שם : סהר יעקב

סמינר : כריית נתונים

314741851

נושא : סיווג ביקורות משתמשים

הצגת נושא הסמינר

<u>הנושא עוסק בניתוח ביקורות של משתמשים וחלוקתם לקלאסטרים לפי ממוצע המרחק הוקטורי שלהם</u>

פילוח ביקורות מוצרים ודפוסי רכישה

•הקוד עוסק בקיבוץ ביקורות מוצרים לפי וקטורי מילים ובאמצעות עיבוד שפה טבעית.

קיבוץ ביקורות בעזרת אלגוריתם KMeans

•אלגוריתם KMeans משמש ליצירת קבוצות (אשכולות) של ביקורות בעלות מאפיינים דומים, על בסיס המרות טקסט למרחב וקטורי .

זיהוי דפוסים ונטיות קנייה

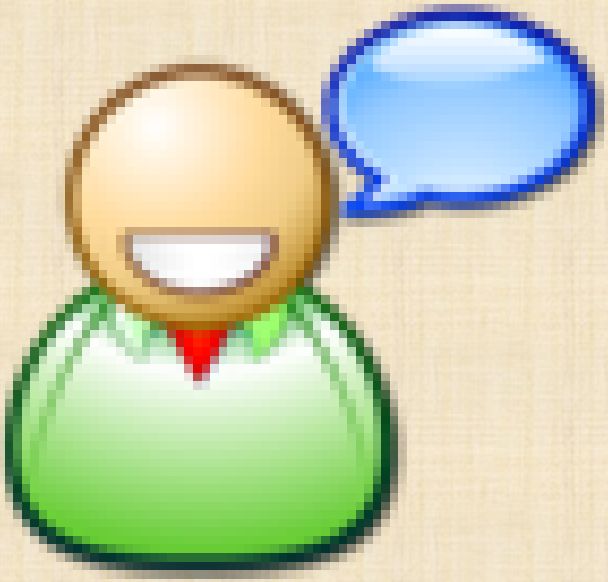
•המודל מזהה אילו מוצרים וביקורות משתייכים לאותה קבוצה, ומספק מידע על דפוסי רכישה חוזרים של לקוחות.

שיפור אסטרטגיות שיווק

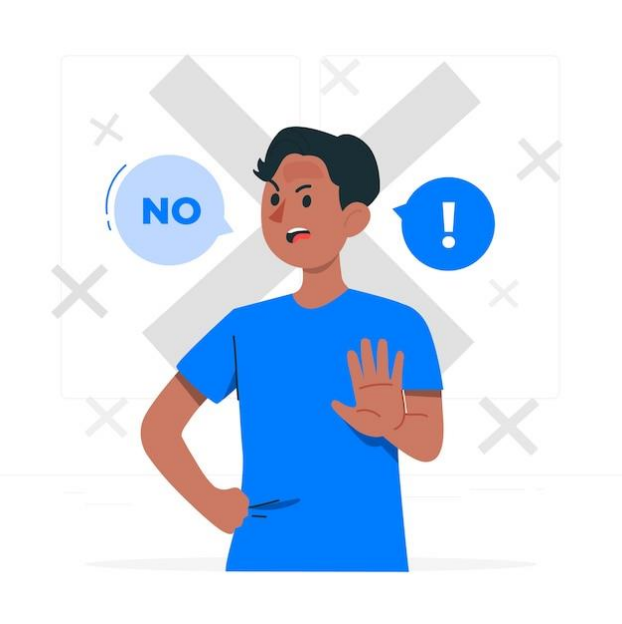
•בעזרת תוצאות הקיבוץ, ניתן להציע המלצות מותאמות אישית, לשפר מיקוד פרסומי ולבנות מבצעי שיווק ממוקדים.

מטרה: לייעל את חווית הקנייה וקיבוץ ביקורות

•זיהוי תבניות חבויות בביקורות עוזר להבין את צורכי הלקוחות, לתקן בעיות, ולשפר את חווית הקנייה הכוללת.



הבעיה העסקית



<u>הבעיה העיקרית היא כמות סוגי הביקורות של המשתמשים וניהולם</u>

- הקוד עוסק בהבנת דפוסי הביקורות של לקוחות על מוצרים, מתוך מטרה לזהות מגמות ודפוסים חוזרים.
- התובנות שמופקות משמשות להפקת הצעות מותאמות אישית, כמו שיפור מיקוד במוצרים מסוימים או הצעת המלצות ללקוחות.
- המערכת עוזרת לשפר את חווית הלקוח על ידי ניתוח והבנה של התגובות והצרכים שלהם, ובכך מאפשרת זיהוי והסתגלות לצרכים משתנים לאורך זמן.

