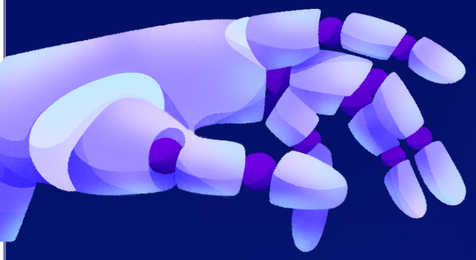


DATA ANALYSIS AND VISUALIZATION PROJECT

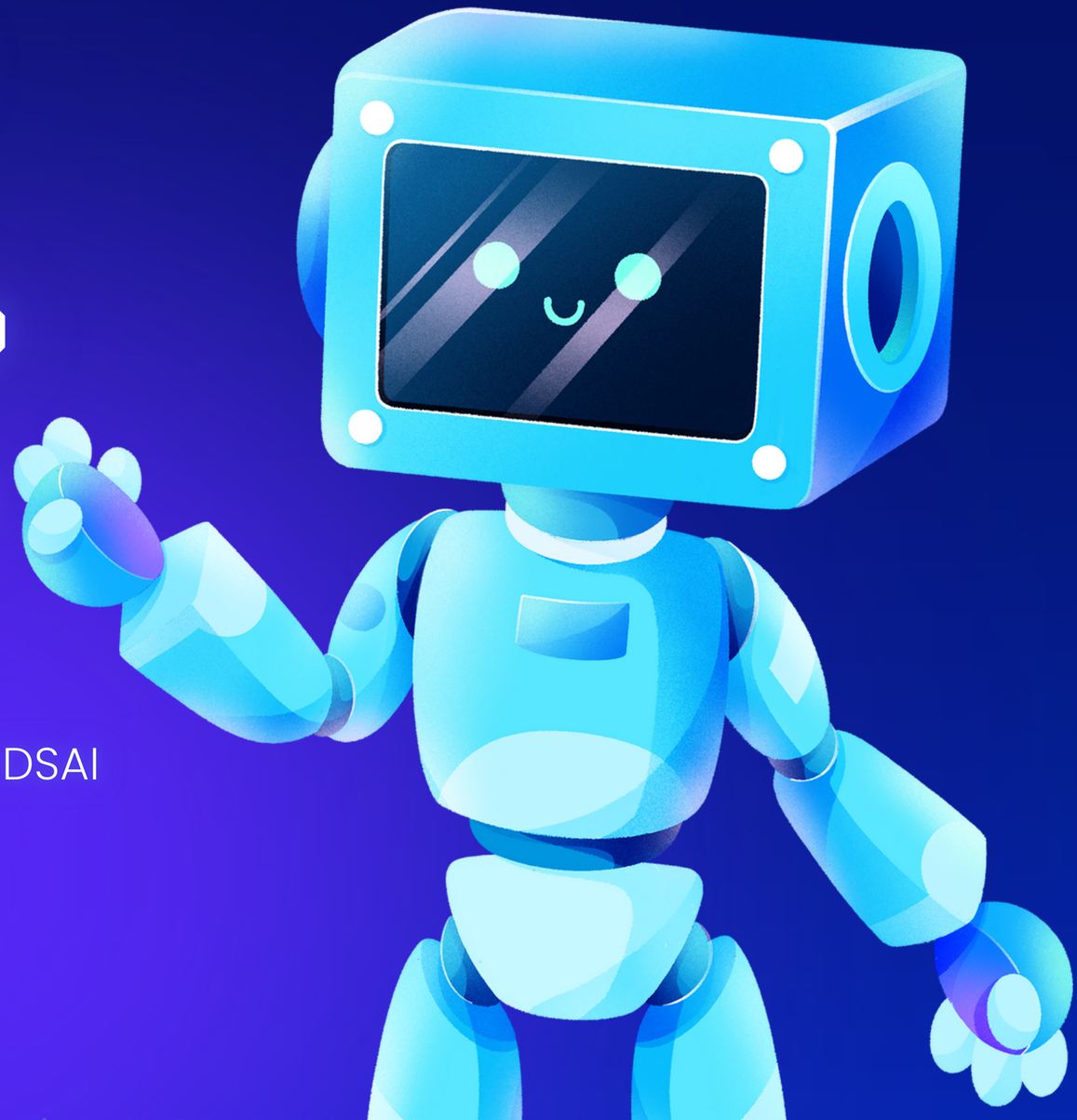
AMERICAN PRESIDENTS' PERSONALITY ANALYSIS AND RANKING PREDICTION THROUGH SEMANTIC SPEECH ANALYSIS

