מטלה בלמידת מכונה מתקדם- סיווג

***עבור K-means:*** *חלקו את הדאטה ל2,3,7,8 אשכולות והציגו את איכות החלוקה על פי הקריטריונים שנלמדו בהרצאה בכיתה.*

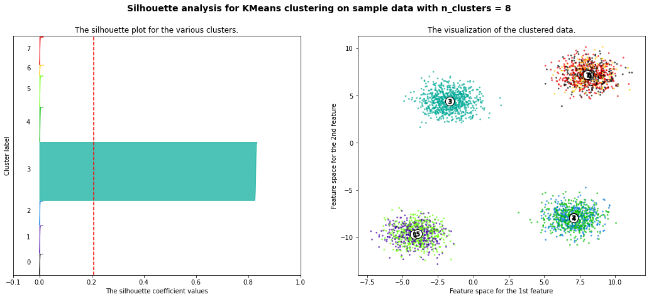
**Silhouette**

איכות החלוקה לא מספיק טובה. ניתן לראות מהגרף שאף אחד מהK לא מדוייק מספיק. עבור K=2 הגרף עומד על הממוצע אבל בעין אנושית אפשר לראות שהחלוקה לא טובה. עבור ל=3 חלק מהסיווגים הם טובים אבל גם חלק קטנים מהממוצע. עבור K=7 ו K=8 יש מסווג אחד שמעל ההמוצע ובאמת "נאמן" לחלוקה אבל כל השאר הרבה מאחורי הממוצע ורואים שיש פה אובר.

A screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generated

**SSE**

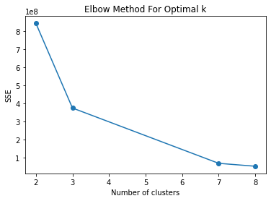
איכות החלוקה לא מספיק טובה.

For n\_clusters = 2 The SSE is: 841594724.019801

For n\_clusters = 3 The SSE is: 374537786.9350555

For n\_clusters = 7 The SSE is: 69520471.71599178

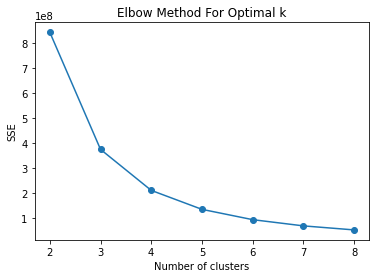
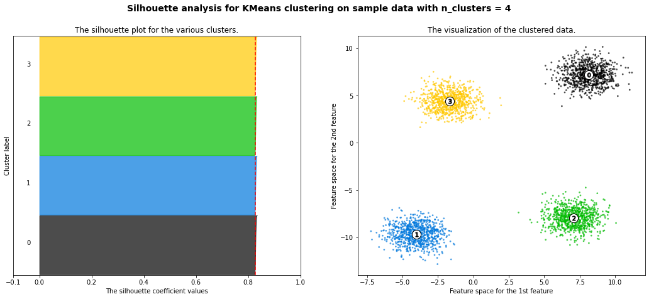
For n\_clusters = 8 The SSE is: 53431441.432874575

עפ"י הנתונים אפשר לראות שאיפשהו בין 3 ל7 יש נקודה מינימלית שבה נרצה להשתמש. 

*בחרו את מספר האשכולות האופטימלי מבחינתכם, הסבירו למה בחרתם ערך זה*

עפ"י Silhouette בראייה ויזואלית ניתן להסיק שהעולם מתחלק ל4 קבוצות, גם המדד מראה שהגרפים מגיעים ממש עד לקו ההמוצע.

עפ"י SSE נראה שK האופטימלי הינו K=5 משום שערך זה מקבל את הSSE הנמוך ביותר. אולם בהשוואה לK=4 במדד הזה אין פער דרמטי(ע"י הנתונים וע"י הגרף) ולכן לפי דעתי כדאי לסווג את הדאטה הנל ל4 קבוצות.

****

For n\_clusters = 2 The SSE is: 841594724.019801

For n\_clusters = 3 The SSE is: 374537786.9350555

For n\_clusters = 4 The SSE is: 211063799.93845418

For n\_clusters = 5 The SSE is: 135400229.71346766

For n\_clusters = 6 The SSE is: 94298782.97384137

For n\_clusters = 7 The SSE is: 69520471.71599178

For n\_clusters = 8 The SSE is: 53431441.432874575

*לאיזה אשכול יש את המספר המקסימלי של דוגמאות ולאיזה יש את המספר המינימלי*

For n\_clusters = 4

Cluster 1 has the minimum number of examples: 856

Cluster 2 has the maximum number of examples: 859

*צרו קובץ CSV המכיל את הנתונים השייכים לכל אשכול בנפרד.  
הסתכלו על הקבצים השונים ונסו להבין לאיזה נושא כל אחד מתייחס, תעדו בתשובה מה הערכים שגרמו לכם להגיע למסקנה*

מערך המילים כל כך גדול שלא הצלחתי להבין איך הסיווג נעשה לצערי.

***עבור המודל ההיררכי:***

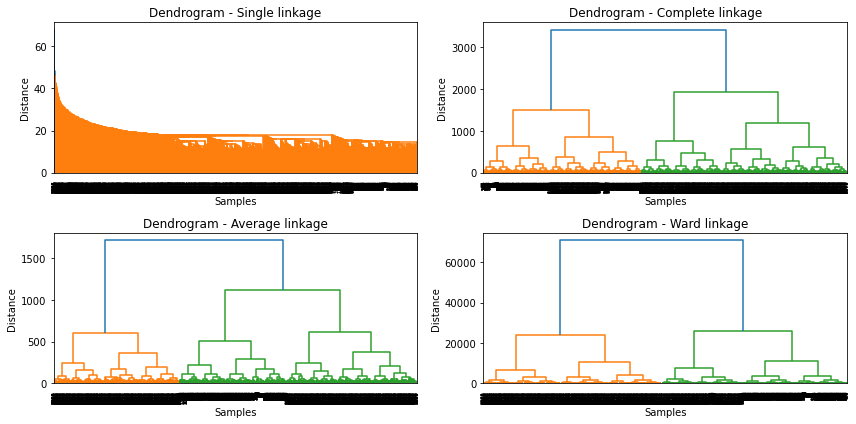
*בצעו סיווג על פי ארבעת ערכי הlinkage בספרייה. השוו את הדנדוגרמים שאתם מקבלים*

עבור ה WORD נקבל דנדוגרם מאוד סימטרי (תכונה שיחסית טבעי לו כי הוא רוצה למזער את המרחק).

עןבר הCOMPLETE נקבל דנדוגרם שבו קבוצה אחת מאוד דומה לשניה אבל פרופורציית הגדלים לא זהה.

עבור הSINGLE נקבל גרף מאוד לא ברור, כנראה נובע מהרבה נתונים שיחסית זהים.

עבור הAVERAGE נקבל גרף לא סימטרי. משהו מגוון ואחר.

**