פלט המערכת

הפלט של הקוד כולל:

- **קובץ CSV** שמאגד את הביקורות, ממוין לפי הקטגוריות (אשכולות) שהוגדרו על ידי אלגוריתם Kmeans.
- **אשכולות ביקורות**: חלוקה של הביקורות לקבוצות דומות על פי התוכן שלהן.
- **גרף Elbow Method**: מספק חיווי על מספר האשכולות האופטימלי לניתוח.
- **תובנות על דפוסי ביקורות**: זיהוי קשרים ותבניות חוזרות בין ביקורות שיכולים לשמש לשיפור חוויית הלקוח והתאמת הצעות שיווקיות.
- **גרפים נוספים**

```
convert try : sen = i love coding with python
[[0.57735027 0.57735027 0.57735027]]
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
```

```
shape : (2220, 2)
                 reviews.title
reviews.rating
1.0                       1132
2.0                        342
3.0                        746
4.0                       2195
5.0                       6136
```

	reviews.rating	reviews.title
51	3.0	Worth the money. Not perfect, but very very go...
52	3.0	Worth the money. Not perfect, but very very go...
97	3.0	Not all fabric or cloth.
106	3.0	KF HDX 8.9 is ok do homework on Prime download...
108	1.0	Dead after 15 months

	0	1	2	3	4	5	6	7	8	9	..
sentences	Worth the money. Not perfect, but very very go...	Worth the money. Not perfect, but very very go...	Not all fabric or cloth.	KF HDX 8.9 is ok do homework on Prime download...	Dead after 15 months	Not everything I expected...	KF HDX 8.9 is ok do homework on Prime download...	KF HDX 8.9 is ok do homework on Prime download...	KF HDX 8.9 is ok do homework on Prime download...	KF HDX 8.9 is ok do homework on Prime download...	.
mean vector	0.226805	0.226805	0.57735	0.377964	0.57735	0.57735	0.251416	0.251416	0.251416	0.251416	.

למה ממוצע?

1. **ייצוג מרכזי**: ממוצע מייצג ביעילות את המיקום הכללי של האשכול.

2. **עמידות לרעש**: מפחית השפעת רעשים קלים באשכולות גדולים.

3. **תואם את האלגוריתם**: ממזער את השגיאה.

4. **מתאים למרחבים מרובי-מימדים**: יעיל לניתוח נתונים מורכבים כמו וקטורי טקסט.

חיסרון בחציון: החציון מתעלם מהמרחקים של הנתונים סביבו ומייצג רק את הערך האמצעי, מה שעלול לגרום למרכז אשכול לא מדויק במבנים גאומטריים מורכבים.
חיסרון בשכיח: כאשר עובדים עם וקטורים, השכיח לא מסוגל להתמודד עם מרחבים רב-ממדיים בצורה אינטואיטיבית או חישובית יעילה.

$$MEAN = \frac{\sum x}{N}$$

Silhouette Score: מודד את התאמת הנקודות לאשכולות שלהן. ערך גבוה (0.85) מציין קיבוץ טוב.

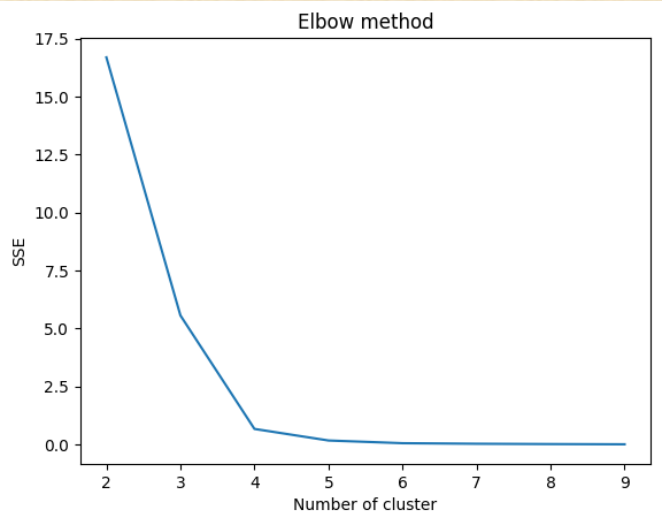
1- מציין קיבוץ טוב
0- מציין קיבוץ מוזר או מוטעה
-1- מציין קיבוץ לא טוב

