**Alternative assessment – final report**

By: Yael Dickstein

1. **Results for k-means**:
   1. Below is the quality of the division into the various clusters according to the SSE and the silhouette criteria:

תמונה שמכילה טקסט, צילום מסך, גופן, מספר

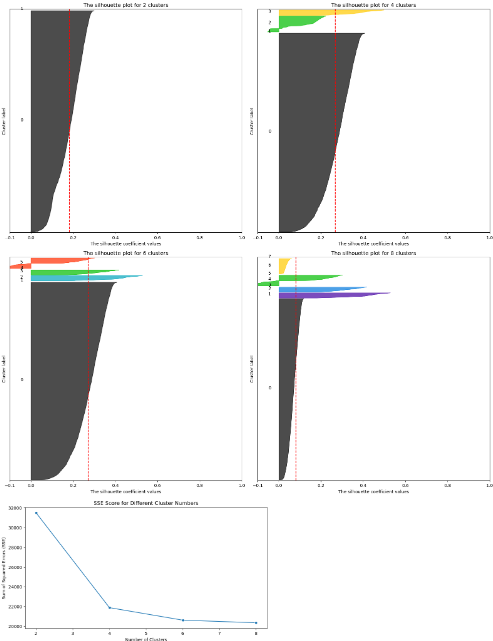
התיאור נוצר באופן אוטומטי

* 1. **Optimal number of clusters:**

Looking at the results (found in clusters\_df), there seems to be a drop in the SSE result from 4 to 6 clusters, but the drop becomes less significant as the number of clusters increases further. The silhouette score (ranges from -1 to 1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters) is the highest for 6 clusters (0.270779), which indicates that the data is well clustered, and as the number of clusters increases, the silhouette score decreases, indicating poorer cluster quality.

In addition, we can see what was also stated in the graphs, according to the graphs of silhouette analysis, it seems that any number of clusters other than 6 or 4 are a bad choice for the data, due to the presence of clusters with silhouette scores below the average, as well as due to the wide fluctuations in the size of the silhouette plots (thickness), according to the graph of SSE Score for different Cluster Numbers There is a drop between 4 and 6 clusters and the more moderate drop between the other clusters.

**Therefore,** **since the silhouette score is higher and the SSE score is lower, it seems** **that 6 clusters may be the optimal choice**.

****

* 1. **Maximum and minimum examples per cluster**:

According to the selection of the optimal number of clusters the **maximum** number of examples is **3098**, and the **minimum** number of examples is **1**.

* 1. **The topic of each cluster: from the words listed in each cluster, it appears that**:

**Cluster 1**: This cluster seems to contain a wide range of words related to **politics, activism, government, and public affairs**. Words like "abortion," "campaign," "election," "democrat," "republican," "president," and "senate" indicate a focus on political discourse. Additionally, words like "protest," "activist," "agenda," and "demonstrate" suggest involvement in social and political activism.

**Cluster 2**: This cluster contains only one word, "hit," which **doesn't provide much insight into the theme** or focus of this cluster. It's possible that this cluster represents a small subset of data with a unique characteristic that is not captured by the words listed.

**Cluster 4**: The words in this cluster are diverse but seem to focus on various topics such as **news, politics, and general discourse**. Words like "mccain," "november," "program," "report," and "war" suggest a mix of political and current events-related content.

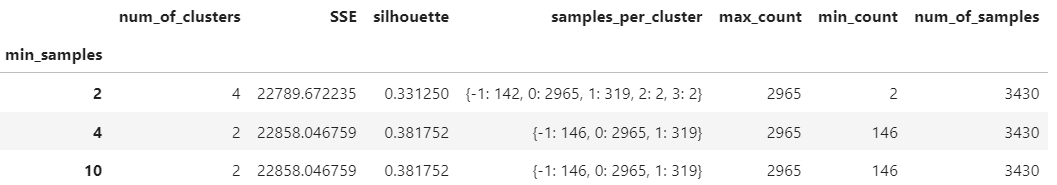
**Cluster 5**: Like Cluster 2, this cluster contains only one word, "brought," which **doesn't offer much insight into the cluster's theme** or focus. It's possible that this cluster represents a small subset of data with unique characteristics not captured by the word listed.

