מגיש : שניאור יום טוב

תקציר על הData בפרוייקט שלי:
הקובץ "dailykos"- מכיל נתונים על 3,430 כתבות חדשות או בלוגים שהתפרסמו ב- Kos Daily, בלוג פוליטי אמריקאי המפרסם מאמרי חדשות ודעה.
מאמרים אלה פורסמו בשנת 2004, והובילו לבחירות לנשיאות ארצות הברית. המועמדים המובילים היו הנשיא המכהן ג'ורג 'וו. בוש (רפובליקני) וג'ון קרי (דמוקרטי).
מדיניות חוץ הייתה נושא דומיננטי בבחירות, ובמיוחד הפלישה לעירק בשנת 2003.
כל אחד מהמשתנים במערך הנתונים הוא מילה שהופיעה בלפחות 50 מאמרים שונים)
1,545 מילים בסך הכל.
(מערך המילים גוזם על פי חלק מהטכניקות הקיימות בניתוח טקסטים)
פיסוק הוסר והוסרו מילות עצירה.
(עבור כל מסמך, ערכי המשתנה הם מספר הפעמים שהמילה הופיעה במסמך).

בדו"ח הבא אני אסביר על המטלה ואתן הסברים על הבחירות שלי ועל הקוד שכתבתי.

הסרתי את העמודה "Document" מכיוון שהיא שונה משאר ה-Data. העמודה אינה מייצגת מילה ואת מספר המילים באותו מאמר כמו שאר העמודות. היא נועדה למספר את המאמרים ויכלה לבלב את הניתוח שלי, ולכן החלטתי להסיר אותה.

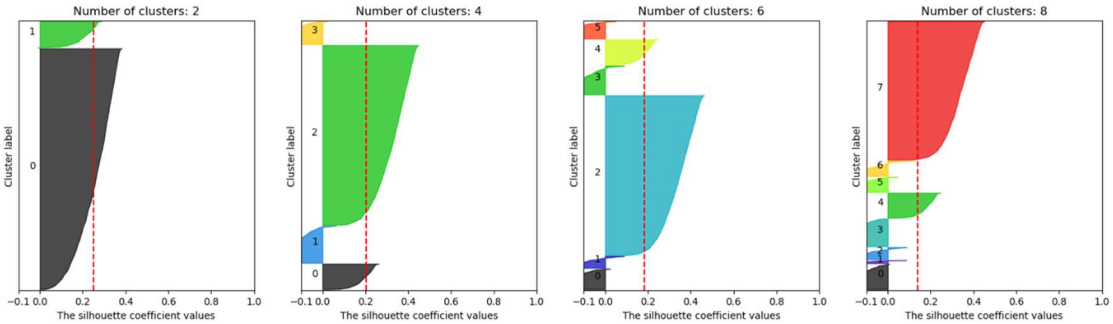
לפי שאסביר על הבחירה של המודלים אני רוצה לציין שהמודלים הם **unsupervised**.
השימוש במודלי unsupervised לניתוח המאמרים מאפשרת לזהות דפוסים וקטגוריות מבלי להסתמך על תוויות מוגדרות מראש. השימוש במודלים אלה מסייע בחלוקת הנתונים לאשכולות המייצגים קבוצות של מאמרים בעלי מאפיינים משותפים. כך ניתן לספק תובנות לגבי המגמות המרכזיות במאמרים שהתפרסמו באותה תקופה. הבחירה במודלים מסוג זה מאפשרת לנו לגשת לנתונים באובייקטיביות, ולמצוא תובנות חדשות ומעניינות.
חשוב לציין שאין זה מדע מדויק ומכאן והלאה הן תובנות שלי בהתאם להבנתי האישית.

מודל k-means:

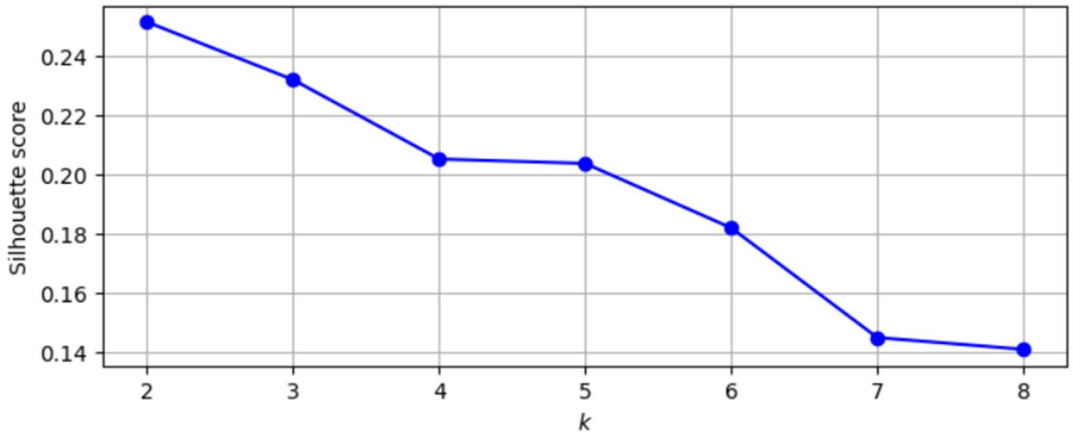
ככלל - הוא פועל טוב יותר כאשר יש מעט מאפיינים ולא כמו במקרה שלנו שיש לנו 1,545 עמודות.

את המודל בחנתי לפי שני מדדים "inertia" (SSE) ו "silhouette" .

- Silhouette

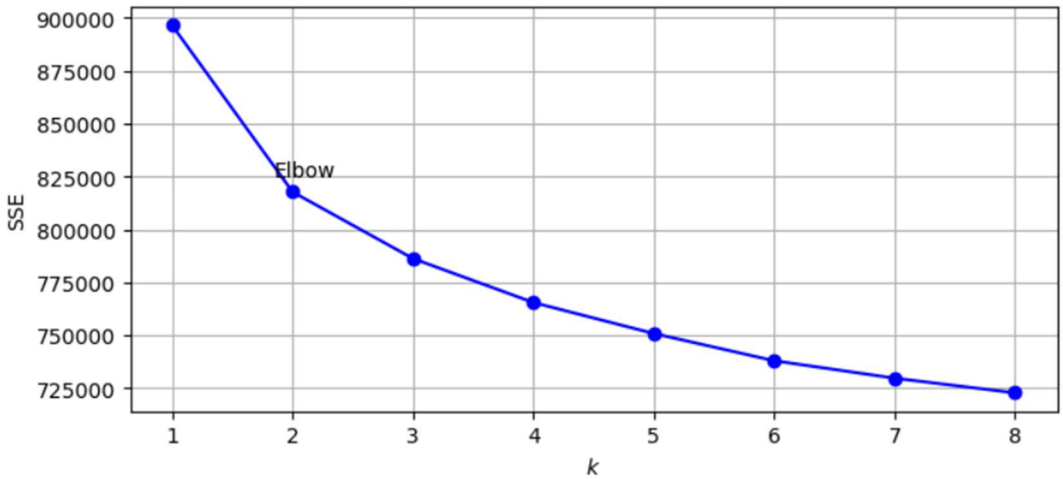


בגרפים בחנתי את מספר האשכולות 2,4,6 ו8 כדי לראות כל חלוקה ביחס לאחרים. הקו האדום האנכי מסמל היכן נמצא ממוצע הסילואט לכל חלוקה. הנתונים של כל אשכול בצבע שונה. לפי התמונה שנוצרה לי כאן נראה שהחלוקה ל2 אשכולות היא העדיפה עלי מכיוון ששם הנתונים עוברים את קו הסילואט מה שמעיד שהנקודות בכל אשכול קרובות יותר אחת לשניה. בניגוד לשאר החלוקות ששם חלק מהאשכולות לא עוברות את קו הסילואט שמעיד על חוסר בהירות בחלוקת האשכולות. שמתי לב שהעובי של אשכול 0 ביחד לאשכול 1 גדול פי כמה ממנו שזה מעיד שגודל האשכולות אינו זהה. בהינתן שגם כאשר חילקת ל 8 ,6 ,4 אשכולות שונים, נשאר תמיד אשכול גדול בהרבה ביחס לשאר האשכולות, הסקתי שזה לא יעזור לי לחלק ליותר אשכולות כדי להגיע לגודל זהה של כל אשכול.



ניתן לראות בגרף שהsilhouette של חלוקה ל2 אשכולות הוא הגבוה ביותר.

