**Assignment for Week 6 - K-Means and Hierarchical Clustering**

**Exerise:**

**KMeans and Hierarchial Clustering**

Collect data from twitter using your choice of hashtag (at least 1000 tweets). To gather your Twitter data, I would recommend using the python library tweepy.

Note: Reddit data can be used instead of Twitter data for this assignment.

* You will use both K-Means and HCA to cluster the collected data
* Draw a dendrogram
* Identify number of optimum clusters, and justify how do you pick number of clusters
* List the top 10 most frequent words
* Visualize using wordclouds
* Summarize your findings

**Assignment Solution**

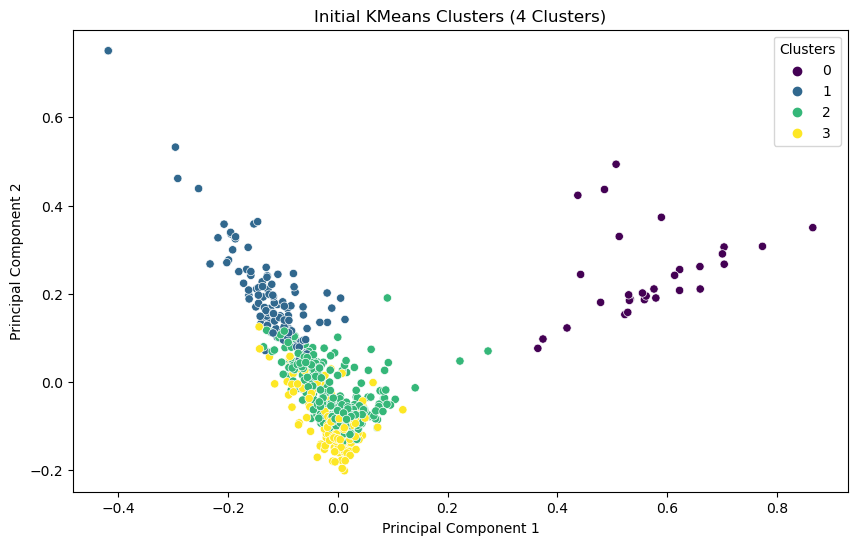
We have considered the posts on ‘Climate Change’ from Reddit and used it to analyze the two major clustering techniques, KMeans and Hierarchical clustering. The 1000 posts are extracted based on posts that have ‘climate change.’

1. **Data collection and preprocessing:**

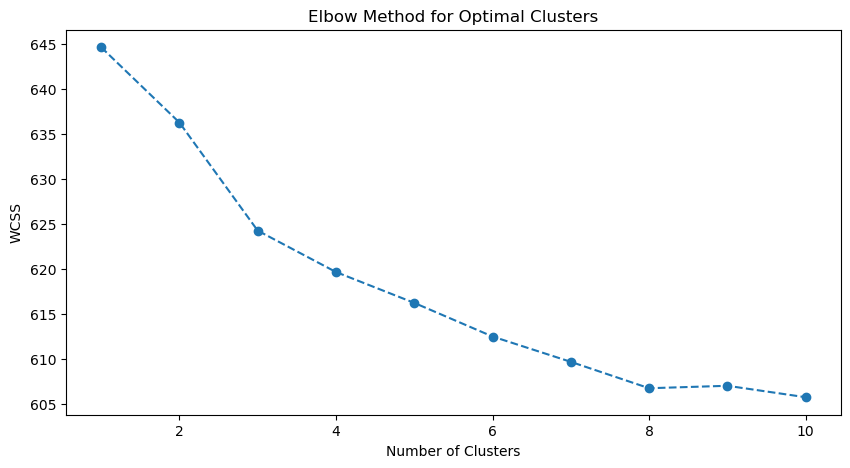
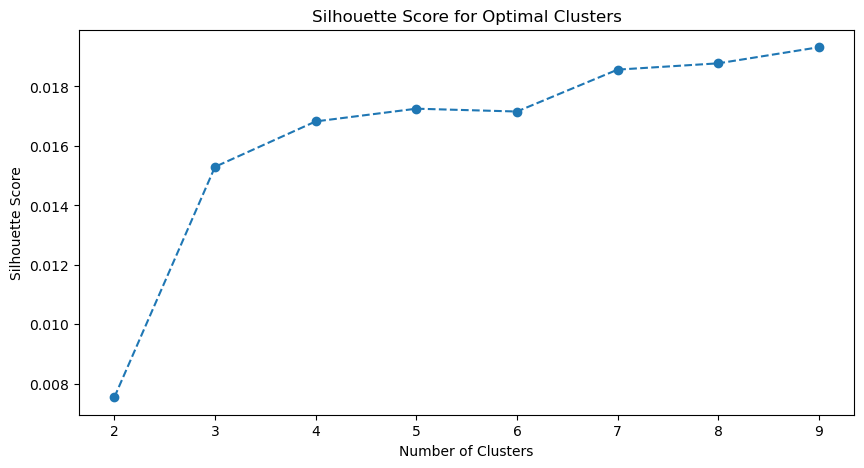
* Data was collected from the "climatechange" subreddit, with 1000 posts retrieved.
* Text data was preprocessed by removing URLs, special characters, and stopwords, and converting to lowercase.
* The TF-IDF vectorization method was used to convert the text data into numerical features, with a maximum of 1000 features.

