תרגיל 3

סיווג

מבוא

משימות סיווג נפוצות מאוד בתחום עיבוד השפות, בעיקר כי בעיות שונות ניתנות לתרגום לבעיות סיווג. משימות סיווג רבות שנראות קשות לעין האנושית יכולות להתבצע באופן מעולה ע"י אמצעי למידת מכונה פשוטים. בתרגיל זה נתנסה בסיווג טקסט מתוך פרוטוקולי הכנסת על-פי סוג הפרוטוקול- ועדה או מליאה. כלומר, נבנה תכנית שתאמן מסווגים שונים לסווג יחידות טקסט לשתי מחלקות: וועדות ומליאות. לצורך תרגיל זה נשתמש בקורפוס ה-כנסת שבניתם בתרגילים הקודמים. כמו כן, ניעזר באובייקטים מספריית scikit-learn.

שלב 1: הגדרת המחלקות

1. בתרגיל זה נשתמש בקורפוס הכנסת שיצרתם בתרגילים הקודמים. להזכירכם, בקובץ הCSV של הקורפוס שלנו, יש עמודה המציינת אם המשפט נלקח מפרוטוקול של וועדה או של מליאה. השתמשו בעמודה זו כדי להגדיר את המחלקה של כל משפט.

שלב 2: חלוקה ליחידות סיווג

מאחר ומשפט הוא יחידת שיח קצרה יחסית, זוהי משימת סיווג קשה לביצוע עבור קלטים כאלו. נרצה שביחידת הסיווג שלנו יהיו מספיק תכונות שיהוו מאפיינים משמעותיים למסווג. לכן, נרצה ליצור מקבצים (chunks) של 5 משפטים מתוך הקורפוס, ואלו ישמשו כיחידות הסיווג.

הערות:

1. אם מספר המשפטים במחלקה אינו מתחלק ב-5, ותרו על שארית המשפטים.
2. כל chunk כמובן יכיל רק משפטים השייכים לאותה מחלקה.

שלב 3: איזון המחלקות

על-מנת לסווג באופן מיטבי, נרצה שהמחלקות תהיינה מאוזנות. לשם כך, עשו down-sampling (רנדומלי) למחלקה הגדולה. כלומר, ביחרו באופן רנדומלי פריטים מהמחלקה הגדולה כמספר הפריטים במחלקה הקטנה וזירקו את יתר הפריטים במחלקה, כך שיתקבלו שתי מחלקות באותו הגודל.

<u>כיתבו בדו"ח מה היה מספר הפריטים בכל מחלקה לפני ואחרי הdown-sampling שביצעתם.</u>

שלב 4: יצירת וקטור מאפיינים (feature vector)

1. Bag of Words: עבור כל chunk יצרו וקטור BoW כוקטור מאפיינים. ניתן להשתמש ב-CountVectorizer. ניתן גם לבחור להשתמש ב-Tfidf. הסבירו (בדו"ח) במה בחרתם ומדוע.

2. צרו וקטור משלכם, עם מאפייני סגנון ותוכן. לשם כך, אתם יכולים להסתכל על הדאטה שיש לכם ולחשוב מה יכול לעזור בסיווג. פיצ'רים יכולים להיות למשל, אורך המשפט הממוצע בchunk , בדיקת קיומן של מילות תוכן מסויימות וכיו"ב. הנכם מוזמנים להשתמש כתכונות גם בעמודות אחרות בדאטה, מלבד עמודת הטקסט*.

* שימו לב **שאסור** להשתמש בעמודות של שם הפרוטוקול או של סוג הפרוטוקול המהוות אינדיקציה ברורה למחלקה, כפיצ'רים.

שלב 5: אימון

1. על מנת לסווג את שני סוגי וקטורי המאפיינים שלכם, אמנו שני סוגי מסווגים:
   i. KNearestNeighbors
   ii. SVM
2. העריכו את דיוק המסווגים ב-2 דרכים:
   i. 10-fold Cross Validation
   ii. חלוקה לקבוצת אימון וקבוצת בדיקה בעזרת sklearn_train_test_split . סט הבדיקה יהווה 10% מהדאטה. עליכם לחלק את הדאטה באופן stratified (קיראו על כך בדוקומנטציה של הפונקציה).
3. הוסיפו לדו"ח classification_report המפרט את תוצאות ההערכה בכל אחת מהדרכים עבור כל מודל ועבור כל וקטור מאפיינים.