-SSE



במדד SSE חיפשתי את נק המרפק - הנקודה (מספר האשכולות) שבא השיפור של סך השגיאה הריבועית בנתונים מפסיק להיות חד ומתחיל להתמתן. כאשר בחנתי רק את אשכולות 2,4,6,8 לא ראיתי נק שאוכל להגיד בבירור שהיא נקודת מרפק ולכן החלטתי שאבחן טווח ארוך יותר של נק. לאחר שבחרתי בטווח ארוך יותר ניתן לראות כאן שהשיפוע מתמתן כאשר הוא מגיע ל2 אשכולות. (ברור שככול שיהיו יותר קלסטרים מדד הSSE יקטן. לדוגמא: אם נבחר את המספר הקלסטרים כמספר הרשומות בדטאה אז לא יהיה בכלל מרחק ריבועי ונגיע ל SSE=0 מה שלא מעיד על חלוקה טובה.)

בכל פעם שהרצתי את אלגוריתם **k-means** בחרתי את הפרמטרים הבאים:

```
KMeans(n_clusters=k, init='k-means++', n_init=10, max_iter=300,  random_state=42).fit(data)
```

init - k-means++ : הוא פרמטר שבו הנקודות ההתחלתיות נבחרות בצורה שמקנה יתרון לתפיסת המבנה האמיתי של הנתונים.

n_init - 10: מציין את מספר הפעמים שבהם ייבחרו נקודות ההתחלה בכל נסיון.

max_iter -300 : בחרתי מספר גדול כך שלא ישפיע עניין הרנדומליות מידי, שכל פעם יתקן את המרכזי אשכולות כך שגם אם היה נבחר בהתחלה נקודות מרכז לא טובות זה יתקן את עצמו.

לפי התוצאות וההסברים שציינתי, באלגוריתם **k-means** הבחירה שלי היא של **2 אשכולות**.

מודל DBSCAN :

במודל זה החלטתי לבחון שלושה min samples שהם: 2,3,21.

	eps	min_samples	n_clusters	n_outliers
15	16.0	2	9	808
16	16.0	3	2	822
17	16.0	21	2	825

```
For min_samples = 2, eps = 16: number of clusters: 9
Silhouette Score is 0.14576005924345706
For min_samples = 3, eps = 16: number of clusters: 2
Silhouette Score is 0.18973641496739918
For min_samples = 21, eps = 16: number of clusters: 2
Silhouette Score is 0.18976284432575555
```

ראיתי שהבחירה ב2 תביא לי את המספר הנמוך ביותר של מספר הלא מזוהים אך זה ייצור לי 9 אשכולות (עם המון אשכולות קטנים של 2 סמפלים באשכול) ועם סילואט נמוך מאוד מהשאר. ולכן החלטתי לוותר עליו. בין הבחירה ב21 ל3 התלבטתי רבות מכיוון שאם הייתי בוחר ב 21 הייתי בעצם יותר מבסס כל אשכול על סמך יותר נק בסביבה שלו ונותן יותר חוזק לכל אשכול בחרתי ב3. כדי לנסות ולהקטין כמה שיותר את הלא מזוהים.

בתמונה כאשר min sample = 3

```
dbscan_cluster
-1      822
 0     2360
 1      248
```

הבחירה ב3 = min sample יצרה לי 2 אשכולות ועוד 822 רשומות(מאמרים) לא מזוהים עם אף אשכול.

מספר דוגמאות המינימאלים והמקסימאלים באשכולות לפי כל אלגוריתם:

```
K-Means: Largest Cluster: 0 (3092 samples), Smallest Cluster: 1 (338 samples)
DBSCAN: Largest Cluster: 0 (2360 samples), Smallest Cluster: 1 (248 samples)
```

יצרתי קבצי csv לכל אשכול וענן מילים לחמש מילים הכי נפוצות.

