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)
ישנן שתי אפשרויות, הראשונה היא המומלצת.

A. from sklearn.datasets import fetch_20newsgroups

B. אוסף הטוויטים שניתן לכם בתחילת הקורס מכיל תיוג לטוויטים שנחשבים דברי נאצה (תוכן בוטה, גזעני או סקסיסטי)

1. הסבירו וסכמו על קיבוץ/אישכול - clustering במילים שלכם, ובהתייחס לנלמד בהרצאה ובמסלול , באופן כללי, וספציפית לקיבוץ טקסט.
   יש להתייחס לסוגי האלגוריתמים השונים שנלמדו בהרצאה, מתי לדעתכם נשתמש בכל אחד, מטריקות קירבה, יתרונות, חסרונות, סקלביליות זמני ריצה וכו'. **(לכל הפחות יש להתייחס ל k-means , hac).** עבור האלגוריתם ההיררכי יש להסביר מהן צורות ה"קישור" LINK השונות
2. בעבור חבילות הפייתון הבאות, קיראו על אלגוריתמי הקיבוץ השונים. בדגש על הנלמד בהרצאה (ובהתייחס לאלגוריתמים בסעיף קודם) סכמו על התכונות, מטריקות הקירבה, אופי הפעולה, וההפעלה ברמה הטכנית של כל אחד מהגרסאות לכלי הנל.
   פרטו מהם הכלים שמקבילים לאלו שנלמדו בהרצאה. שימו לב בספריות השונות יש יותר מ -2 גרסאות שונות, וגם לגרסאות מקבילות יש קלט שונה והגדרות שונות.
   1. sklearn
   2. NLTK
3. הסבירו וסכמו במילים שלכם, בהתאם לנלמד בהרצאה, איך מבצעים הערכת מודל בקיבוץ טקסט, בעזרת אילו מדדים נשתמש.
4. איך ניתן לבצע וויזואליזציה לכל אחד מאלגוריתמי הקיבוץ ?

5. חקר אלגוריתמי הקיבוץ :
   1. בחלק זה יהיה עליכם להשתמש באלגוריתמים לקיבוץ מ 2 ספריות מוכנות - , k-means hac . סה"כ 4 כלים.
   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 **מנומקות** כמידת האפשר, לדוג' קוסינוס, אוקלידי ומנהטן)

3. שיטות קישור (link) שונות עבור האלגוריתם ההיררכי (2 אפשרויות)

ii. לא לשכוח חלוקה ל train ו test
iii. יש להציג זמני ריצה עבור כל תהליך
iv. יש להציג את מבני הנתונים המייצגים את המודל
v. יש להציג וויזואליזציות רלוונטיות לקיבוץ.
רמז ורפרנס:

vi. scipy.cluster.hierarchy.dendrogram

vii.
```
pca = PCA(n_components=2, random_state=21)
reduced_features = pca.fit_transform(X.toarray())
# reduce the cluster centers 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')
```

6. יש להציג מדדי הערכה לכל מודל (על פי הנלמד בהרצאה) ניתן להשתמש בספריות פייתון המיועדות לכך (כמו sklearn.metrics).
7. סכמו את התוצאות באופן השוואתי וברור והסבירו במילים שלכם (יש להתייחס לנקודות הבאות): את התוצאות יש להציג גם בטבלה השוואתית בנוסף לויזואליזציות 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

8. 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 !