ROHIT ASHWANI

BTech(Honors) CSE with specialization in DSAI

III Year Undergraduate

Indian Institute of Technology Bhilai



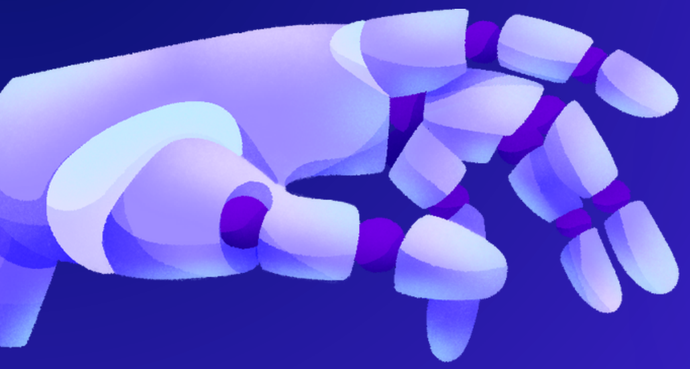
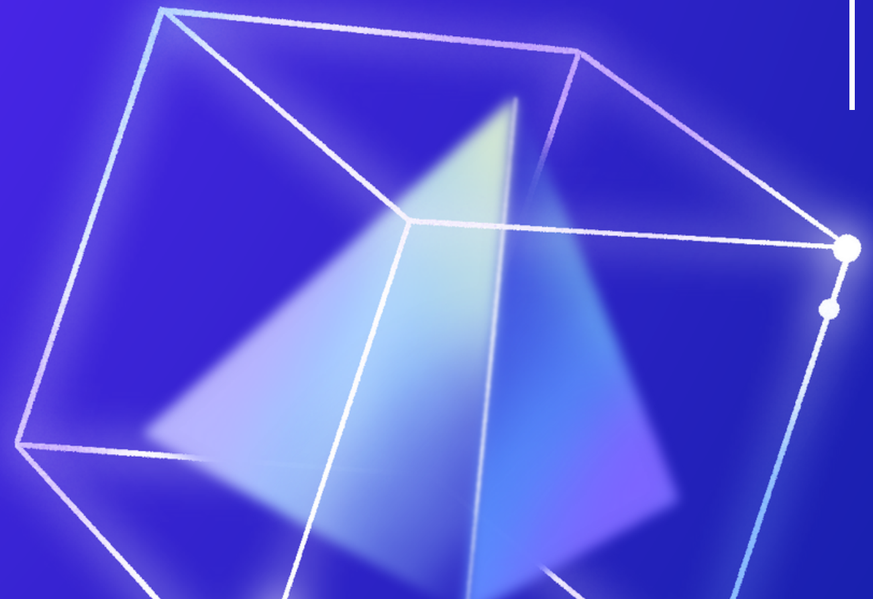


TABLE OF CONTENTS

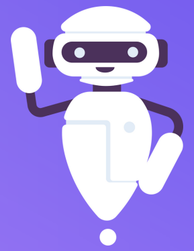


PROBLEM STATEMENT

THE OBJECTIVE IS TO EMPLOY SEMANTIC ANALYSIS TECHNIQUES ON HISTORICAL SPEECHES DELIVERED BY AMERICAN PRESIDENTS TO GROUP THEM INTO CLUSTERS BASED ON THE CONTENT AND UNDERLYING THEMES. ADDITIONALLY, THE GOAL IS TO CREATE A PREDICTIVE MODEL THAT CAN DETERMINE THE CLUSTER AFFILIATION OF A NEW PRESIDENTIAL CANDIDATE BASED ON THEIR SPEECH CONTENT, LEVERAGING PATTERNS IDENTIFIED FROM PREVIOUS PRESIDENTIAL SPEECHES.

