מעבדה 5 - הנחיות הגשה

עליכם להגיש סיכום של המעבדה **במסמך PDF או HTML אחד** כתוצר של מחברת ג'ופיטר - ניתן יהיה להגיש רק קובץ אחד.
המחברת תכלול קוד, טבלאות רלוונטיות, ויזואליזציות, את הסיכומים והתשובות ניתן לכתוב בכתב יד ולהוסיף כתמונות למסמך (בתנאי וקריא) או להקלידן. יש לתעד את הקוד באופן ברור עבור כל פעולה.
יש לתת כותרת ברורה וקריאה עבור כל גרף וטבלה (ולא כותרת כללית לכלל הגרפים מאותו סוג)
מקרא עבור הגרפים (או משפט הסבר)
במידה והתבקשתם להציג ערכים, בבקשה לכתוב במשפט את המסקנה מהגרף.
מומלץ לבנות את המחברת בצורה כללית כך שתהיה נכונה לכל אוסף מסמכים (ותהפוך להיות כלי עבודה עבור כל ניתוח טקסטואלי )

כנתונים עליכם להשתמש במסמכים מסווגים (מתויגים)

```
from sklearn.datasets import fetch_20newsgroups
```

עליכם לבחור 4 קטגוריות מתוך אוסף זה

```
categories = ['talk.politics.misc', 'talk.religion.misc', 'comp.graphics', 'sci.space']
```

1. בעבור SKLEARN , קיראו על אלגוריתמי הקיבוץ השונים. בדגש על הנלמד בהרצאה (**k-means ,hac**) שימו לב לתכונות, מטריקות הקירבה, אופי הפעולה, וההפעלה ברמה הטכנית של כל אחד מהגרסאות לכלי הנל.

```
from sklearn.cluster import AgglomerativeClustering
from sklearn.cluster import KMeans
```

2. חקר אלגוריתמי הקיבוץ :
   1. בחלק זה יהיה עליכם להשתמש באלגוריתמים לקיבוץ מ SKLEARN
   2. יש לטפל במסמכים בהיבט העיבוד המקדים שימו לב ניתן להיעזר בכלים שפותחו בעבר, או להיעזר בספריות פייתון המיועדות לכך. (מצופה ייצוג TF-IDF )
   **יש להציג את הגדלים השונים של מבני הנתונים שנוצרו לאחר תהליכי העיבוד הנבחרים.**
   3. יש לבנות מודלים לקיבוץ המסמכים בעזרת כל אלגוריתם.
      i. עליכם להריץ מספר ניסויים אמפיריים בין האלגוריתמים (כמה שניתן להשוות בינהם)
         1. מספר האשכולות ( 3 אפשרויות שונות **ומנומקות**)
         2. מטריקת המרחק/דימיון (3 אפשרויות **-** קוסינוס, אוקלידי ומנהטן)
         3. שיטות קישור (link) שונות עבור האלגוריתם **ההיררכי** (2 אפשרויות)

      ii. לא לשכוח חלוקה (30/70) ל train ו test
      iii. יש להציג זמני ריצה עבור כל תהליך
      iv. יש להציג את מבני הנתונים המייצגים את המודל
      v. יש להציג ויזואליזציות רלוונטיות לקיבוץ.
      רפרנס:

vi. scipy.cluster.hierarchy.dendrogram

vii.
```
pca = PCA(n_components=2, random_state=21)
reduced_features = pca.fit_transform(X.toarray())
# reduce the cluster centre to 2D
reduced_cluster_centers = pca.transform(model.cluster_centers_)
plt.scatter(reduced_features[:,0], reduced_features[:,1],
c=model.predict(X))
plt.scatter(reduced_cluster_centers[:, 0], reduced_cluster_centers[:,1],
marker='x', s=150, c='b')
```

3. יש להציג מדדי הערכה לכל מודל (על פי הנלמד בהרצאה RAND and Purity) ישלהשתמש בספריות פייתון המיועדות לכך (כמו sklearn.metrics ) שימו לב למטריקות הערכה לאישכול .

https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics

4. סכמו את התוצאות באופן השוואתי וברור והסבירו במילים שלכם (יש להתייחס לנקודות הבאות): את התוצאות יש להציג גם בטבלה השוואתית בנוסף ויזואליזציות מתאימות להערכת מודל וניתוח התוצאות .
   i. גודל מבני נתונים לפני הקיבוץ (גודל מטריצת הקלט)
   ii. זמני בניית מודל
   iii. ניתוח של תוצאות ההערכה של המודלים השונים.
   iv. מסקנות

5. חלק ג' - סיכום
   1. עליכם לכתוב סיכום קצר במילים שלכם (4-5 פסקאות), מתוך התייחסות לידע האישי שלכם בתחום מדעי הנתונים, והחומר הנלמד עד היום בתואר.
   2. מה למדתם ממעבדה זו - יש להציג טבלאות השוואתיות וגרפים (גדלים ומדידות זמנים).
   3. יש לדון בתוצאות הניסויים השונים שלכם.
   4. יש לסכם את המסקנות הנובעות מההשוואות השונות.
   5. עליכם להסביר מעבדה זו בהתייחס לחומר התיאורטי שנלמד עד היום בתחום מדעי הנתונים.

בהצלחה !