1. **K-Means Clustering:**

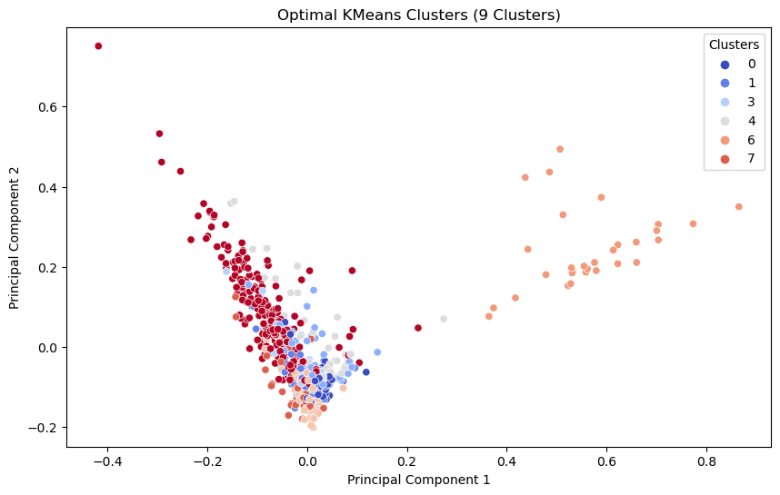
* Initially, K-Means clustering was performed with 4 clusters.
* The visualization of ‘4 clusters ‘ is done using PCA



* The optimal number of clusters is identified by iteratively changing the number of clusters and evaluating the WCC and silhoutte score
* A silhouette score analysis was conducted to determine the optimal number of clusters, testing from 2 to 9 clusters.



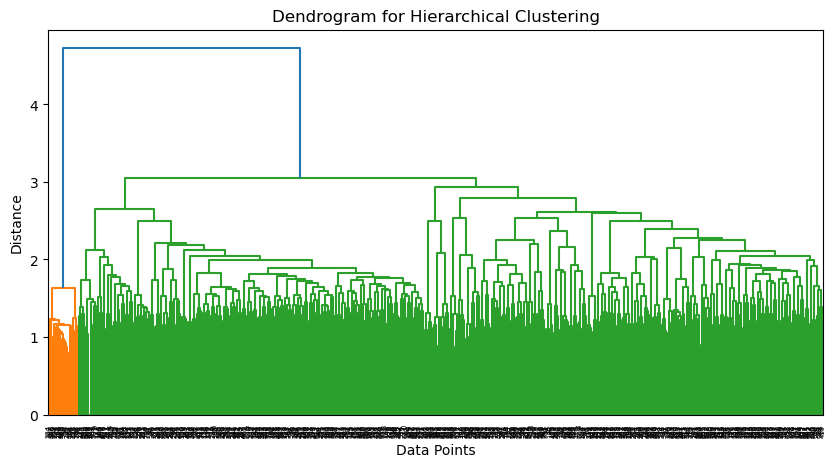
* The silhouette score plot shows that the optimal number of clusters is 9, as the score decreases after that point.
* However, we can say that the rise is sharp up to clusters ‘4’ and there is low rate after that.
* The visualization of ‘9’ clusters is given below



Clearly, the clusters are mixing, only cluster 6 seems separable from others and rest of clusters are close to each other.