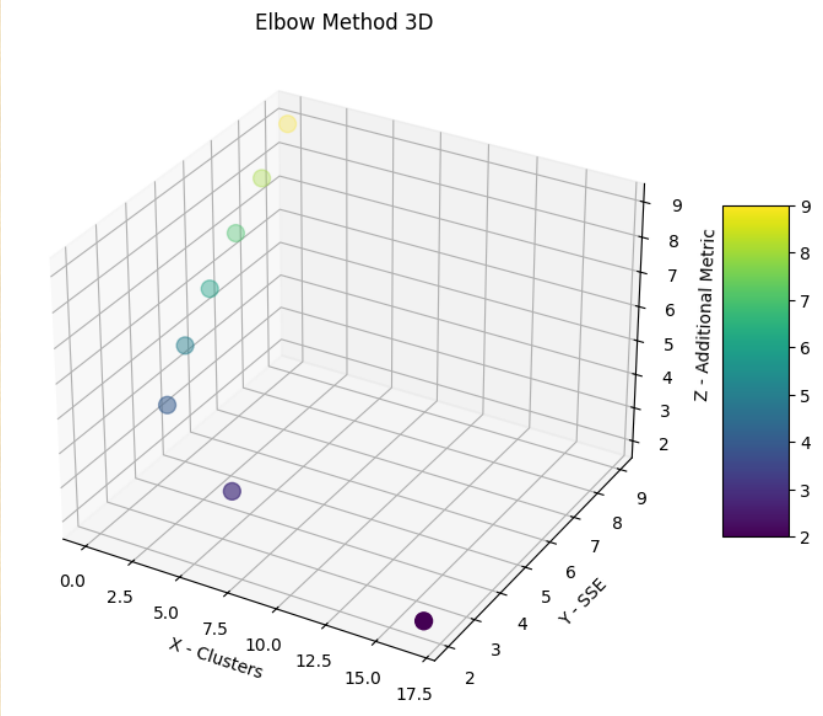
$$\frac{b(i) - a(i)}{\max(a(i), b(i))} = S(i)$$

מתבצע ע"י חישוב המרחק בין כל נקודה וגם נקודת המרכז .
בעזרת הפונקציה מתמטית הוא מעריך את המדד של כל נקודה.
ככל שיש יותר קלאסטרים המדד יותר טוב והוא יעלה (בדומה לsse, כאן המדד יותר טוב כי הוא קטן)

```
Silhouette Score: 0.9805

Final Data with Clusters:
                                                sentences mean vector  cluster
0  Worth the money. Not perfect, but very very go...    0.226805        4
1  Worth the money. Not perfect, but very very go...    0.226805        4
2  Worth the money. Not perfect, but very very go...    0.226805        4
3  Worth the money. Not perfect, but very very go...    0.226805        4
4                           Not all fabric or cloth.     0.57735        3
cluster 5 done successfully
```


Elbow method
SSE
Number of cluster
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
2
3
4
5
6
7
8
9


Elbow Method 3D
X - Clusters
Y - SSE
Z - Additional Metric
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
2
3
4
5
6
7
8
9

מדדים נוספים:

```
Number of clusters: 6
        SSE  Silhouette Score  Davies-Bouldin Index  Calinski-Harabasz Index
0  0.056814          0.985174              0.284697            566900.002124
col name of 6 clusters :
Index(['sentences', 'mean vector', 'cluster'], dtype='object')
```

SSE (Sum of Squared Errors):
מודד את המרחקים בין הדגימות למרכזי הקיבוץ. ערך נמוך מציין קיבוצים צפופים.

Silhouette Score:
מדד זה מודד את איכות הקיבוץ בהשוואה לדגימות שנמצאות **בקיבוצים שונים**. ערך גבוה (קרוב ל-1) מציין קיבוצים ברורים ומופרדים היטב.

Davies-Bouldin Index:
מדד זה מודד את היחס בין **היקפי הקיבוצים למרחקים בין מרכזי הקיבוץ**. ככל שהמדד נמוך יותר, כך איכות הקיבוץ גבוהה יותר. לפי צפיפות , ערך נמוך מציין קיבוצים מופרדים היטב.