Dbscan-

```
Top words in DBSCAN Cluster 1:
november     2568
poll         1170
vote         1028
challenge    1009
democrat      656
dtype: int64
```

```
Top words in DBSCAN Cluster 0:
bush         3961
kerry        2808
poll         2144
democrat     1923
republican   1400
dtype: int64
```

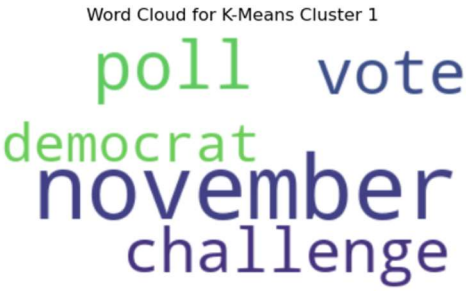
נראה שהמאמרים **באשכול 0** מדברים בעיקר על המועמדים עצמם ועל המפלגות האם של המועמדים. (**בוש וקרי** אלו המועמדים לבחירות בשנה שנה שנלקחו המאמרים, **poll זה סקר** אך זה משותף לשני האשכולות ויש גם **דמוקרטית ורפובליקנית** שדמוקרטיה היא מילה עם משמעות נוספת, להערכתי כאן זה כאלה שגם מדברים על המפלגה הדמוקרטית וגם על המילה דמוקרטיה כמו באשכול השני.)

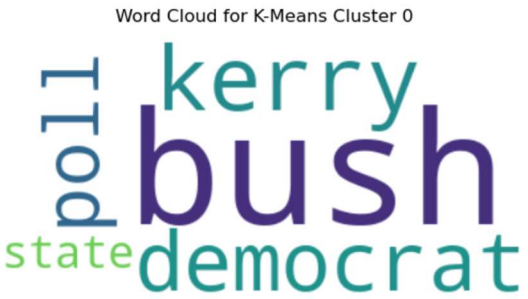
בעוד שלדעתי המאמרים **באשכול 1** מדברים על מועד(**november**) הבחירות(**vote**) של אותה שנה והאתגר(**challenge**) של הבחירות ועל הדמוקרטיה בכללי. ופחות מתייחסים למועמדים עצמם כמו באשכול הקודם.

- K-Means

```
Top words in K-Means Cluster 1:
november     3438
poll         1658
vote         1496
challenge    1387
democrat     1093
dtype: int64
```

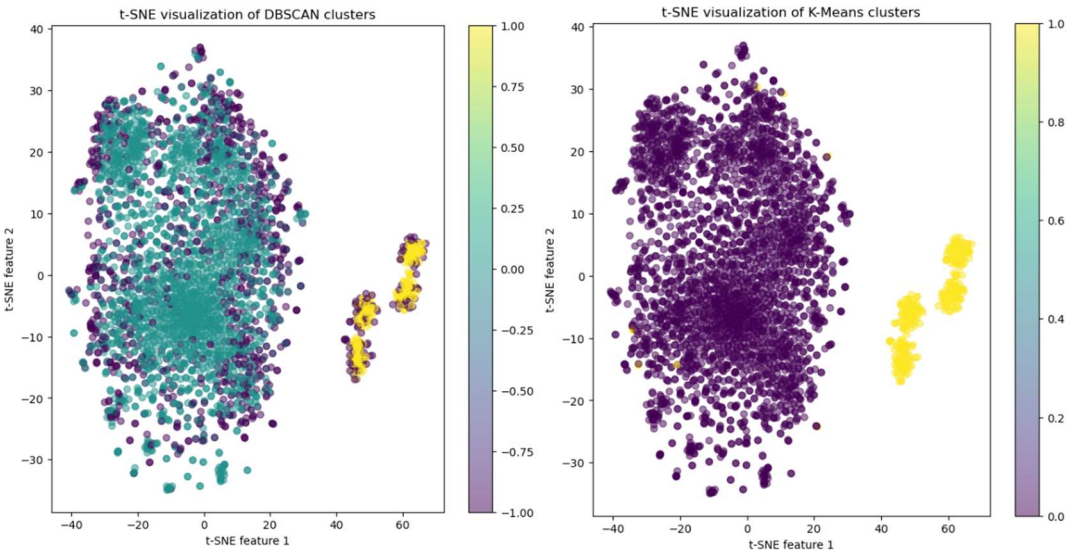
```
Top words in K-Means Cluster 0:
bush         6878
kerry        4488
democrat     3760
poll         3160
state        2718
dtype: int64
```





גם כאן נראה שהמילים הכי נפוצות סה"כ זהות חוץ מאשכול 0 שהמילה republican יצאה ובמקומה נכנסה לחמישייה המילה state. (המילה republican לפי K-Means היא במקום השישי באשכול הראשון.) להערכתי מדובר באותה חלוקה. באשכול הראשון מדברים על המועמדים עצמם בוש וקרי במספרים גדולים מהרבה מהאשכול השני בעוד שבאשכול השני מדובר על חודש הבחירות שיתקיימו והאתגר שהוא תומן בתוכו.

