COVID Literature Clustering

Abhilash Hemaraj (001305283)
Niyati Chopra (001443559)

Introduction

Allen Institute for AI in collaboration with 6 other organizations, released a dataset containing over 125 thousand literature articles about information on the COVID-19 pandemic

The main aim of our project is to perform clustering of our articles to better classify them based on the topics they cover. The literatures could contain research papers, blog articles, studies, etc based on a variety of topics like virus behaviour, symptoms, prevention, vaccination, etc.

Dataset: https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

Dataset

-The dataset was originally in the form of .json files

-We extracted the text corpus to get 128915 rows of data

-Out of over 30 possible column values, we chose this as our final dataset:

	paper_id	title	authors	abstract	body_text
0	0000028b5cc154f68b8a269f6578f21e31f62977	"Multi-faceted" COVID-19: Russian experience	NaN	NaN	['According to current live statistics at the ...
1	0001418189999fea7f7cbe3e82703d71c85a6fe5	Absence of surface expression of feline infect...	['E Cornelissen', 'H Dewerchin', 'E Hamme', 'H...	['Feline infectious peritonitis virus (FIPV) p...	['Feline infectious peritonitis (FIP) is a fat...
2	00033d5a12240a8684cfe943954132b43434cf48	Detection of Severe Acute Respiratory Syndrome...	['Petra Wandernoth', 'Katharina Kriegsmann', '...	['Background: Amplification of viral ribonucle...	['Severe acute respiratory syndrome coronaviru...
3	0003793cf9e709bc2b9d0c8111186f78fb73fc04	Title: Rethinking high-risk groups in COVID-19	['Anastasia Vishnevetsky', 'Michael Levy']	NaN	["How do we protect our 'high-risk' patient po...
4	000379d7a7f37a2ccb978862b9f2016bd03259ea	ScienceDirect ScienceDirect Effect of Nanomate...	['Harish Devaraj', 'Rajiv Malhotra']	['approach. The NM shape in the conformal circ...	['Integration of functional electronic devices...

Data cleaning

There were quite a few NaN values in our dataset. Since all the values are string, we converted the NaN values to None so that we can use it for EDA and model fitting

Another problem was that our body had a lot of null strings. We also removed those rows too.

We also removed any papers where the number of authors > 30, since it is unlikely that a single paper have too many authors We performed our analysis on 20% of our data chosen randomly

```
paper_id        0
title       14023
authors     12465
abstract    41063
body_text       0
```

Natural Language Processing

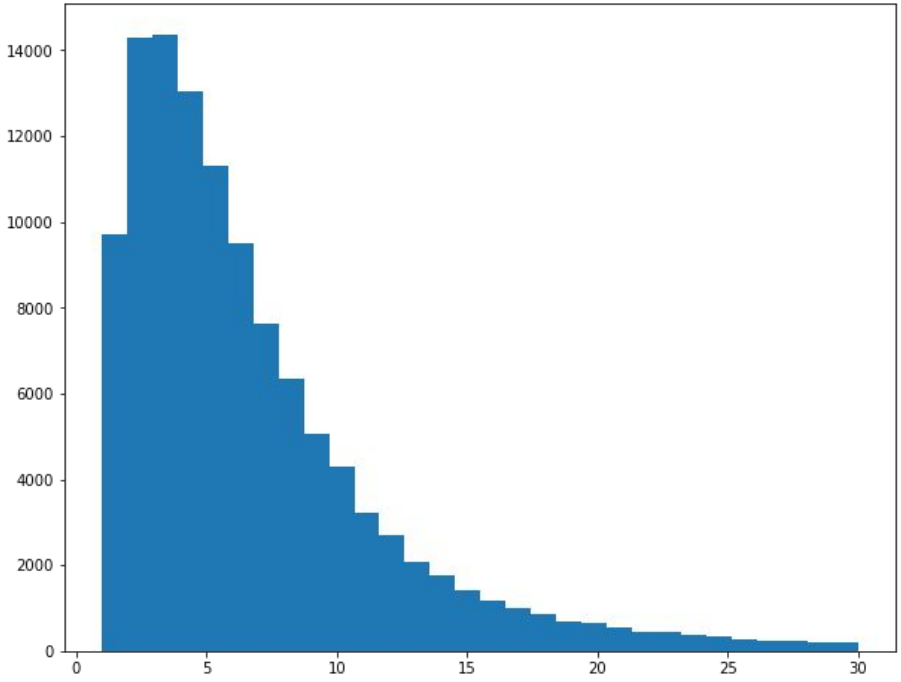
```
TfidfVectorizer(max_df=0.8, max_features=None,
                min_df=0.2, stop_words=stop_words_nltk,
                use_idf=True,smooth_idf=1,sublinear_tf=1, tokenizer= tokenize_and_stem,
                ngram_range=(1,3),lowercase=True)
```