1. **Hierarchical Clustering:**

* Incase of Hierarchial clustering, there is no need to mentioned the number of clusters the algorithm will split the data hierarchially.
* dendrogram was created using Ward's method for hierarchical clustering.



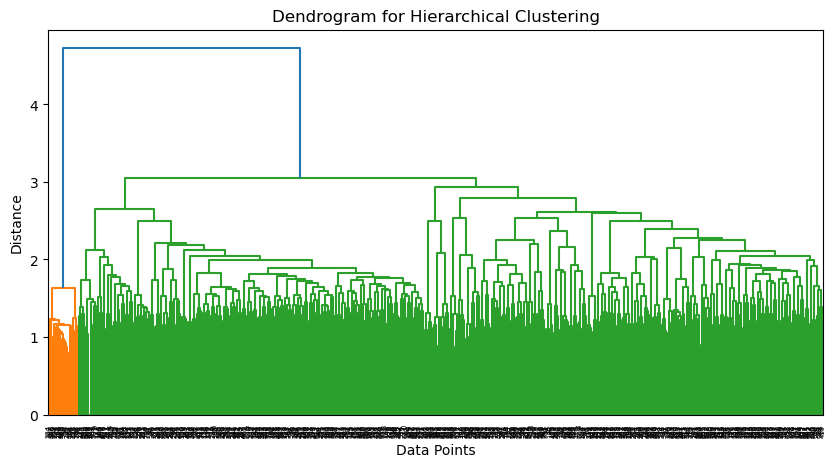
* The dendrogram provides a visual representation of how the data points are grouped at different levels of similarity.
* It shows several distinct clusters at different heights, which can guide the choice of the number of clusters.

1. **Optimal Number of Clusters:**

* **Kmeans:** The silhouette score analysis suggests that the optimal number of clusters is 9, as it has the highest score and lowest WCSS. However, the score is still low, and the data is not separated efficiently. This might be because very little data (1000 posts) were collected using a single key word, ‘climate change’. We might increase the number of posts, selecting the posts using other relevant key words such as ‘global warming’, ‘temperature rise’ etc.

Using Kmeans, the optimal number of clusters is shown as ‘9’ based on Shilotte scores however the score is not satisfactory, and the rate of increase in score is much less than ‘4’

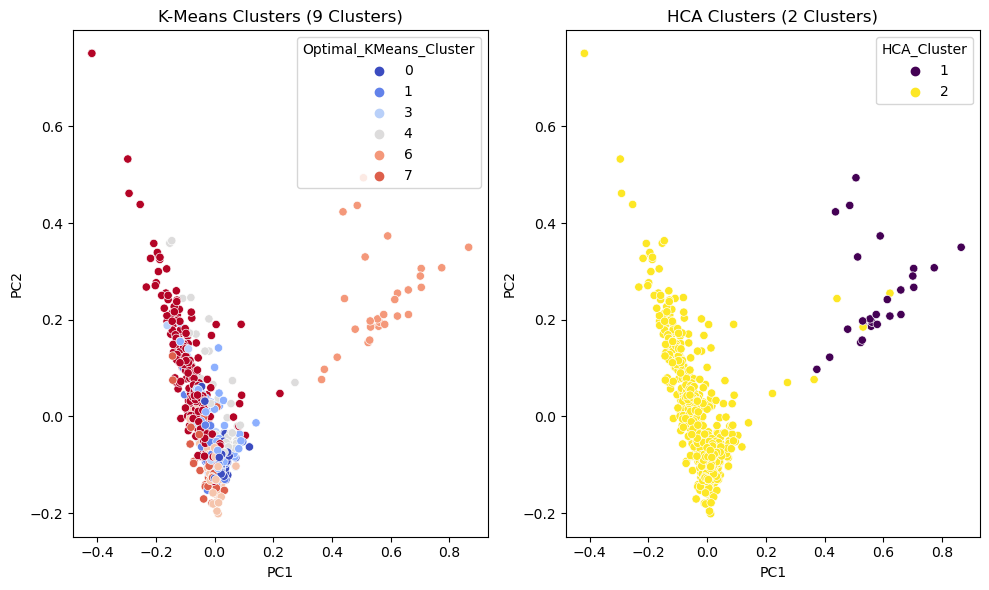
* **Hierarchial:** In hierarchical clustering, the optimal number of clusters is determined by analyzing the dendrogram and using quantitative methods. The dendrogram represents how data points are merged into clusters at different levels. To determine the optimal number of clusters Look for the largest vertical distance (gap) between two merged clusters that is not crossed by horizontal lines. Draw a horizontal line at this largest gap; the number of vertical lines intersected gives the number of clusters.



