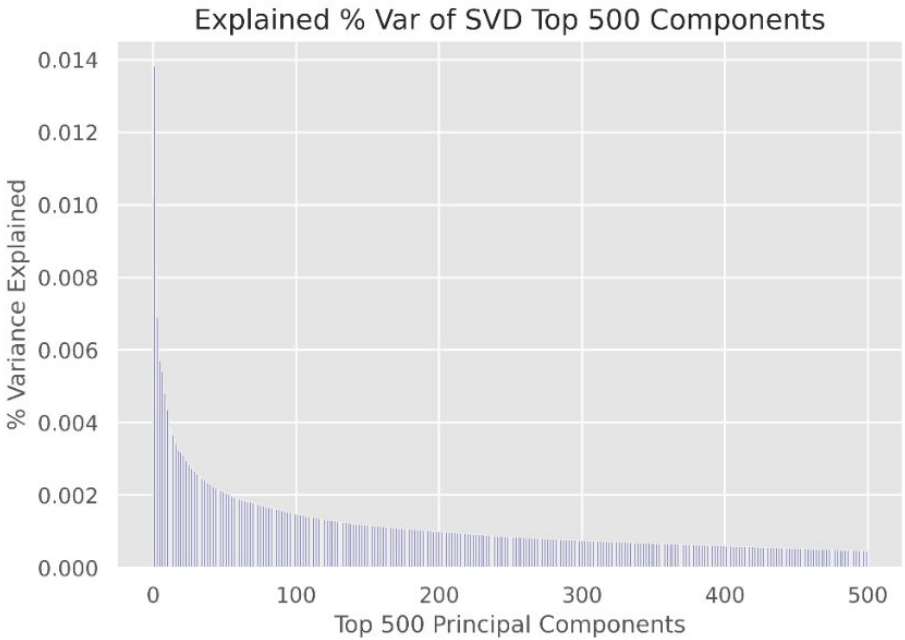
Natural Language Processing (NLP) steps:

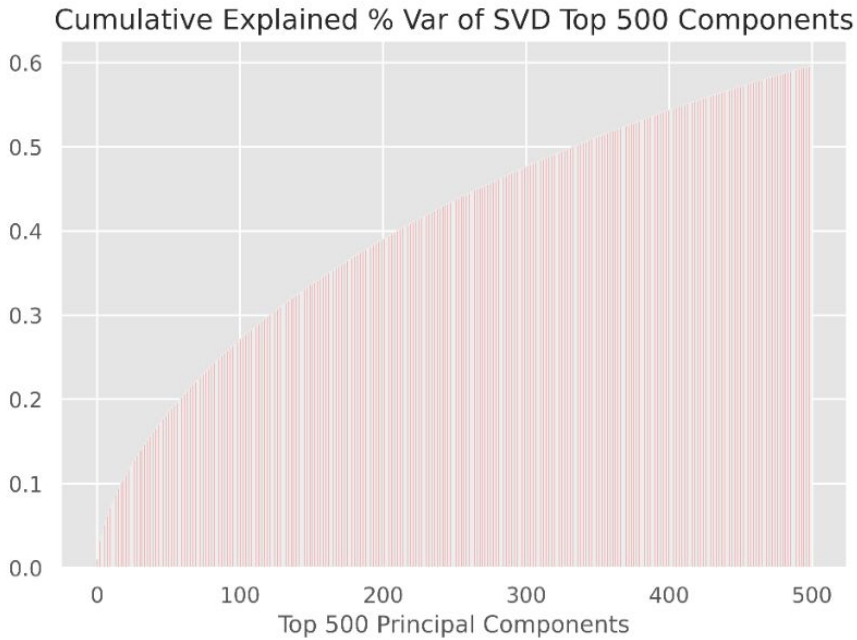
1. Used scispaCy library biomedical term model to identify bio terms
2. Removed stop-words and lemmatized

Pre-processing steps:

1. Vectorized using TF-IDF Vectorizer
2. Generated word cloud to see most common words after NLP steps

Feature Selection and Importance





Dimensionality Reduction: Truncated Singular Value Decomposition (SVD) was used instead of PCA due to memory (RAM) issues, in order to **reduce the number of features from 2185 to 500**. This speeds up calculations but still **retains approximately 60% of the overall variance** of the vectorized form of the data.