FURTHERMORE, THE PROJECT AIMS TO DEVELOP A RANKING PREDICTION MODEL THAT ESTIMATES THE POTENTIAL RANK OR POSITION OF A NEW PRESIDENTIAL CANDIDATE WITHIN THE ESTABLISHED CLUSTERS. THIS PREDICTION WILL BE BASED ON THE RANKINGS OR CLASSIFICATIONS DERIVED FROM THE SPEECHES OF THE EXISTING AMERICAN PRESIDENTS.

PROJECT OBJECTIVES



OBJECTIVE 01

To group the American Presidents into clusters based on the semantic analysis of the speeches given by them

OBJECTIVE 02

If a new presidential candidate gives some speech we should be able to predict the cluster to which he belongs to based on the previous provided speeches

OBJECTIVE 03

To predict the ranking of the new presidential candidate based on the rankings of the provided presidents.

METHODOLOGY

Data Collection

- Data collected by scraping the official website of University of Virginia (Miller Center) which contains the records of speeches given by American Presidents
- Data was also collected from the datasets avaliable at Kaggle.

Merging of Data and Data Exploration

- Data from these different sources were not compatible to merge with each other.
- Merged them by making appropriate conversions.
- The final dataset was then explored to see the amount of data collected and checking whether the data is representative.

METHODOLOGY (CNTD.)

Data Preprocessing

- **Data Cleaning** : Removal of punctuation marks , removal of ASCII characters , removal of urls , etc. Making all the chracters lowercase.
- **Tokenizing the Data :** Breaking it down into smaller units, such as words or phrases, to analyze or process it more effectively.
- **Removing the stop words and other unnecessary words**
- **Stemming the Data :** Reducing words to their root form, aiding in text analysis by grouping variations of a word together.
- **Converting the tokens into string and unidecoding it**

METHODOLOGY (CNTD.)

Data Preprocessing

- **Vectorizing the Data :** Converting text data into numerical vectors based on term frequency-inverse document frequency, representing each term's importance within a speech corpus.

Finally the data preprocessing part of the data is completed and we will now move on to the data modelling part.

METHODOLOGY (CNTD.)

Unsupervised Data Clustering

- **We will use K-means clustering to cluster our data into small groups.**

Why to use this ???

- **No prior assumptions:** K-means doesn't require prior knowledge about the dataset, making it suitable for unsupervised learning.
- **Automatic pattern detection:** It automatically identifies patterns in data by iteratively assigning data points to clusters and optimizing centroids.
- **Fast convergence:** Typically converges quickly, providing results even with high-dimensional data.

But still there is an issue ??

METHODOLOGY (CNTD.)