**Cluster 6**: This cluster appears to focus on topics related to d**ata analysis, elections, and possibly marketing or advertising**. Words like "calculate," "electoral," "affiliate," "article," "campaign," and "debate" suggest a focus on data-driven decision-making, electoral processes, and promotional activities.

Based on this analysis, it's clear that Cluster 1 and Cluster 6 have more meaningful words that provide insights into their themes or focuses. Cluster 1 seems to focus on political discourse and activism, while Cluster 6 appears to involve data analysis and electoral processes. Clusters 2, 4, and 5 contain fewer words or words that don't offer clear insights into their themes.

Furthermore, it seems that there are no more than 5 clusters, but the values we were asked to enter are 2,4,6,8, so one cluster has no values at all. In addition, since there are no values in this cluster, the minimum count in the previous section for the cluster is 1 because it does not find a cluster and fails to count zero.

1. **Results for DBSCAN**:
   1. Below is the effect of three values for the min\_samples parameter on the clustering:

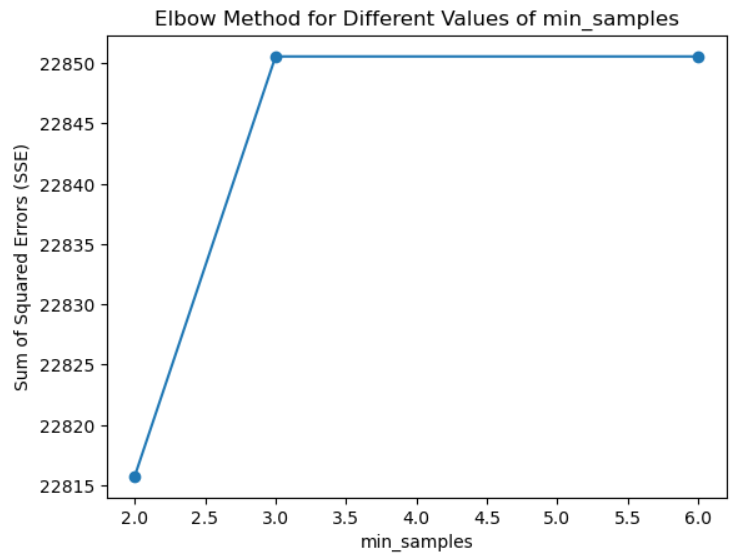


* 1. **Optimal value of min\_samples**:

The min\_samples parameter in DBSCAN affects the clustering result by specifying the minimum number of samples required to form a dense region (i.e., a core point), a high value of min\_samples will result in fewer clusters and a low value of min\_samples will result in more clusters.

In the case of my data there is a slight increase in SSE for the value min\_samples=3 and min\_samples=2, however this increase is small compared to the increase in the silhouette score which is higher for min\_samples=3 (0.405253) (the ratio of the increase in the SSE score is 1.001 compared to the ratio of the increase in the the parallelogram which is 1.219).

**Therefore, I will choose the value min\_samples=3 as an optimal value**.



* 1. **Maximum and minimum examples per cluster**:

According to the selection of the optimal value of min\_samples the **maximum** number of examples is **3006**, and the **minimum** number of examples is **104**.

* 1. **The topic of each cluster**:

**Cluster -1 (Noise)**:The words in Cluster -1 **don't seem to follow a specific theme or topic**, they appear to be a mixture of various words that don't necessarily relate to a specific concept or category. Some of these words include "acknowledge," "cases," "credit," "pressure," "trade," and "weekly." This indicates that Cluster -1 represents noise or outliers in the data.

**Cluster 0**:The words in Cluster 0 seem to relate to **political and governmental topics**. There are mentions of political figures such as "bush," "cheney," "kerry," and "lieberman." Additionally, there are words related to politics and governance, such as "campaign," "congress," "election," "government," "policy," and "presidential." This suggests that Cluster 0 may represent discussions or articles related to politics, elections, and government affairs.

**Cluster 1**:The words in Cluster 1 seem to cover a range of topics, **including news, events, and general discussions**. There are mentions of specific events or incidents such as "debate," "attack," "memorial," and "release." Additionally, there are words related to media and communication, such as "article," "blog," "conversation," "magazine," and "report." This cluster seems to encompass a broader range of topics compared to the more politically focused Cluster 0.