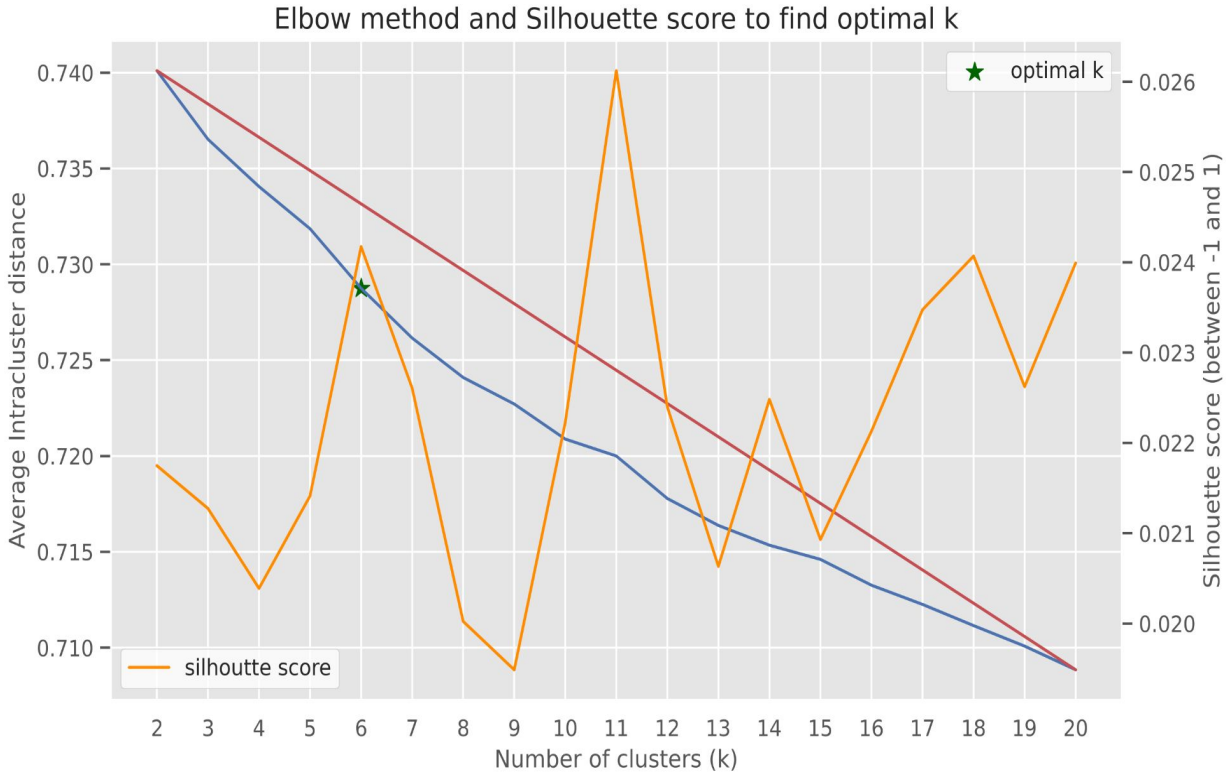
KMean Unsupervised Clustering Model Tuning

For the KMean clustering model, the hyperparameter to tune is the **number of clusters (k)**.

This was tuned using two methods (both represented on graph on the right):

1. **Elbow Method (blue line).** Calculate the average (intra-cluster) distance of all data points to their nearest centroid for each k-value. **Optimal k-value occurs when elbow starts flattening (ie. negative slope decreases).**

2. **Silhouette Score**. Calculates the ratio of intra-cluster distance (as above) and inter-cluster distance (between clusters) and gives a **value between -1 and 1**. A higher value means a small intra-cluster distance, meaning tighter clustering.