– T-SNE



נראה שמודל k-means חילק בצורה טובה את האשכולות חוץ מכמה אחדים צהובים בין הסגול נראה שיש חלוקה ברורה בין המאמרים. מה שתמוהה לי מיכון שהוא מודל שלא נהוג להשתמש בו עבור מספר רק של מאפיינים.

במודל dbscan הוא התקשה ללקט אליו את הקצוות של האשכולות (כאן הסגול הוא הבלתי מסווג) אפילו שבחרתי את ה min sample הקטן יותר (3) ולא הגדול (21) עדיין הן נשארות לא מסווגות. כאשר הגדלתי את eps מ16 ל17 זה הפך את כל הסמפלים להיות אשכול אחד.

לאחר התבוננות בגרף t-SNE נראה לי שיהיה נכון יותר לחלק את האשכולות או לשלושה או לחמישה אשכולות.

-PCA

PCA הוא הורדת מימדים. הוא נודע להקל בין היתר על איסוף הדאטה וזמן הרצת המודלים השונים, ולתת לנו אפשרות להגיע לאותם תוצאות או לעיתים לתוצאות טובות יותר בכמות משמעותית פחות של מאפיינים. כאן נתבקשתי לבצע הורדת מימדים שיסבירו 80% מהדאטה.

```
מספר הרכיבים הנדרשים כדי להסביר 80% מהווריאנס הוא : 275
מספר המאפיינים שנפלו בתהליך: 1270
```

החלטתי להריץ את כל השלבים מהתחלה דווקא על מודל **DBSCAN** מכיוון שהוא בשלבים הקודמים פעל בצורה פחות טובה מ K means ורציתי לראות האם הוא משפר את המודל DBSCAN או לא.

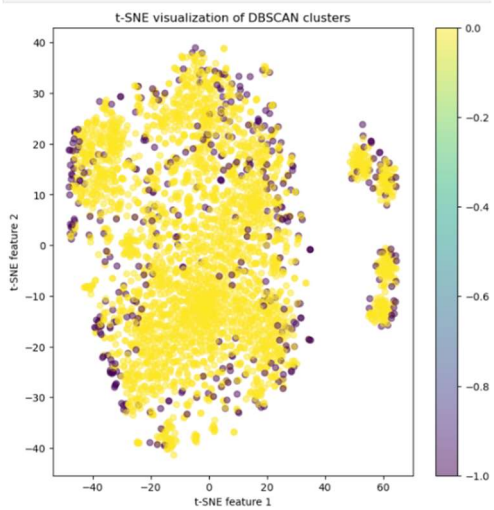
בחנתי שלושה ערכים של min_samples 2,3 ו21 .

כאשר ה eps היה שווה ל16:

```
For min_samples = 2, eps = 16: number of clusters: 3
Silhouette Score is 0.23322486934924136
For min_samples = 3, eps = 16: number of clusters: 1
For min_samples = 21, eps = 16: number of clusters: 1
```

וגם כאשר המשכתי עם הפרמטרים של ההתחלה (min_samples =3, eps = 16)

יצא t-SNE : שהכל מסווג כאשכול אחד ועוד כמה נק לא מוגדרות.



כדי להשוות למה שקרה לפני ההורדת ממדים. הורדתי את eps ל15 ואת min sample השארתי 3.