Tokenization and stemming converts all words into their root word and removing word extensions (like -ing and -ers) in order to standardize the text corpus

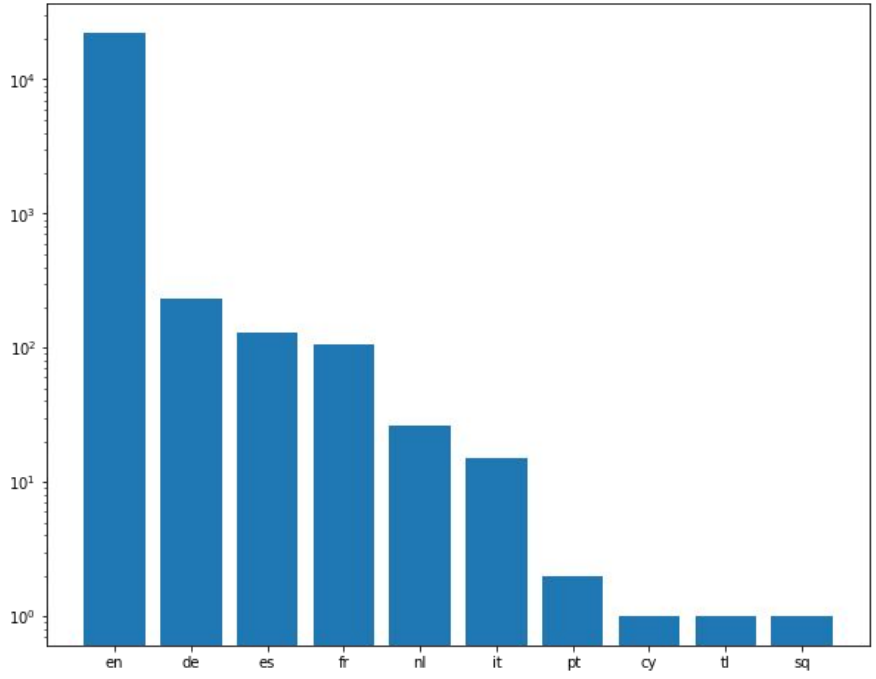
The TF-IDF Vectorizer is used for converting our text corpus to a vector representation of the words. It also removes all unnecessary stop words