Calinski-Harabasz Index:
מדד זה מודד את היחס בין המרחק בין מרכזי הקיבוצים לבין הצפיפות **בתוך כל קבוצה**. ערך גבוה מציין קיבוצים איכותיים וברורים.

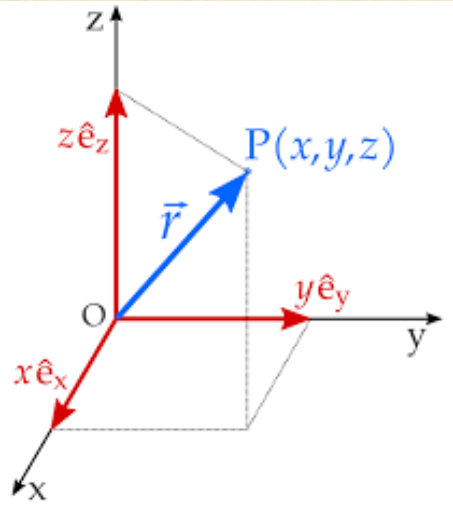
מסקנה - **Davies-Bouldin Index** (0.2) מציין הפרדה ברורה בין הקבוצות
Silhouette Score(0.97) מעיד על קיבוצים מופרדים היטב עם דגימות קרובות בתוך כל קבוצה
ו- **Calinski-Harabasz Index** (כ-560,000) מציין קיבוצים צפופים ומופרדים באופן איכותי.

נתונים ודרכי אסיפתם

- איסוף נתונים ממאגרי נתונים קיימים של קגל
- פורמט CSV לקריאה קלה בסביבת פייתון
- עיבוד נתונים בסביבת פייתון
- שימוש ב-Google Colab
- הכנת נתונים לניתוח וייצוגם כוקטור
- איכות הנתונים האמיתיים
- חשיבות הנתונים הנכונים

כלים לניתוח סטטיסטי

- הקוד עושה שימוש בספריית **TfidfVectorizer** להמרת מילים לוקטורים לצורך ניתוח טקסט.
- באמצעות אלגוריתם **KMeans**, הוא מקבץ את הביקורות לאשכולות דומים.
- הנתונים מוצגים בצורה גרפית, כמו בגרף ה- Elbow Method, לזיהוי מספר האשכולות האופטימלי.
- באמצעות נאמפיי, הקוד ממיר את גרף המרפק לפונקציה פולינומית ממעלה 3 (בגלל שיש שני נקודות פיתול) אז הערך k שבניהם הוא האופטימלי.
- תובנות אלו מסייעות בשיפור תהליכי קבלת החלטות ובהבנת הקשרים בין הביקורות על סמך הוקטורים המייצגים את המילים.

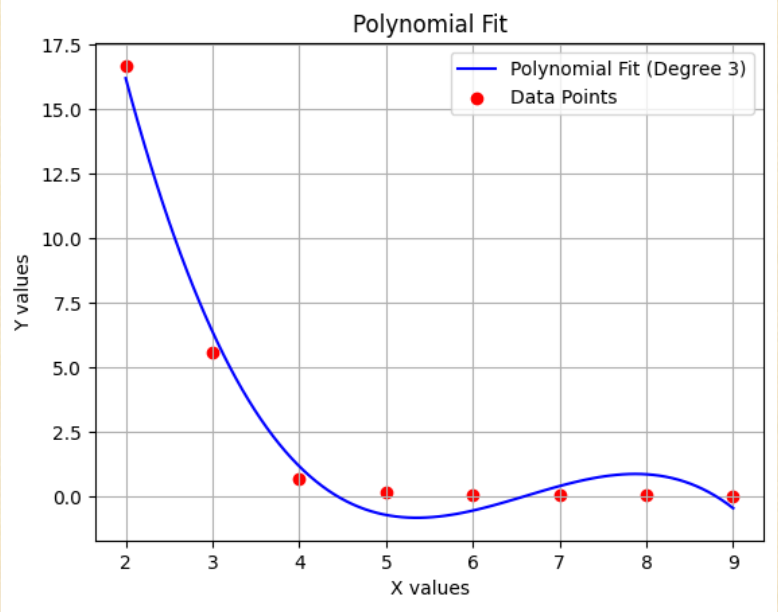


```
coeff = np.polyfit(list(sse.keys()), list(sse.values()), 3)  # דרגת הפולינום 3
p = np.poly1d(coeff)
print(f'f(x) = \n\n')
print(f'{p}')
```

```
f(x) = -0.2126 x^3 + 4.217 x^2 - 26.87 x + 54.78
```

```
[54] x = sympy.Symbol('x')
     fun = ''
     for i,conf in enumerate(p):
       fun += str(conf)+ f'*x**{poly_deg - i}+'
     fun = fun[:-1]
     sympy_fun = sympy.sympify(fun)
     diff_x = sympy.diff(sympy_fun, x)
     proof = sympy.solve(diff_x, x)
     print(proof)
```

```
[5.35286627111674, 7.87182857528473]
```



```python
for key,val in sse.items():
  if proof[0]<key<proof[1]:
    print(f'best cluster is {key} , sse : {val}\n\n')
    evaluate_clustering(meanvec_reshape, n_clusters=key)
    break
```

```
best cluster is 6 , sse : 0.056814254524235136


Number of clusters: 6
        SSE  Silhouette Score  Davies-Bouldin Index  Calinski-Harabasz Index
0  0.056814          0.985174              0.284697            566900.002124
```