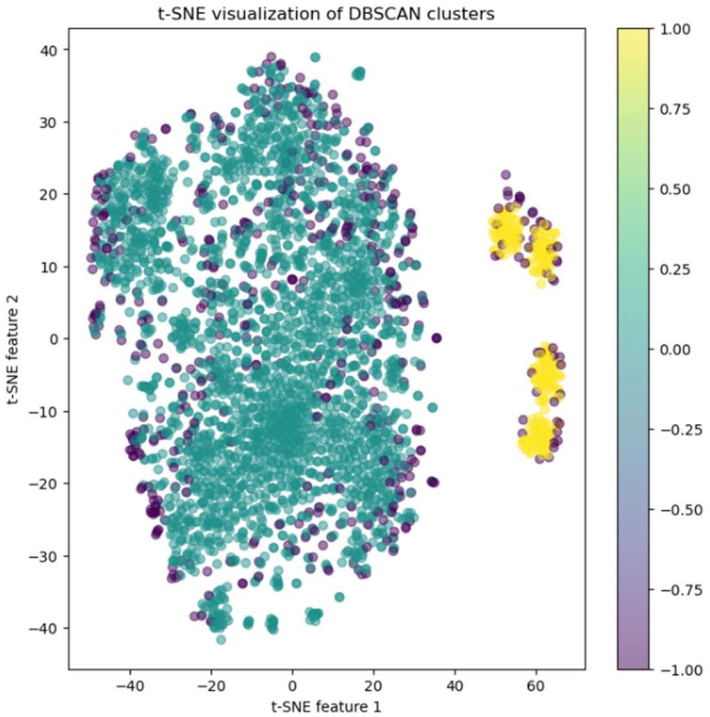
המספר הנמוך ביותר של סמפלים והמספר הגבוהה ביותר של סמפלים בכל אשכול הן:

```
DBSCAN: Largest Cluster: 0 (2678 samples), Smallest Cluster: 1 (277 samples)
```

ניתן לראות גם צמצום בכמעט חצי של הנקודות שסווגו כ 1-

dbscan_cluster	
-1	475
0	2678
1	277

לאחר השימוש בt-SNE:



אך עדיין זה לא מסווג לי לשלושה או לחמישה אשכולות.

לכן החלטתי לשפר את eps ולבחון בקוד מספר רב של eps ו min sample.

לאור המסכנה שנראה שבגרף t-SNE שישנם לפחות שלושה אשכולות החלטתי לצמצם כבר בקוד את התוצאות שמחזירות לי פחות משלושה אשכולות. בהתחלה הרצתי את eps על טווח 2-15 ואת min samples בטווח של 5-150 בקפיצות של 5, אך לאחר מיכן צמצמתי את הטווחים לטובת חיסכון בזמן והשארתי רק את הטווחים שהחזירו תוצאות בטבלה המודפסת.

הקוד עם הטווח המקוצר שהשארתי :

```
for eps in range(6,9):
    for min_samples in range(20,55,5):
        dbscan = DBSCAN(eps=eps, min_samples=min_samples)
```

להלן התוצאות:

	min_samples	eps	n_clusters	silhouette_avg	smallest_cluster_size	n_noise_samples
0	20	6	4	-0.074260	21	2480
1	25	6	3	-0.080626	26	2511
2	25	7	3	0.011946	58	2134
3	30	7	3	0.011132	58	2138
4	35	7	3	0.004488	35	2171
5	35	8	3	0.086439	70	1808
6	40	8	3	0.084429	60	1811
7	45	8	3	0.083736	65	1821
8	50	8	3	0.079468	51	1838

כאן אני בוחר להמשיך עם eps=8 ועם min sample=35 שבהן יש לי במספר המינימלי של האשכול הקטן את מספר הסמפלים הגבוהה ביותר (70). (שורה 5)

הפעלתי את אלגוריתם dbscan עם הפרמטרים הללו.

```
# הפעלת DBSCAN
dbscan = DBSCAN(min_samples =35, eps =8)
new_data_after_pca['dbscan_cluster'] = dbscan.fit_predict(new_data_after_pca)
```