Exploratory Data Analysis

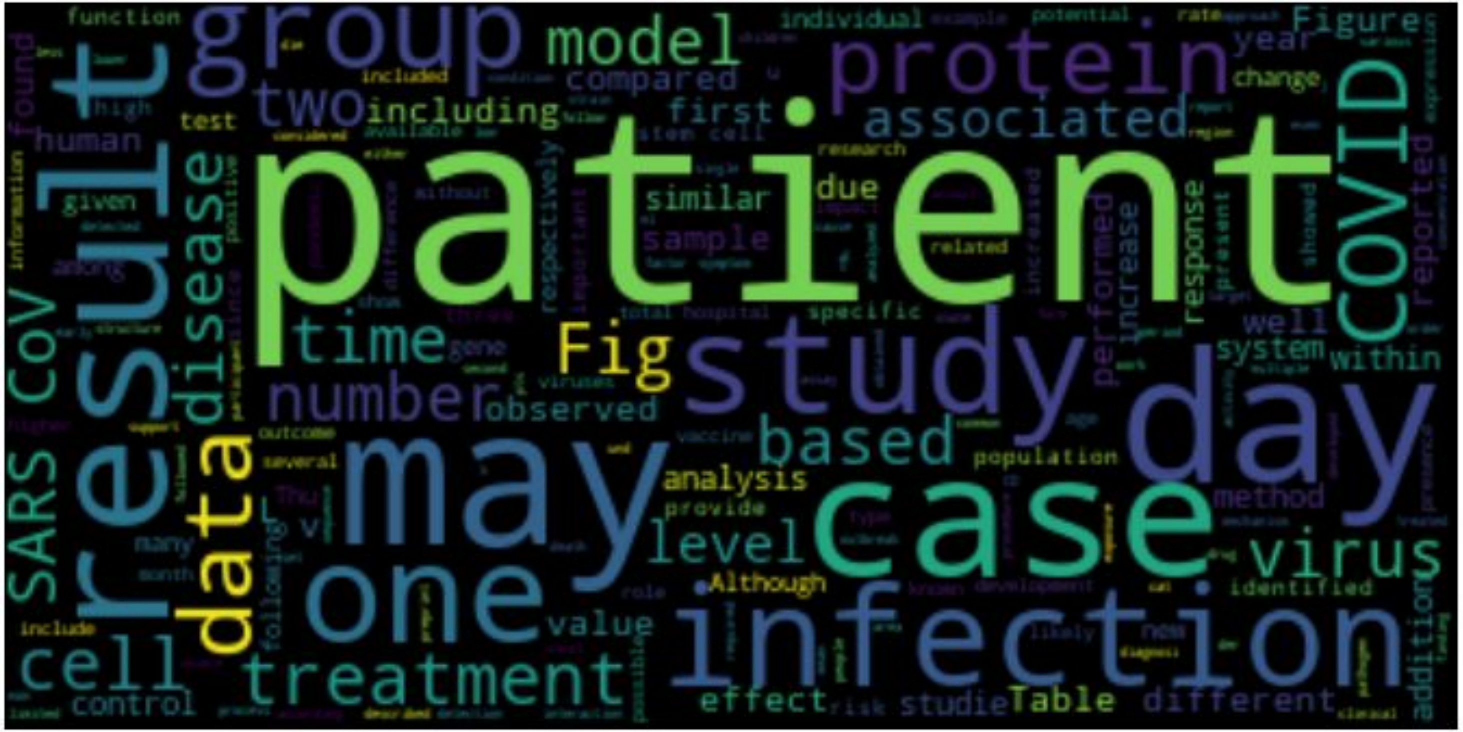
Distribution of number of authors per paper



