מעבדה 4 - 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. האפשרות הראשונה היא from sklearn.datasets import fetch_20newsgroups
ניתן לבחור מספר קטגוריות בודד לסיווג.
B. אוסף הטוויטים שניתן לכם בתחילת הקורס מכיל תיוג לטוויטים שנחשבים דברי נאצה (תוכן בוטה, גזעני או סקסיסטי)

1. הסבירו וסכמו על סיווג naive bayes במילים שלכם, ובהתייחס לנלמד בהרצאה ובמסלול , באופן כללי וספציפית לסיווג טקסט.
יש להתייחס לסוגים השונים שנלמדו בהרצאה, מתי לדעתכם נשתמש בכל אחד, יתרונות, חסרונות, סקלביליות זמני ריצה וכו'.
2. בעבור חבילות הפייתון הבאות, קיראו על גרסאות naive bayes השונות , סכמו על התכונות, אופי הפעולה, וההפעלה ברמה הטכנית של כל אחד מהגרסאות למסווג הנל.
פרטו מהם המסווגים שמקבילים לאלו שנלמדו בהרצאה, ומה יתאים יותר למטרת סיווג טקסט.
שימו לב בספריות השונות יש יותר מ -2 גרסאות שונות, וגם לגרסאות מקבילות יש קלט שונה והגדרות שונות.
    1. sklearn.naive_bayes
    2. NLTK
    3. NLTK with sklearn wrapper (nltk.classify.scikitlearn module)
    להתייחס לדרך הפעלת ה wrapper
    4. Textblob.classifiers
3. הסבירו וסכמו על feature selection במילים שלכם ובהתייחס לנלמד בהרצאה ובמסלול, באופן כללי וספציפית לטיפול בטקסטים.
יש להתייחס לשיטות השונות שנלמדו (שלוש שיטות שנלמדו בהרצאה)
יש להתייחס ליתרונות וחסרונות של כל שיטה ביחס לאחרות.
יש להתייחס ל feature selection בהקשר של naive bayes לאיזה סוגים ספציפיים הוא עדיף ומה המשמעות של הפעולה עבור המסווג.
4. הסבירו וסכמו במילים שלכם, בהתאם לנלמד בהרצאה, איך מבצעים הערכת מודל בסיווג טקסט, בעזרת אילו מדדים נשתמש.

5. יש לטפל במסמכים בהיבט העיבוד המקדים שימו לב ניתן ורצוי להיעזר בכלים שפותחו בעבר
יש להציג את הגדלים השונים של מבני הנתונים שנוצרו לפני תחילת כל סיווג.
ניתן גם להיעזר בספריות פייתון המיועדות לכך, המטרה היא להגיע לשני ייצוגים:
   1. ייצוג של BOW - count vectors
      from sklearn.feature_extraction.text import CountVectorizer
   2. ייצוג TFIDF
      from sklearn.feature_extraction.text import TfidfVectorizer
6. יש לממש מסווג בייסיאני בעזרת ספריות סטנדרטיות, numpy, pandas .
מומלץ לתכנן את המסווגים הנל לאחר למידה יסודית של פונקציות הספריה לסיווג איתן תעבדו.
בנוסף מומלץ להתחיל בתכנון הממשקים בין המודולים השונים בקוד שלכם כדי לחסוך עבודה בהמשך של התאמות של מבני נתונים שונים ככקלט למסווג.
   1. מולטי נומינלי
   2. מולטי וריאנטי
רפרנס כללי למימוש (רעיון בסיסי למימוש, מחייב התאמות ותשומת לב):
https://towardsdatascience.com/implementing-naive-bayes-in-2-minutes-with-python-3ecd788803fe
7. יש לבנות מודל לסיווג המסמכים בעזרת כל מסווג מסעיף 6
   1. לא לשכוח חלוקה ל train ו test
   2. יש להציג זמני ריצה עבור כל סיווג
   3. יש להציג את גודל מבני הנתונים המייצגים את המודל
   4. יש לבנות מודל בעזרת:
      i. מולטי נומינלי
      ii. מולטי וריאנטי
8. יש לבנות מודל לסיווג המסמכים, בעזרת מסווגים מקבילים מספריות פייתון.
   1. לא לשכוח חלוקה ל train ו test
   2. יש להציג זמני ריצה עבור כל סיווג
   3. יש להציג את גודל מבני הנתונים המייצגים את המודל
   4. יש לסווג בעזרת: רק NLTK
      i. מולטי נומינלי (**משתי ספריות שונות**)
      ii. מולטי וריאנטי (**משתי ספריות שונות**)
9. יש להציג מדדי הערכה לכל מסווג (על פי הנלמד בהרצאה) ניתן להשתמש בספריות פייתון המיועדות לכך (כמו sklearn.metrics).
10. סכמו את התוצאות 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 ואחרי הסיווג (גודל מטריצת הקלט וגודל מטריצת המודל)
   ii. זמני בניית מודל
   iii. אופי הסיווג שהתקבל בכל אחת מהספריות השונות של פייתון וההבדל בינהם לדעתכם.
   iv. ניתוח של תוצאות ההערכה של המודל.
   v. מסקנות

11. 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 !