Based on HCA optimal number of clusters is found to be ‘2’. This method is subjective but provides an intuitive understanding.

1. **Choice of optimal number of clusters:**

We have found out that optimal number of clusters lie between ‘4’ and ‘9’ based on Kmeans. Using HCA the optimal number of clusters is found to be ‘2’ . The visualization i.e., scatter plots of PCA is used to select the optimal number of clusters.



Based on this visualizations, I support optimal number of clusters as ‘2’ they seem to differentiate the data appropriately, where as 9 clusters shows the inseparable clusters.

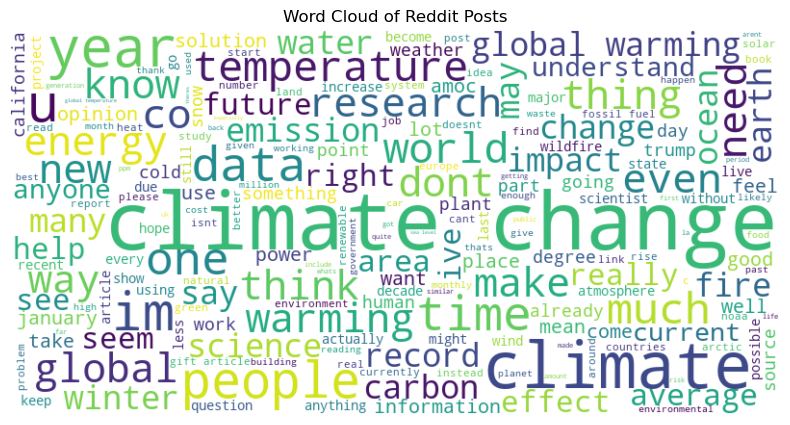
1. **Top 10 Most Frequent Words:**

* The top 10 most frequent words in the preprocessed text data were identified using the word counts.
* These words likely represent the most common themes or topics discussed in the "climatechange" subreddit.
* Top 10 most frequent words and their frequencies are

Top 10 words: [('climate', 567), ('change', 362), ('would', 175), ('global', 155), ('like', 142), ('im', 142), ('warming', 134), ('years', 118), ('us', 117), ('data', 115)]

1. **Word Cloud Visualization:**

* A word cloud was generated to visually represent the most frequent words in the dataset.
* This visualization provides an intuitive way to understand the main topics and themes in the climate change discussions.



1. **Summary & Conclusions:**
2. Clustering Performance:
   * Both K-Means and Hierarchical Clustering methods were successfully applied to the Reddit data.
   * The optimal number of clusters appears to be between 4 and 9, based on the silhouette score analysis.
   * The dendrogram from hierarchical clustering found that the optimal number of clusters is ‘2’
   * Overall, we have selected the optimal number of clusters as ‘2’ as there is clear visualization of separation of clusters when clusters is ‘2’
3. Topic Identification:
   * The top 10 most frequent words and the word cloud visualization reveal the main topics of discussion in the "climatechange" subreddit.
   * These likely include themes related to global warming, environmental policies, people, carbon, temperature.
4. Data Characteristics:
   * The Reddit posts from the "climatechange" subreddit show distinct groupings, suggesting several main themes or perspectives in the discussions.
   * The relatively low optimal number of clusters (2) indicates that the discussions may be polarized or focused on a few main topics.

**Summary:**

This analysis of Reddit posts from the "climatechange" subreddit using K-Means and Hierarchical Clustering techniques has revealed distinct groupings in the discussions. The optimal number of clusters appears to be between 2, suggesting that the conversations may be focused on a few main themes or perspectives. The most frequent words and word cloud visualization provide insights into the key discussion topics, likely including global warming, environment, temperature and count.

The clustering methods effectively grouped similar posts together, allowing for the identification of main themes in the climate change discussions. This approach can be valuable for understanding public discourse on climate change, identifying key concerns and topics, and potentially tracking how these discussions evolve over time. The analysis can be improved further by extracting more posts and removing some trivial words such as Iam, would. Also, including other keywords relevant to climate change, such as ‘global warming,’’ temperature rise’, etc., can improve the analysis.