Distribution of languages



Word Cloud



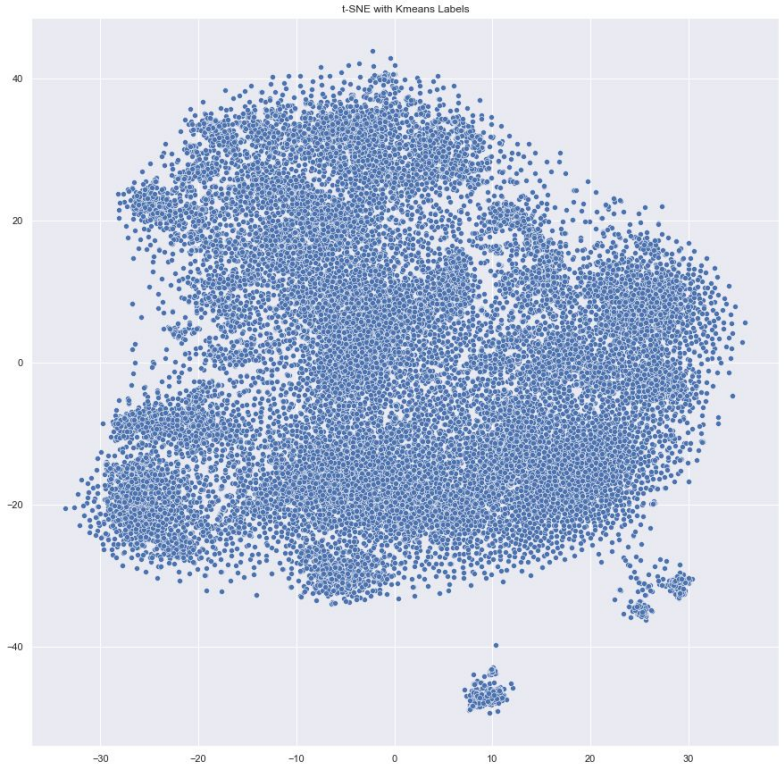
patient
result
may
case
infection
day
study
protein
group
one
data
disease
COVID
SARS
CoV
treatment
cell
model
two
time
Fig
number
associated
including
compared
first
individual
potential
rate
Figure
year
change
reported
function
human
found
test
high
similar
due
sample
based
level
virus
value
analysis
provide
Although
effect
Table
different
control
include
identified
well
system
within
method
population
specific
observed
response
increase
performed
increased
research
related
information
given
among
following
several
outcome
many
month
role
respectively
important
stem cell
vaccine
likely
addition
present
showed
gene
viruses
total hospital
risk
studie
known
development
included
available

Dimensionality Reduction

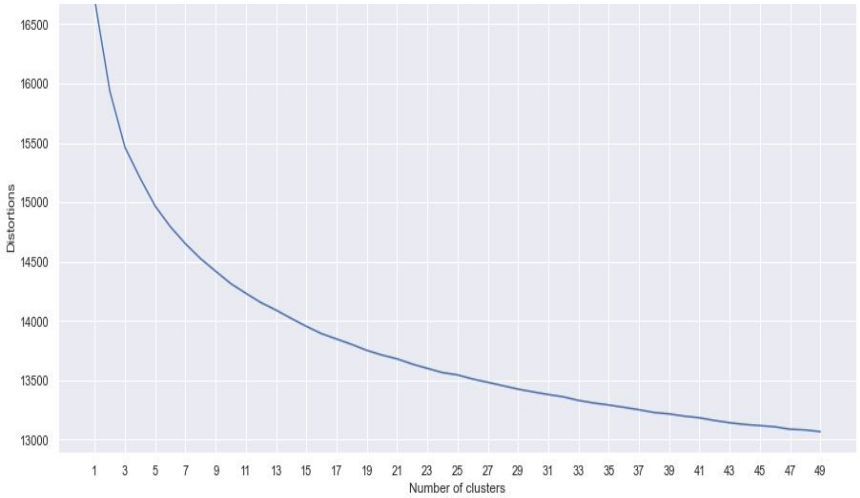
- PCA : Fit a PCA model to reduce dimensions while retaining 95% of the variance in our data. We got the number of principal components to be 644

```
shape after PCA: (25783, 644)
```