**In summary**, based on the words associated with each cluster:

**Cluster -1 represents noise or outliers.**

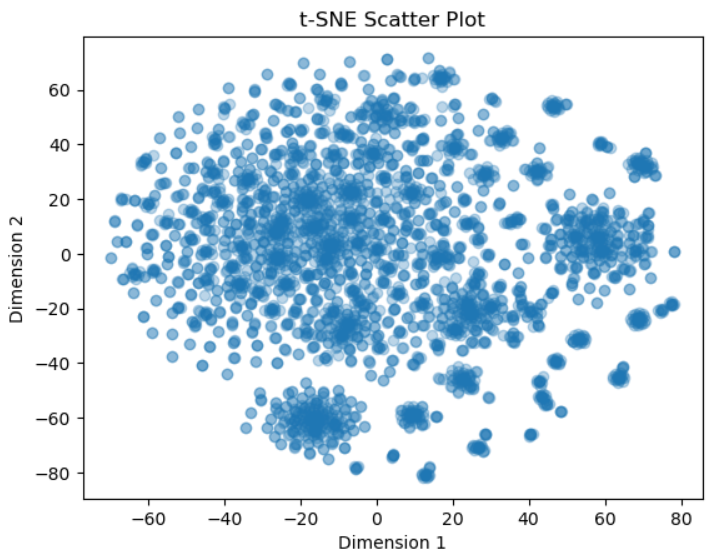
**Cluster 0 likely focuses on political and governmental topics.**

**Cluster 1 covers a broader range of topics, such as news, events, and general discussions.**

1. **Results for TSNE**:
2. In many cases, t-SNE produces meaningful embeddings even without normalization, therefore I'll try this algorithm on normalized and denormalized data.

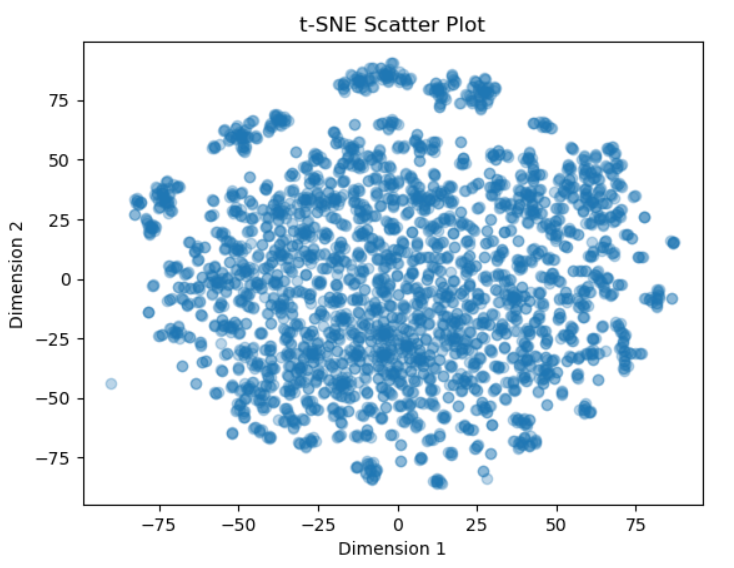
**Understanding the number of clusters from the visualization** **for the normalized data**:

According to the two-dimensional scatter plot we received for the t-SNE algorithm, it seems that our information has cluster 3 or 4.



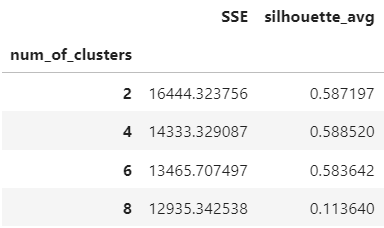
**Understanding the number of clusters from the visualization** **for the denormalized data**:

According to the two-dimensional scatter plot we received for the t-SNE algorithm, it seems that our information has cluster 2 or 4.



The visualization gives a general direction regarding the number of clusters with and without normalization. It seems that in both cases there are between 2-4 clusters, so far in the assignment we have received for K-Means 3 significant clusters and for DBSCAN 2 therefore **the visualization supports the findings but does not help to make them more precise**.