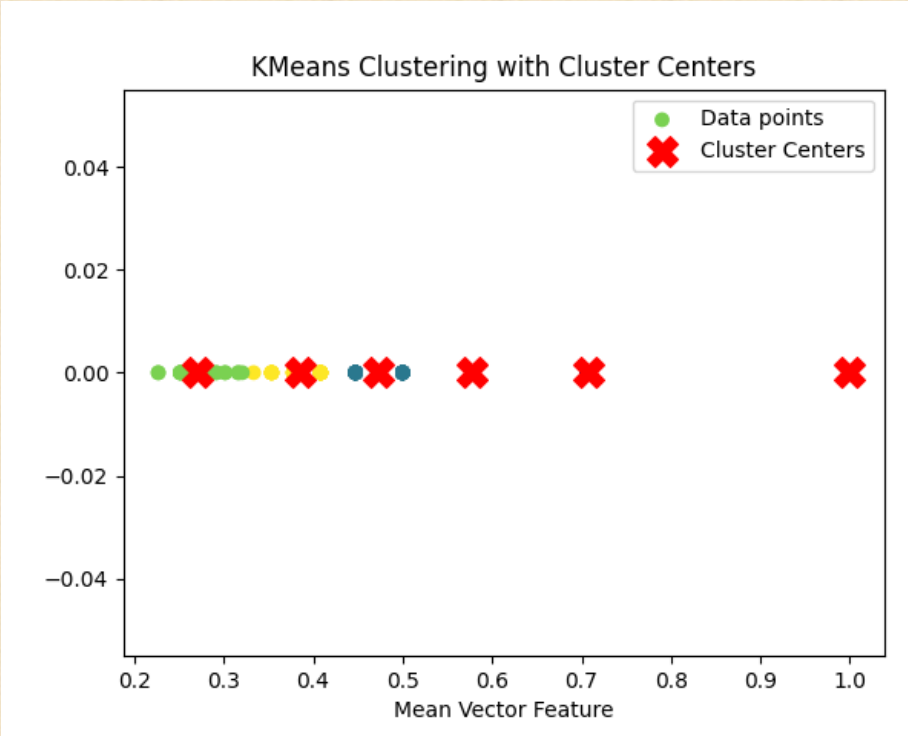
האלגוריתם KMeans

- קיבוץ ביקורות דומות לפי אותו ממוצע וקטור

הקוד מקבץ ביקורות דומות באמצעות חישוב ממוצע וקטור המילים שלהן, מה שמאפשר לזהות קטגוריות משותפות בין הביקורות.

- זיהוי קטגוריות משותפות של ביקורות

תהליך זה מסייע להבין אילו ביקורות חולקות נושאים דומים ולמיין אותן באופן אופטימלי לאשכולות רלוונטיים.