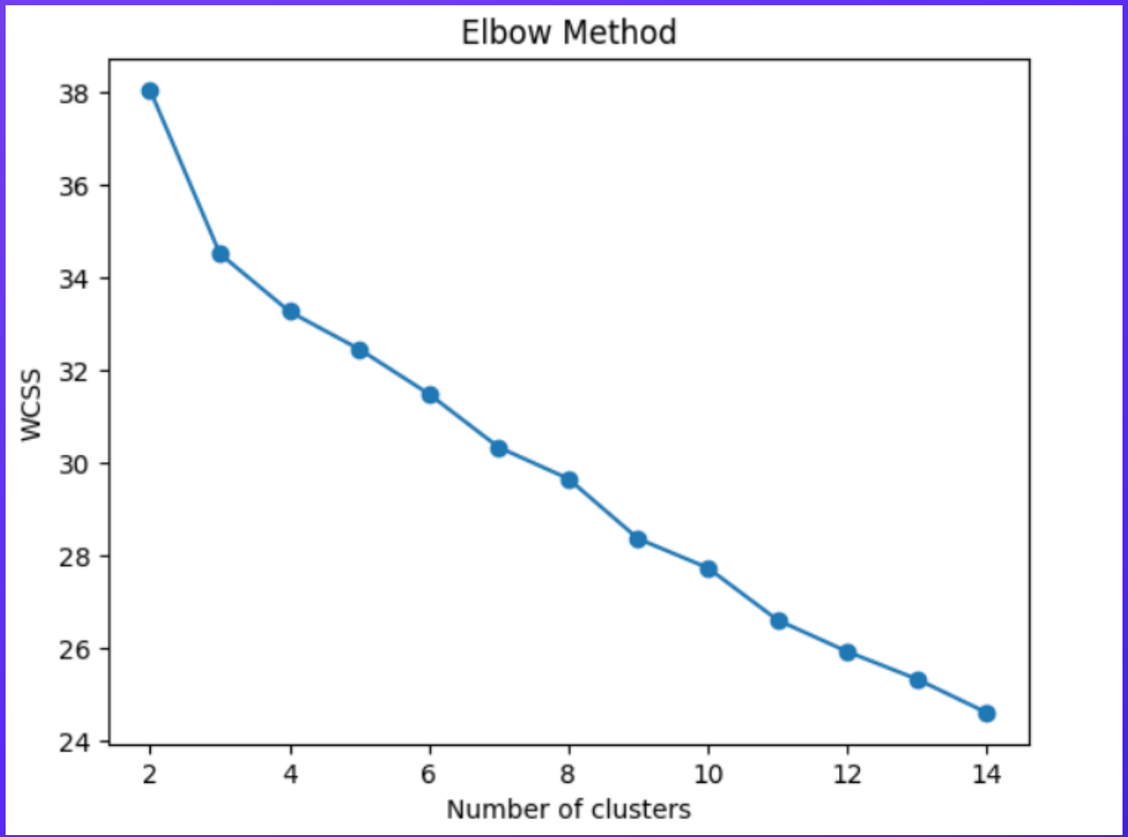
Selecting an appropriate value of K

We do not know that how many clusters of data are there in our data. How to know that ?

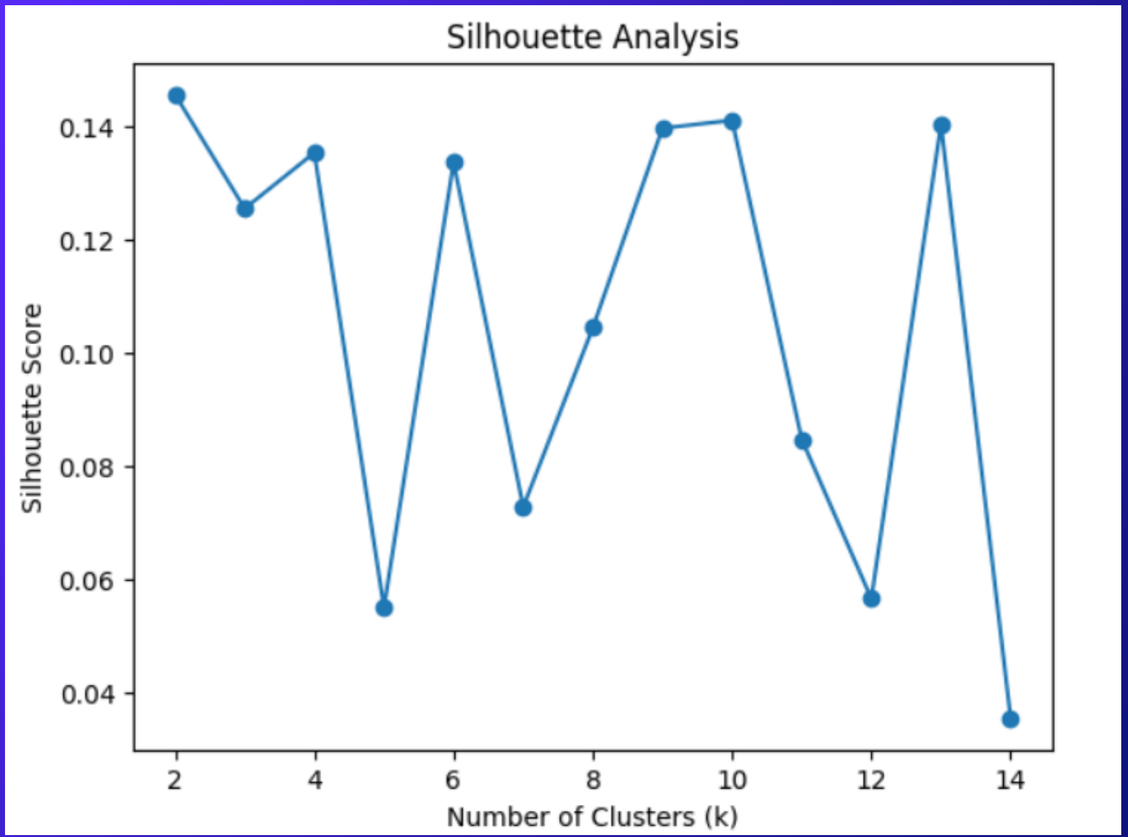
- **Elbow Method**
- **Sihouette Score**

METHODOLOGY (CNTD.)