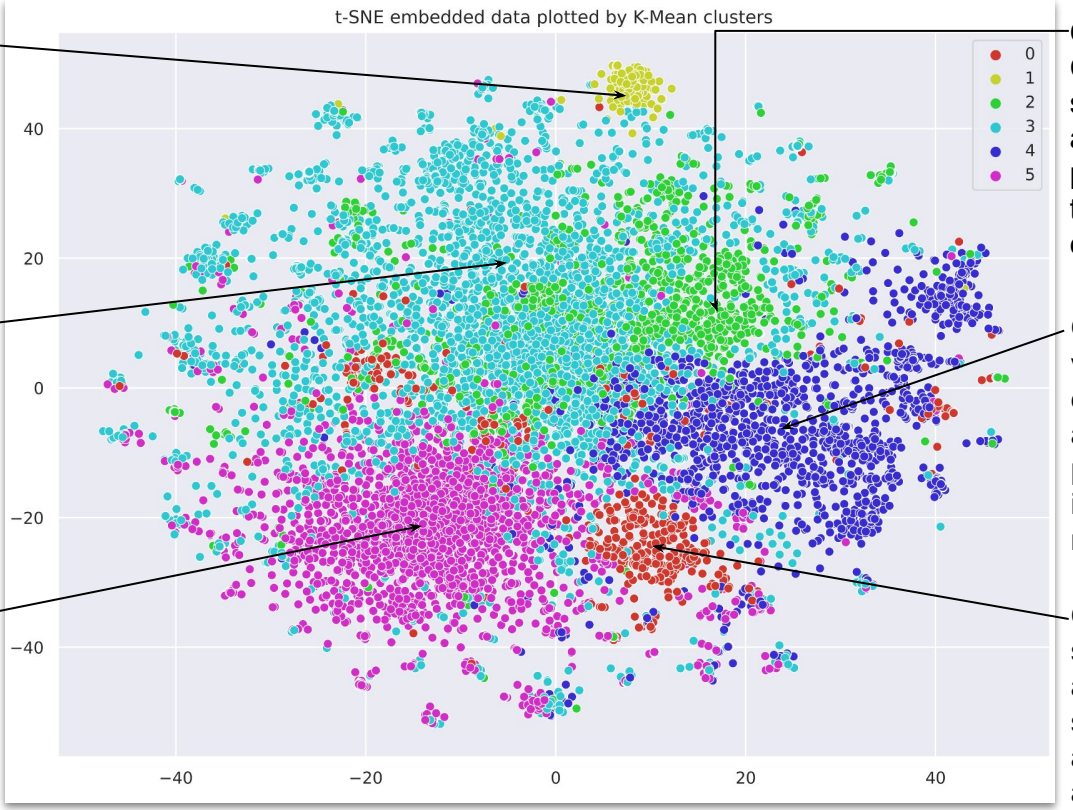
The optimal k-value = 6 was selected based on a balance of silhouette score and elbow method. Higher k-values lead to lower intracluster distances but also risk overfitting the data.

KMean Clustering and LDA Topic Modelling Results

Cluster 1 (Education):
Student, pandemic, anxiety, college, medical student, age, design, staff, patient, study, young, vulnerability, learning activity

Cluster 3 (General Impact):
pandemic, patient, impact, increase, level, anxiety, child, women, day care, public, information, transmission, risk, ppe concentration

Cluster 5 (Mortality Rate):
mortality, age group, patient, treatment, severe, symptom, mechanical ventilation, fever, tocilizumab, il6, crp, ards



Cluster 2 (Transmission):
Country, outbreak, policy, social distancing, epidemic, asymptomatic, lockdown, population, temperature, transmission, travel, city, contact, measure

Cluster 4 (Vaccines):
vaccine, treatment, expression, drug, interaction, antibody, efficacy, ace2, protease, concentration, inhibit, gene, target, mutation, replication, chloroquine

Cluster 0 (Detection):
sensitivity, specificity, sample, assay, saliva, asymptomatic, seroprevalence, detection, igg, accuracy, rate, diagnosis, antibody, viral load

From these topic keywords it seems that each cluster covers distinct topics and could inform policy.

Insights and Potential Policy Guidance

Insights that can be drawn from the topic modelling keyword results (previous slide):

- **Cluster 0 - COVID-19 Detection & Screening:** testing saliva using assay and igg antibody tests are related to improving detection and accuracy. Further, seroprevalence (viral load), has a relation to detection and asymptomatic patients. This information could be used by hospitals and governments to improve testing.
- **Cluster 1 - COVID-19 Impact on Education:** there is a concentration of studies that look at impact on elementary aged students, adolescents and college students as well as education methods such as in-person learning, interactive learning and learning activities. These could be used by governments to assess impact and risk of schools remaining open.
- **Cluster 2 - Factors Affecting Transmission:** policy measures such as social distancing, lockdown, contact tracing, travel restrictions and testing may help reduce transmission.
- **Cluster 3 - General Impact:** significant collection of articles discuss increase in anxiety in the population, and need to train and staff health care workers, medical residents and researchers.
- **Cluster 4 - Vaccine-Virus Interaction:** contains research on how the virus targets and binds to the ace2 receptor, how drugs could potentially target virus protesases, and whether chloroquine is an effective treatment.
- **Cluster 5 - Factors affecting Mortality:** age group, treatment type, ventilator access, c-reactive protein levels (CRP), availability of tocilizumab receptor antibodies and IL-6 inhibitor levels could affect mortality and respiratory illnesses (ARDS).