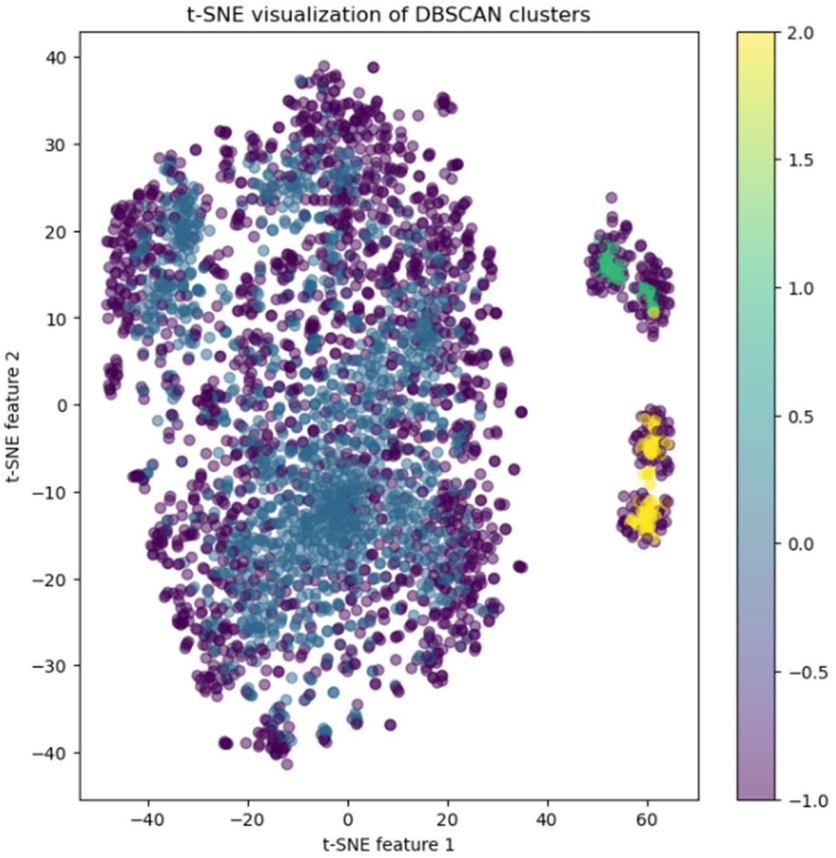
מספר הסמפלים הנמוך והגבוהה ביותר באשכול הם:

```
DBSCAN: Largest Cluster: 0 (1465 samples), Smallest Cluster: 1 (70 samples)
```

ניתן לראות שגם כאן מספר הלוא מסווגים (רעש) גדל משמעותית

```
dbscan_cluster
-1    1808
 0    1465
 1      70
 2      87
```

אך זה נתן לנו את הליבה של כל אשכול לפי גרף t-SNE :



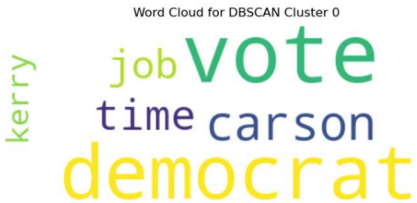
הסגולים הם הבלתי מסווגים.

ביצעתי ענן מילים לכל אשכול עם חמשת המילים הנפוצות ביותר.

```
Top words in DBSCAN Cluster 2:
november    1070.731318
democrat     165.455951
state        143.469594
percent      102.950463
dean          84.657255
dtype: float64
```


```
Top words in DBSCAN Cluster 1:
november    884.111330
vote         83.520217
report       82.957057
kerry        80.486171
job          60.935552
dtype: float64
```


```
Top words in DBSCAN Cluster 0:
vote        378.481421
democrat    370.604950
carson       97.354622
time         73.381365
job          65.080851
dtype: float64
```



באשכול 0, המילים הנפוצות הן "vote", "democrat", "carson", "time", "job". מתוך המילים הללו ניתן להסיק כי המאמרים באשכול זה עשויים לעסוק בנושאים של **פוליטיקה פנימית ופעילות מדינית**, כגון הצבעה, דיונים פוליטיים והשפעת מפלגות דמוקרטיות על המדינה. ניתן להניח שהמאמרים מתמקדים בדיווחים על פעילות מדינית ופוליטית, דיונים על תוכניות עבודה והשפעה פוליטית על החברה והכלכלה.

באשכול 1, המילים הנפוצות הן "november", "vote", "report", "kerry", "job". מתוך המילים הללו ניתן להסיק כי המאמרים באשכול זה עשויים להתמקד בנושאים **פוליטיים, בעיקר קשורים לאירועים נוכחיים ולדיונים בין מפלגות ומועמדים**. המאמרים עשויים לכלול דיווחים על בחירות, דיונים על הצבעות בקונגרס או בפני הציבור, ודיונים על תוכניות מדיניות פוליטיות.

באשכול 2, המילים הנפוצות הן "november", "democrat", "state", "percent", "dean". מתוך המילים הללו ניתן להסיק כי המאמרים באשכול זה עשויים להתמקד בנושאים כגון **תוצאות בחירות, נתונים סטטיסטיים על התפלגות הקולות, וניתוחים של מדיניות ציבורית**. ייתכן ויהיו דיווחים על תוצאות הבחירות, דיונים על השפעת התוצאות על מדיניות פנים וחוץ, והשוואות בין מפלגות ומועמדים.

תודה רבה!

קישור לפרוייקט בגיט