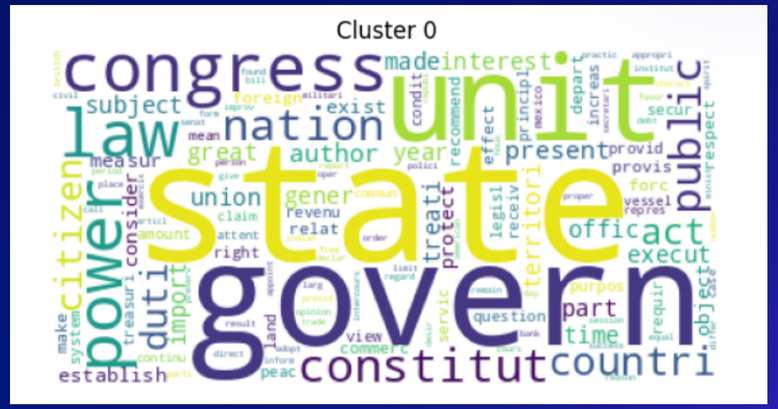
ELBOW METHOD

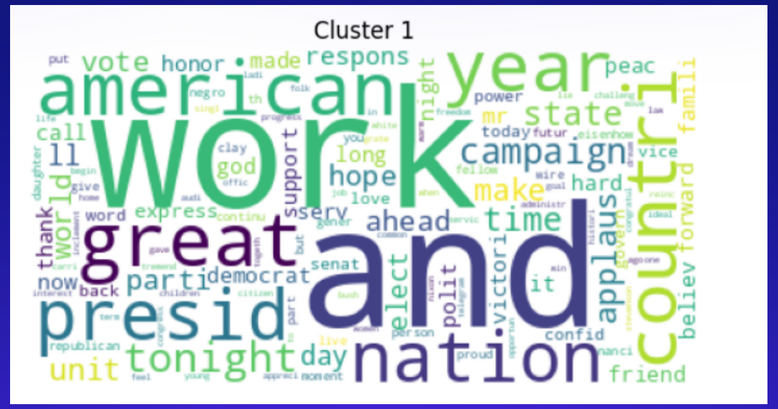


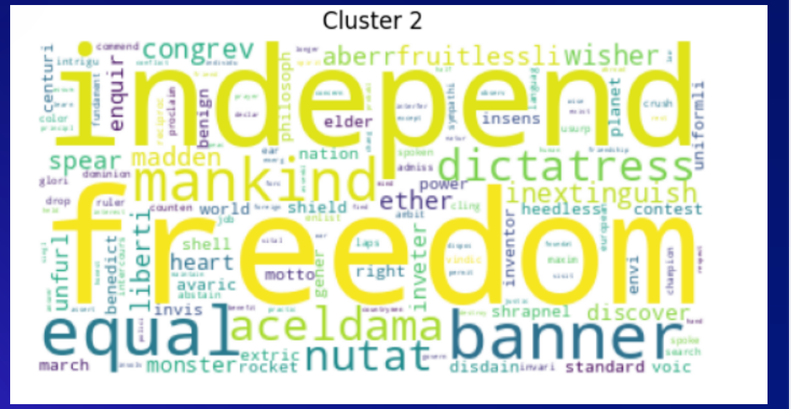
Silhouette Score

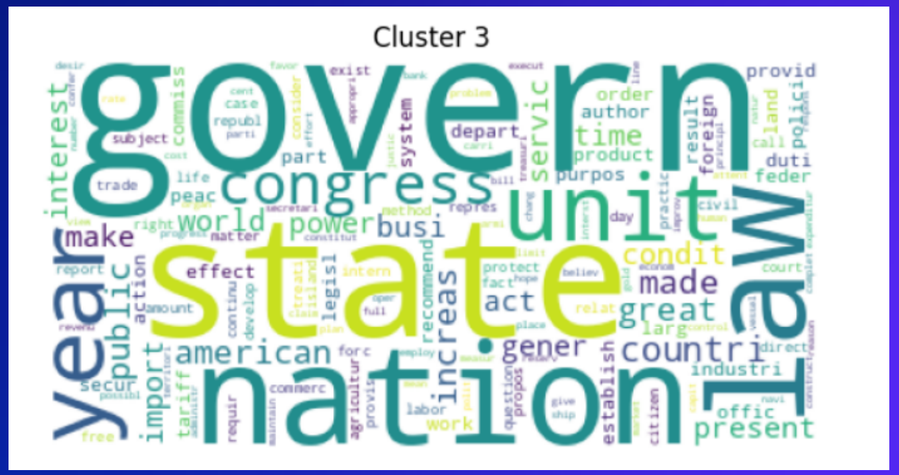


RESULT & ANALYSIS

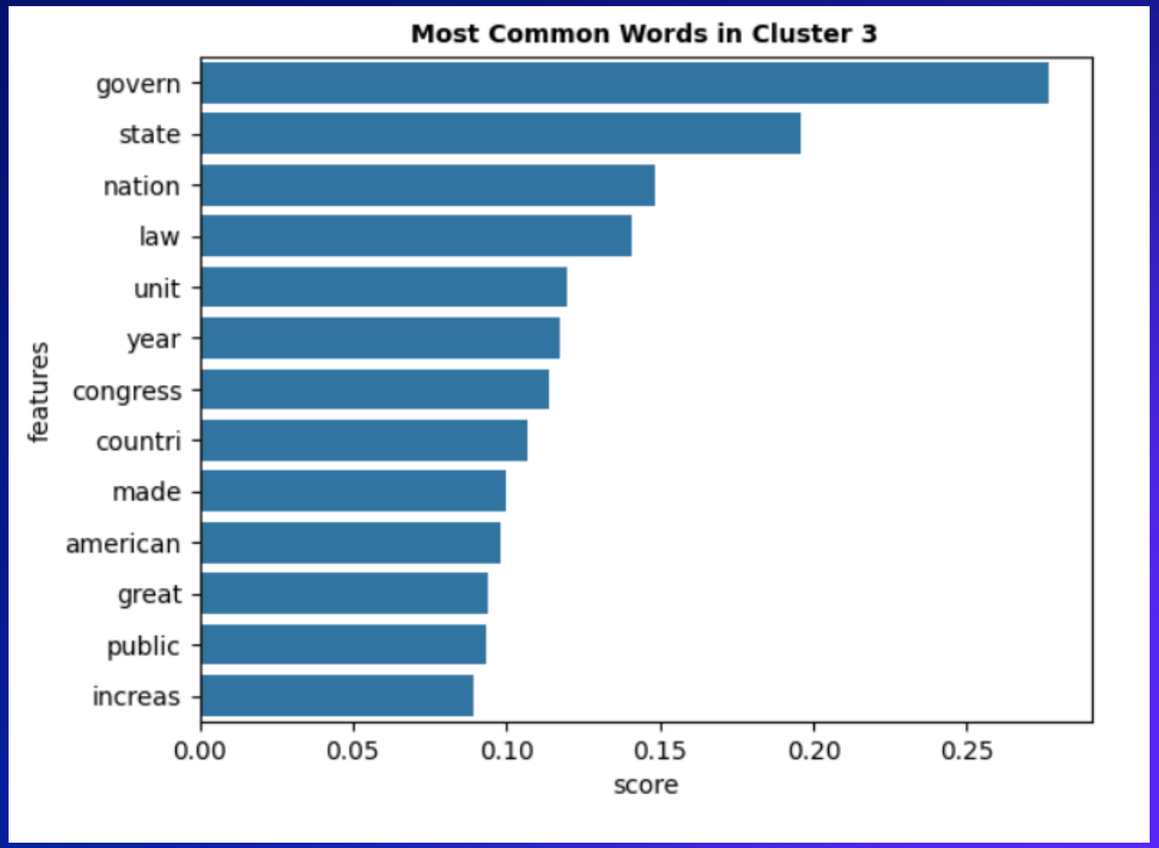








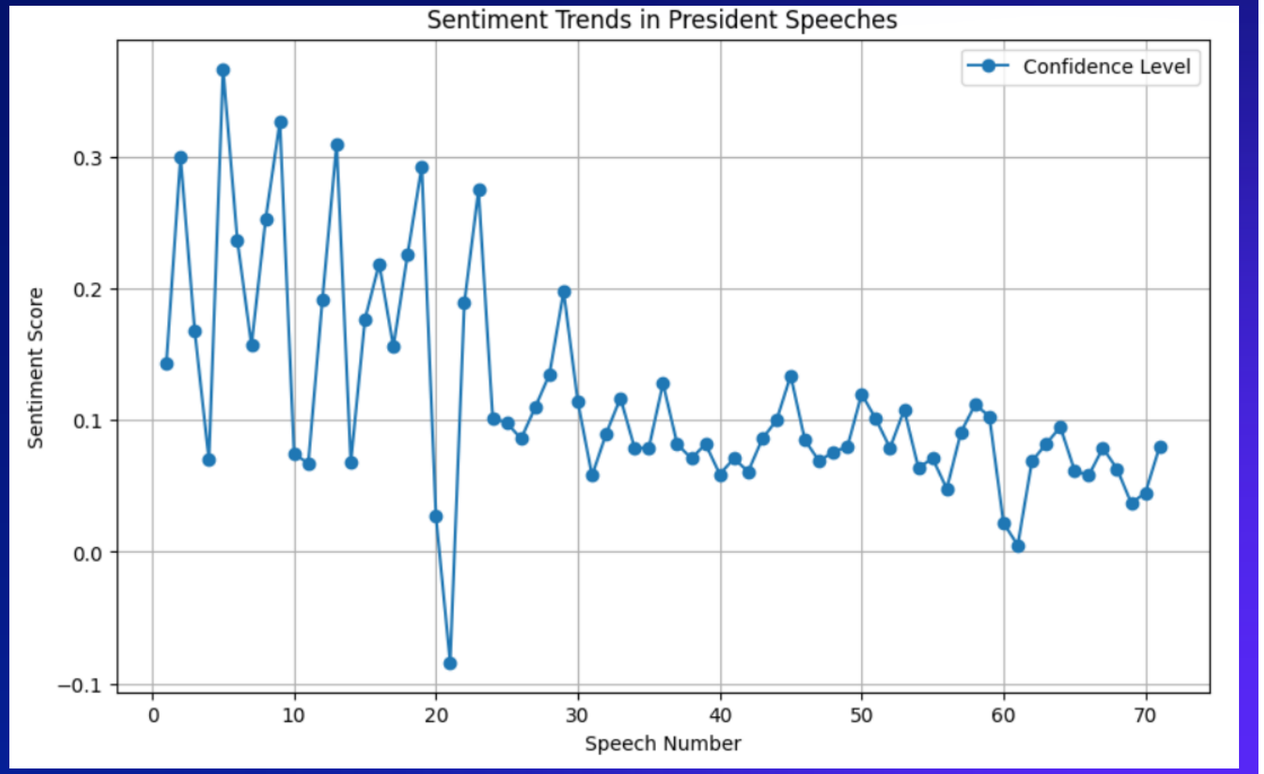
RESULT & ANALYSIS

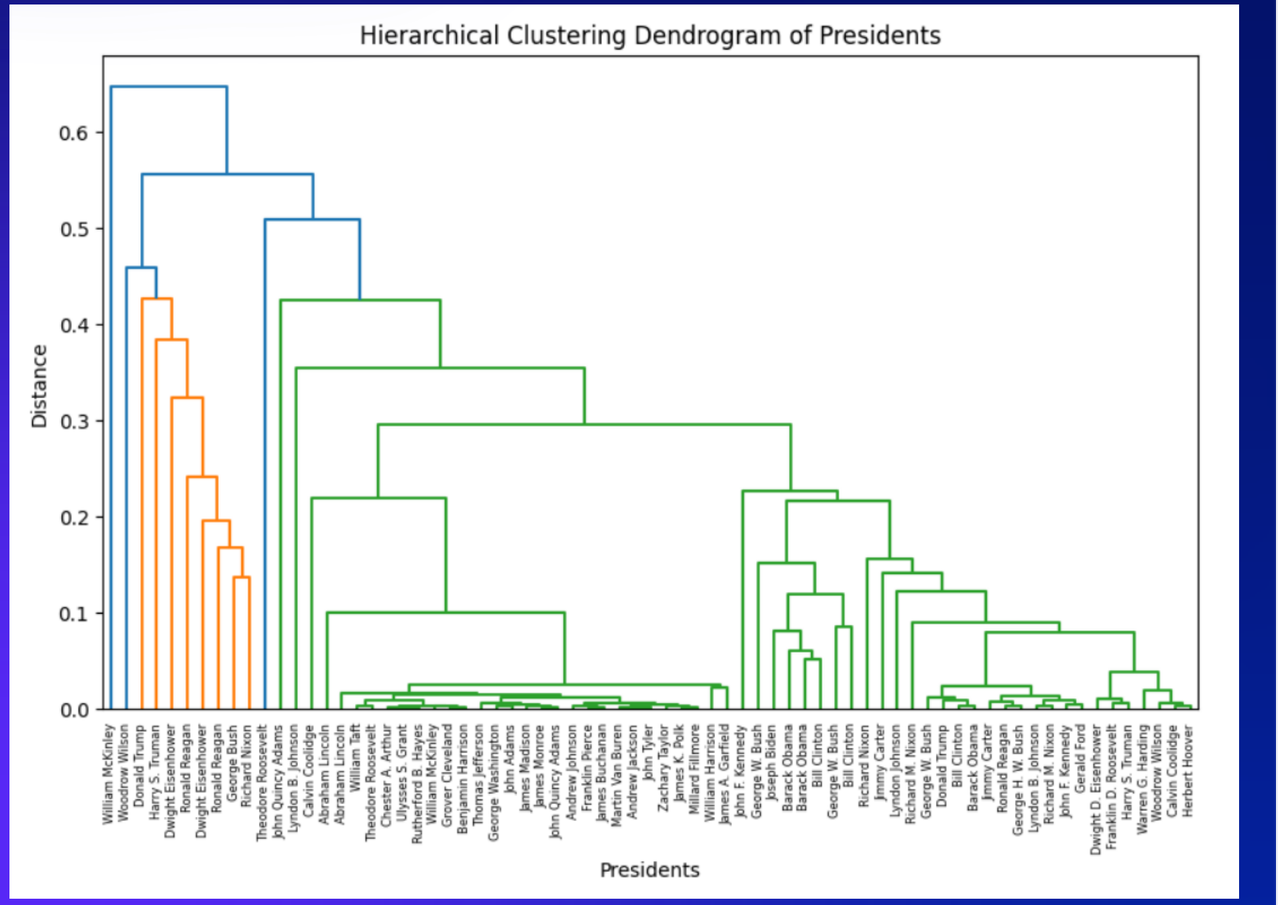



```
cluster3
```

```
['Grover Cleveland',
 'Benjamin Harrison',
 'William McKinley',
 'William Taft',
 'Woodrow Wilson',
 'Warren G. Harding',
 'Calvin Coolidge',
 'Herbert Hoover',
 'Theodore Roosevelt']
```

RESULT & ANALYSIS





CLUSTER AND RANKING PREDICTION

In order to predict a new speech is given to our model.

- It will determine the cluster by using K-means method and place the speech in that particular cluster.
- For predicting the ranking we are going to find the top 5 nearest neighbors of the speech and then do the averag of the ranking of those speakers of those five speeches to find the ranking of the speaker of the new speech.

CLUSTER AND RANKING PREDICTION

```
# Speech by George Washington
speech = "Whereas it appears that a state of war exists between Austria, Prussia,
tell_cluster_and_popscore(speech)
```

```
Speech shows resemblance to Cluster  0
Predicted Popularity Score of this speech :  567.2
```

It detected correct cluster as George Washington belonged to the first cluster

CHANLLENGES FACED

01

- To include all the leaders of the world is an easy task. As the speeches are not easy to scrap from some websites because of the poor structures of the websites. Moreover the speeches are not avaliable in one particular language making this dataset difficult to get.

02

- The ranking system which I have used is not universal. The fact is there is no such universal ranking system which inlcude all the leaders of the world.

THANK YOU!