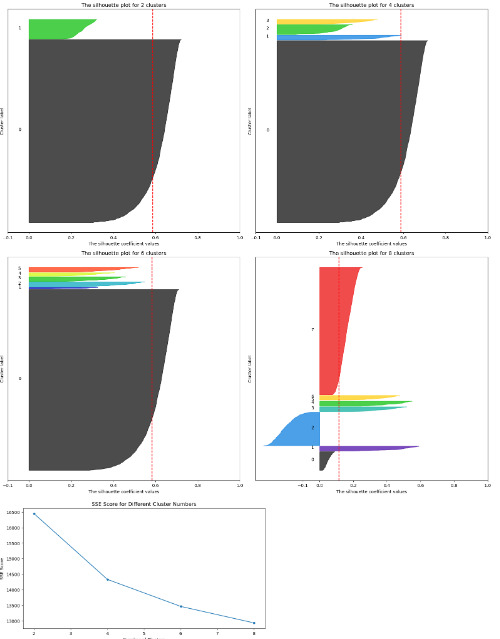
1. **Results for PCA and k-means**:
   1. Below is the quality of the division into the various clusters according to the SSE and the silhouette criteria:



* 1. **Optimal number of clusters**:

As in the first part where we ran K-Means without PCA the deliberation is between 6 clusters and 4 clusters, this time the relative decrease (1.064) in the SSE score between 4 and 6 clusters is greater than the relative decrease in the silhouette score (1.008), therefore I will still choose - 6 clusters as an optimal number of clusters.

**Therefore,** **even though the silhouette score is lower the SSE score is lower too, it seems** **that 6 clusters may be the optimal choice**.



* 1. **Maximum and minimum examples per cluster**:

According to the selection of the optimal number of clusters the **maximum** number of examples is **3099**, and the **minimum** number of examples is **24**.

* 1. **From the words listed in each cluster, it appears that**:

**Cluster 1**: This cluster includes words related to **politics, activism, government, and societal matters**. Words like "abortion," "activist," "administration," "bush," "campaign," "congress," "democracy," "election," "iraq," "president," and "terrorism" indicate a focus on political and social issues.

**Cluster 3**: This cluster seems to revolve around words related to **electoral processes, campaigning, and political strategy**. Words like "candidate," "debate," "electoral," "campaign," "endorsement," "poll," "vote," and "win" suggest a focus on election-related topics.

**Cluster 4**: This cluster contains a diverse range of words that don't seem to fit into a specific thematic category. It includes words like "magazine," "mccain," "mother," "newspaper," "san," "war," "web," and "wrong." These words cover a **broad spectrum of topics**, indicating a less obvious theme compared to the other clusters.

Here too we see that there are empty clusters as we saw last time without PCA.

1. Number of features before PCA: 1545

Number of features after PCA: 344

**Number of features lost during PCA: 1201**

**The use of PCA** helped refine the result, obtained by the k-means algorithm without PCA. The current step increased the silhouette score and decreased the SSE score and therefore **helped to achieve a better, and more decisive clustering**.

* All the algorithms produced similar results – dividing the original data into two major topics – namely **politics, activism, government** and **electoral processes, campaigning, and political strategy**.

**It can be established that according to the number of clusters – there is consistency in the assignment**. The **K-Means algorithm, ran without PCA** has shown that although 6 clusters seemed to be the optimal choice, 5 clusters were produced and one empty cluster, from the 5 produced 2 clusters were one-word clusters, and 1 cluster with broad and unfocused set of words, which means **2 main clusters in terms of their topics and the values** ​​entered into them; when the same algorithm ran **with PCA** it was easier to see the distribution of the clusters since this time 3 clusters came out with values ​​(although again the value 6 was optimal, maybe I would change the number of clusters in the work according to the findings and run again for testing, however the run was performed with the requested values) **and 2 of them were main** because the third was a set of words without a clear theme.

In **DBSCAN we've got 2 clusters from the beginning** with clear subjects and another cluster with noises. **The t-SNE supported the findings with the help of the visualization** (2 to 4 clusters) with and without normalization.