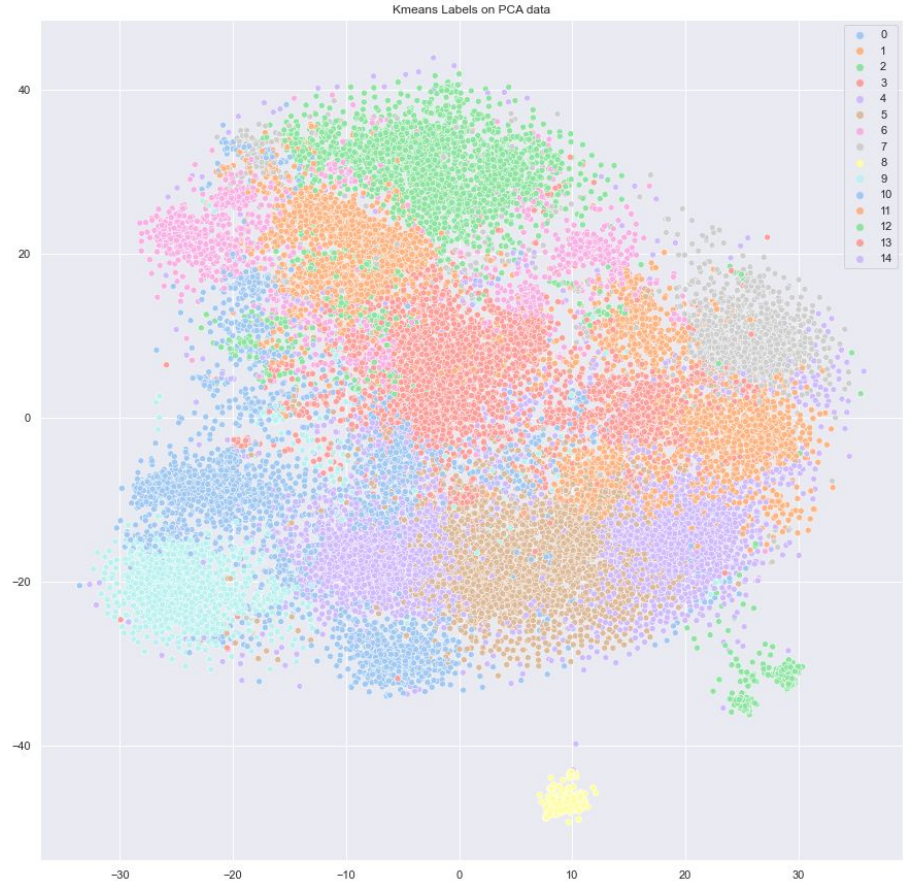
- TSNE : Fit a TSNE model to our data to reduce it to two dimensions with the resulting graph looking like:



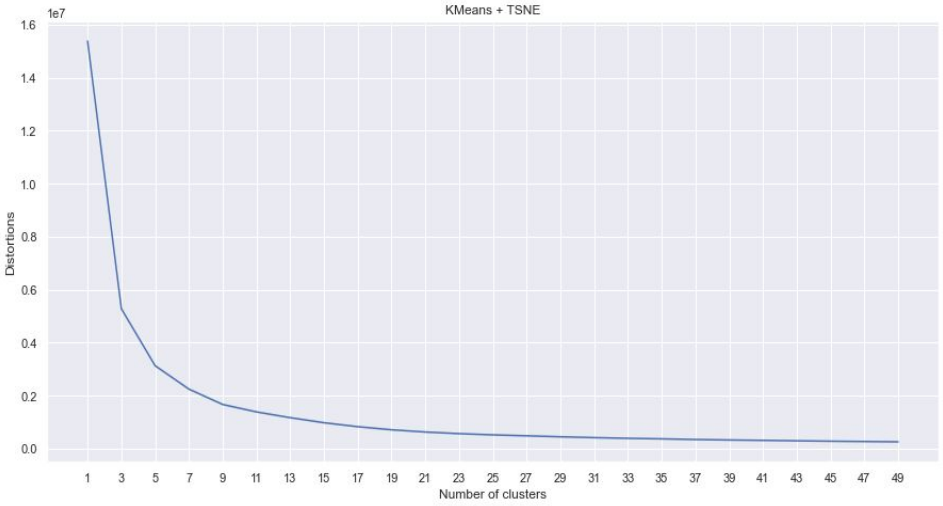
Clustering: KMeans + PCA



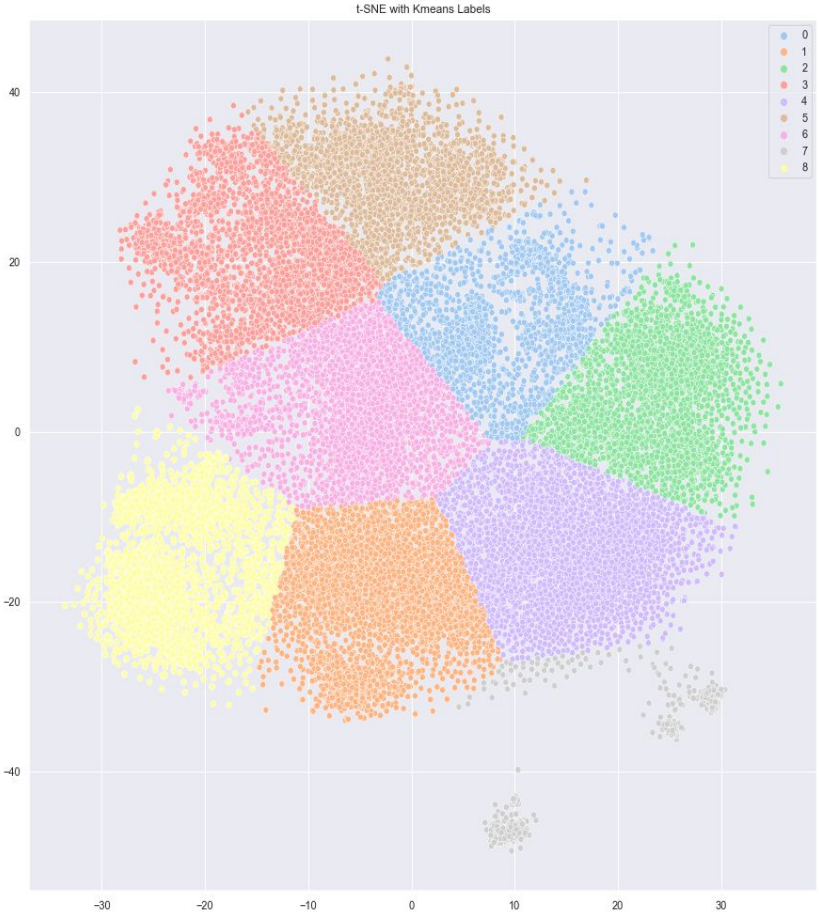
Optimal number of clusters = 15



Clustering: KMeans + TSNE



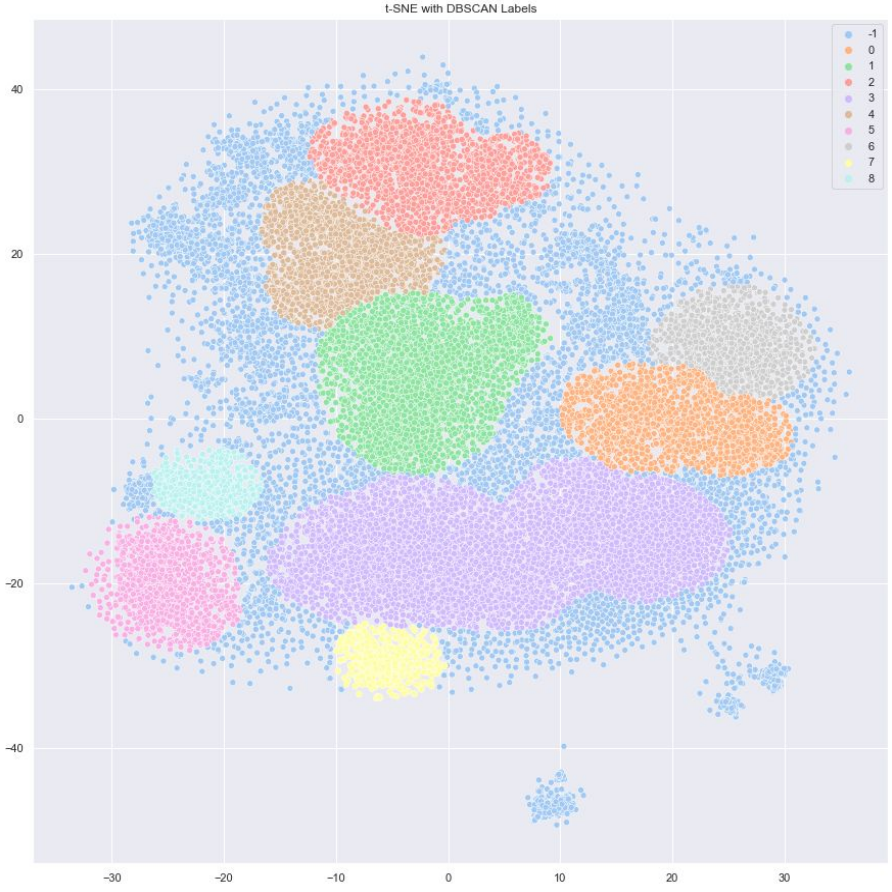
Optimal number of clusters = 9



Clustering: DBSCAN + TSNE

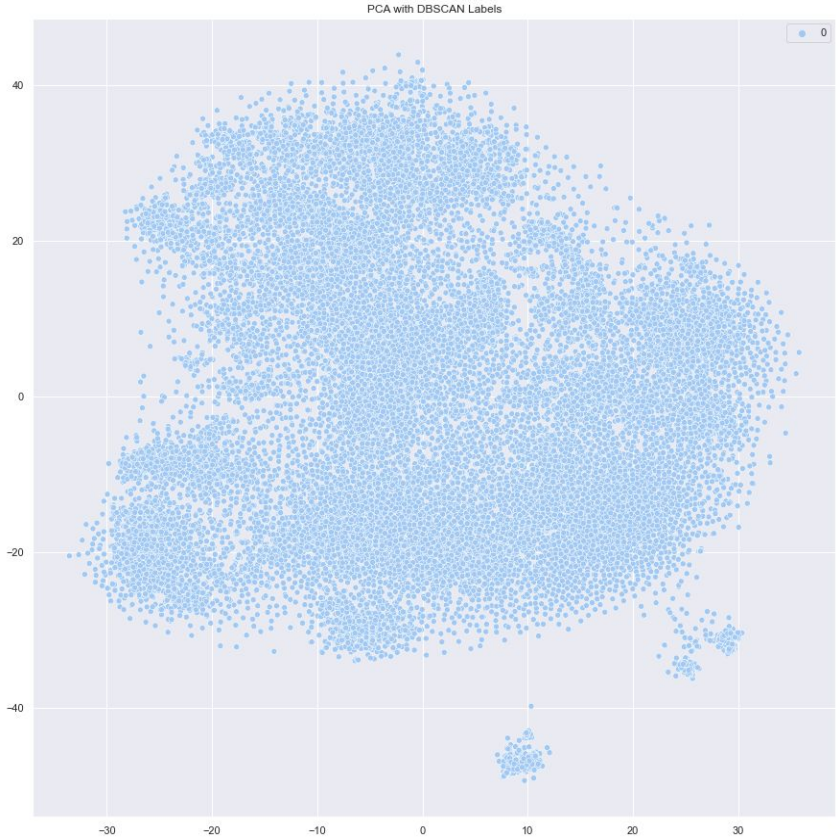
Chose value of eps and min_samples by trial and error

eps = 4 ; min_samples = 500



Clustering: DBSCAN + PCA

Due to computational limitations, we were unable to fit an appropriate DBSCAN on our PCA reduced data. This is the we got after using similar parameters as KMeans



Conclusions and future scope

For the text corpus that we chose,

- TSNE performed much better than PCA, however TSNE ends up losing a lot of the variance in the data. Therefore, PCA is a better dimensionality reduction technique to capture variance.

- KMeans fit better as we got a defined number of clusters, whereas there was a lot of trial and error involved in DBSCAN

- There are ways to tune the hyperparameters (eps, min_samples) of DBSCAN or use other options, like the OPTICS algorithm, to get the best possible values of DBSCAN, which are a part of the future scope of this project.

Thank you

Questions?