יצירת עץ החלטה וחיזוי

decision tree information

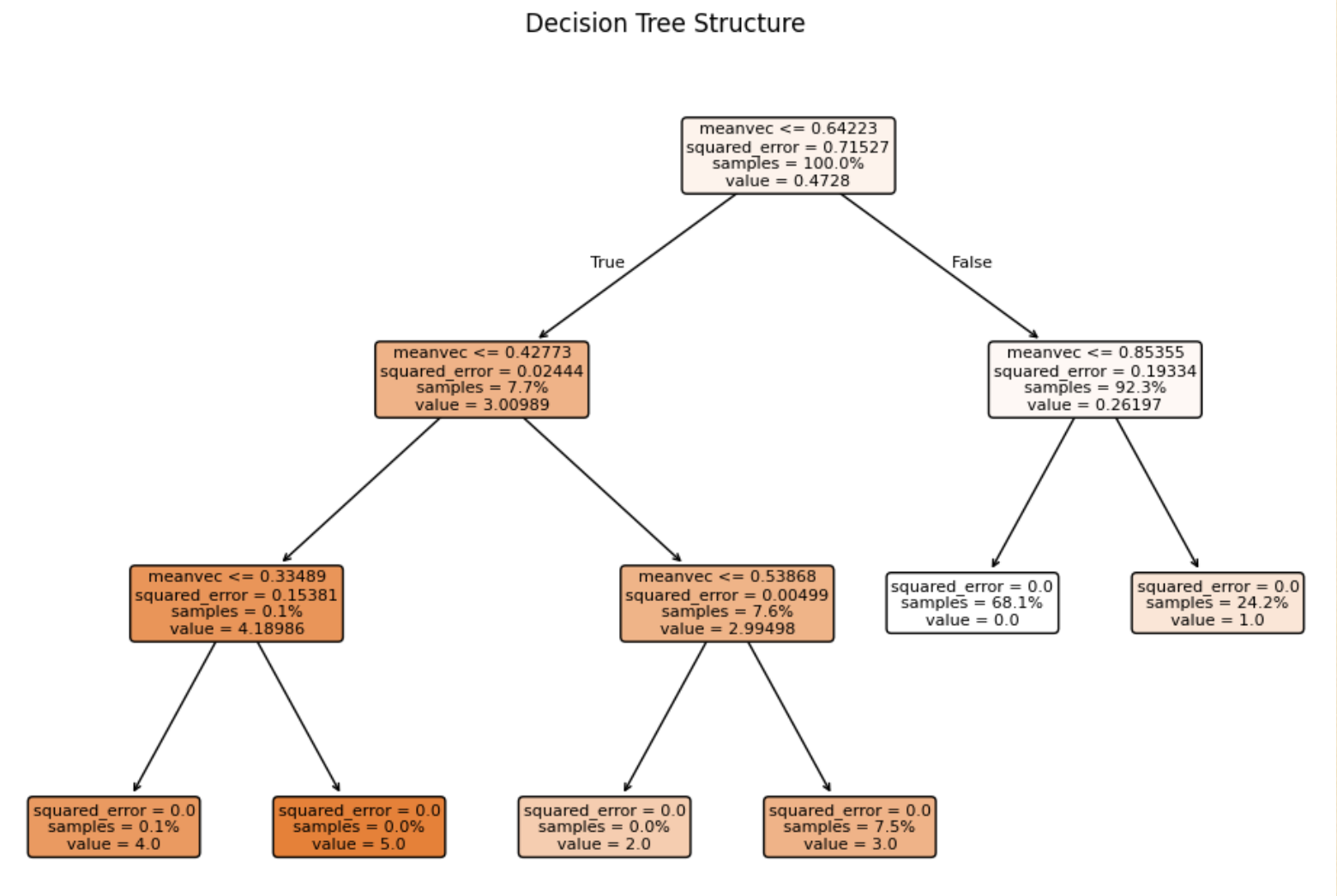
Decision tree MSE: 1.9190211457e-06
Decision tree SSE: 0.999999999999999999
R-squared (R^2): 9.9999731991e-01

Sentence - It's not worth the price at all. It didn't meet my expectations.

The vector of sentences: [[0.33333333 0.33333333 0.66666667 0.33333333 0.33333333 0.33333333]]

The reshaped vector: [[0.38888889]]

Prediction: 5.0


Decision Tree Structure
meanvec <= 0.64223
squared_error = 0.71527
samples = 100.0%
value = 0.4728
True
False
meanvec <= 0.42773
squared_error = 0.02444
samples = 7.7%
value = 3.00989
meanvec <= 0.85355
squared_error = 0.19334
samples = 92.3%
value = 0.26197
meanvec <= 0.33489
squared_error = 0.15381
samples = 0.1%
value = 4.18986
meanvec <= 0.53868
squared_error = 0.00499
samples = 7.6%
value = 2.99498
squared_error = 0.0
samples = 68.1%
value = 0.0
squared_error = 0.0
samples = 24.2%
value = 1.0
squared_error = 0.0
samples = 0.1%
value = 4.0
squared_error = 0.0
samples = 0.0%
value = 5.0
squared_error = 0.0
samples = 0.0%
value = 2.0
squared_error = 0.0
samples = 7.5%
value = 3.0