שלב 6: סיווג

לתרגיל מצורף קובץ בשם knesset_text_chunks.txt, המכיל בכל שורה chunk של טקסטים מהכנסת. עליכם לסווג כל chunk לאחת המחלקות plenary (מליאה) או committee (ועדה) בעזרת המודל שעבורו קיבלתם את התוצאות הטובות ביותר, ולכתוב את הסיווגים לקובץ בשם classification_results.txt .
כל שורה בקובץ תתייחס לchunk שבאותה שורה בקובץ המקורי, ותכיל רק את תוצאת הסיווג "plenary" או "committee".
למשל:

committee
plenary
plenary
committee

...

הערות:

1. שימו לב, שבקובץ הקלט בשלב 6 מופיעים רק הטקסטים עצמם ולא ערכים התואמים לעמודות אחרות, לכן אם השתמשתם באלו בוקטור המאפיינים שיצרתם, לא תוכלו לסווג את הדוגמאות האלו בעזרתו. ביחרו מודל שכן מתאים למשימה.
2. לאורך הקוד יש מספר מקומות בהם יש מידת אקראיות. עליכם להשתמש ב random.seed() וב numpy.random.seed() עם מספר קבוע, על מנת לקבע את התוצאות שלכם, אחרת הם ישתנו בכל ריצה. לשם כך, הוסיפו בתחילת הקוד

```
import random
import numpy as np
random.seed(42)
np.random.seed(42)
```

שאלות

ענו בדו"ח על השאלות הבאות:

1. האם היו הבדלים ב-percision ו-recall בין המחלקות? אם כן, מה ניתן להסיק מהם?
2. האם תוצאות הסיווג בחלוקת אימון-בדיקה של 10%-90% דומות לתוצאות ה-cross validation? בין אם כן ובין אם לאו, נסו לשער מדוע.
3. הסבירו מהם היתרונות והחסרונות של שני סוגי המסווגים KNN, SVM בהם השתמשתם. האם לדעתכם אחד מהם עדיף על פני השני, עבור משימת הסיווג שבתרגיל?
4. פרטו את היתרונות והחסרונות ליצירת יחידות הסיווג. מה יהיו ההשלכות אם נגדיל ואם נקטין אותן באופן משמעותי?

שאלת בונוס:

התנסו בגדלים שונים של יחידות סיווג (chunks) והסבירו מה לדעתכם מספר המשפטים האידיאלי למשימת הסיווג בתרגיל זה? ענו על כך בדו"ח.

ספריות מותרות לשימוש

אתם יכולים להשתמש ב
Pandas, Numpy, scikit-learn
ובכל ספריה סטנדרטית של python.
אתם יכולים לחפש שם של ספריה ב https://docs.python.org/3/library/index.html על מנת לבדוק אם זו ספריה סטנדרטית. לא יהיה מענה על שאלות לגבי שימוש בספריות ספציפיות.

אופן ההגשה

1. ההגשה היא בזוגות בלבד.

2. עליכם להגיש קובץ zip בשם hw3_<id1>_<id2>.zip (כאשר <id1>,<id2> הם מספרי תעודות הזהות של הסטודנט הראשון והשני בהתאמה), המכיל את הקבצים הבאים:
   a. קובץ python בשם **knesset_protocol_classification.py** המכיל את כל הקוד הנדרש כדי לממש את שלבים 1-6.
   b. קובץ text בשם **classification_results.txt** כפי שתואר בשלב 6.
   c. קובץ PDF בשם **<id1>_<id2>_hw3_report.pdf** ובו דו"ח המפרט על הקוד, על ההחלטות שקיבלתם במהלך העבודה על התרגיל, גודל המחלקות כפי שתואר בשלב 3, ה calssification reports כפי שתואר בשלב 5, ומענה על השאלות. אל תשכחו לציין בתחילת הדו"ח את שמותיכם ותעודות הזהות שלכם.

יש להקפיד על עבודה עצמית, צוות הקורס יתייחס בחומרה להעתקות או שיתופי קוד, כמו גם שימוש בכלי AI דוגמת chatGPT.

ניתן לשאול שאלות על התרגיל בפורום הייעודי לכך במודל.

יש להגיש את התרגיל עד לתאריך 25.2.24 בשעה 23:59.

בהצלחה!