התאמות שבוצעו

- הנתונים טופלו על ידי הסרת ערכים חסרים ומילים לא חשובות, כדי להבטיח תוצאה אחידה.
- הנתונים הוסבו לווקטורים ויוצאו לקובץ אקסל.
- שיפור איכות הנתונים כלל ניקוי תווים מיותרים ושיפור הפורמט.
- כל תהליך העיבוד תועד, תוך שמירה על התאמות קריטיות לניתוח עתידי.

אתגר ראשון

- בחירת מספר הקבוצות האופטימלי היא קריטית בתהליך הקיבוץ.
- השוואת תוצאות בין ערכים שונים של K ולבחון את השפעתו על איכות הקיבוצים.
- האתגר הוא למצוא את K המתאים, שכן מספר קטן או גדול מדי עלול להוביל לתוצאות לא מדויקות.
- ניתוח מעמיק של SSEחשוב כדי לבחור את ה- K האופטימלי.

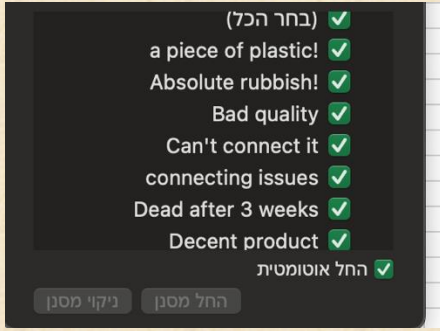
```python
for key,val in sse.items():
  if proof[0]<key<proof[1]:
    print(f'best cluster is {key} , sse : {val}\n\n')
    evaluate_clustering(meanvec_reshape, n_clusters=key)
    break
```

```
best cluster is 6 , sse : 0.056814254524235136


Number of clusters: 6
        SSE  Silhouette Score  Davies-Bouldin Index  Calinski-Harabasz Index
0  0.056814          0.985174              0.284697            566900.002124
```

Worth the money. Not perfect, but very very good for start to finish novels in good light 11,965 people found this helpful. Was this review helpful to you Yes No	0.22680460581325723	4
Worth the money. Not perfect, but very very good for start to finish novels in good light 11,965 people found this helpful. Was this review helpful to you Yes No	0.22680460581325723	4
Not all fabric or cloth.	0.5773502691896258	3
Not all fabric or cloth.	0.5773502691896258	3
Not all fabric or cloth.	0.5773502691896258	3

3	0.57735027	Maybe I was	3004
3	0.57735027	Maybe I was	3005
3	0.57735027	Maybe I was	3006
3	0.57735027	Maybe I was	3007
1	1	Disappointed	4096
1	1	Disappointed	4097



0
0
0
0
0
0
0
0
0
0
0
0
0

אתגר שני

- עיבוד נתונים במהלך חישוב יכול להתרחש כאשר המידע לא מומר כראוי או אם קיימת אי התאמה בין הנתונים השונים.
- כדי להתמודד עם בעיה זו, ניתן להשתמש בפתרונות מבוססי פונקציות מתקדמות, כמו טור טיילור או קירוב פונקציות פולינומית שמאפשרים חישוב מדויק יותר של דמיון בין נתונים. טור טיילור מספק קירוב פונקציונלי למודלים מתמטיים, ומאפשר לעבד נתונים בצורה מדויקת יותר, תוך צמצום איבוד המידע בתהליך החישוב , בנוסף בחנתי תהליך ממוצע וקטור על משפט דוגמא.

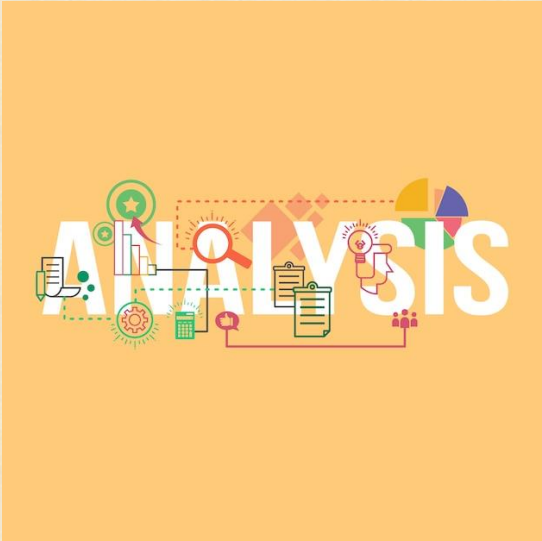
```python
sen = "I love coding with Python because it allows me to quickly build powerful."
print(f'convert try : sen = {sen}')
trySen=create_vector_from_list([clear_stopWord(sen.split())])
print(trySen)
```

```
convert try : sen = I love coding with Python because it allows me to quickly build powerful.
[[0.37796447 0.37796447 0.37796447 0.37796447 0.37796447 0.37796447
  0.37796447]]
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

```python
[74] print(np.mean(trySen))
```

```
0.37796447300922725
```

אתגר שלישי



- זיהוי נכון של המדד הסטטיסטי הוא שלב חשוב בתהליך עיבוד נתונים, שכן הוא מאפשר למדוד את הדמיון או הקשר בין נתונים בצורה מדויקת.
- חישוב המדד על הווקטור המומר של המילה חשוב להבטיח שההשוואות יהיו נכונות ויעילות.
- כשממירים מילים לווקטורים, חישוב המדד הסטטיסטי על הווקטור המומר מאפשר לבצע חילוק לקבוצות (קלאסטרים) בצורה הטובה ביותר, כך שניתן לזהות דפוסים וקטגוריות עם דיוק גבוה.

mean std variance
median mode

מסקנות עיקריות

1. **קיבוץ דעות דומות**: קיבוץ ביקורות מאפשר לאגד דעות דומות ולזהות קבוצות עם תחושות משותפות לגבי המוצר או השירות.
2. **גילוי תבניות וטרנדים**: חילוק לקבוצות מאפשר חשיפת תבניות בעדויות, כמו ביקורות חיוביות מול שליליות, שיכולות להצביע על בעיות או יתרונות במוצר.
3. **שיפור תהליך קבלת ההחלטות**: קיבוץ ביקורות מסייע למנהלים ומפתחים להתמקד בתחומים הדורשים שיפור או תיקון, מה שעוזר בקבלת החלטות ממוקדות.
4. **איתור בעיות או הזדמנויות**: החילוק לקבוצות עוזר לזהות בעיות משותפות או תחומים שדורשים שיפור, ובכך מקדם הזדמנויות לפיתוח ושדרוג.
5. **חילוק לקבוצות בהתבסס על מאפיינים שונים**: ניתן לבצע חלוקה לקבוצות על פי קריטריונים מגוונים, כמו דעות חיוביות או שליליות, ולהתאים את אסטרטגיות השיווק לפי הקבוצות השונות.

הקוד המלא

```
from sklearn.cluster import KMeans
def kMean_model(n_class):
  n_class = n_class
  rand_state = 42

  kMean = KMeans(n_clusters=n_class, random_state=rand_state) # model
  my_k_mean = kMean.fit(meanvec_reshape) # train kMean
  sse = kMean.inertia_ # sum of squared errors

  dict_cluster = {}
  dict_cluster['mean vector'] = meanvec
  dict_cluster['cluster'] = my_k_mean.labels_
  dict_cluster_df = pd.DataFrame(dict_cluster)
  dict_cluster_df.head()

  final_dataClustering_whith_sen = pd.merge(pro_df_with_sentences.transpose() , dict_cluster_df , on='mean vector')
  copy_f = final_dataClustering_whith_sen.copy()
  final_dataClustering_whith_sen # organize reviews sentences with meanVec and clusters
#print results

  return (copy_f , sse)
```

הקוד המלא

```
def load_data(path):
  data = pandas.read_csv(path)
  data = pd.DataFrame(data)
  return data
#change to vector
def create_vector_from_list(col_values):
  vectorizer = TfidfVectorizer()
  X = vectorizer.fit_transform(col_values)
  X = X.toarray()
  return X
#delete unimprotant words
nltk.download('stopwords')
def clear_stopWord(col_val):
  stop_words = set(stopwords.words('english'))
  filtered_text = [word for word in col_val if word not in stop_words] # מוחק את המילים אם לא חשובות
  txt = " ".join(filtered_text)
  return txt
```

תודה על ההקשבה

- סיום המצגת
- הודיה לקהל (התחוויה)
- הזמנה לדון על התוצאות
- שיתוף פעולה עתידי (השקעה)
- הבנת חשיבות הניתוח