תרגיל 3

סיווג

מבוא

משימות סיווג נפוצות מאוד בתחום עיבוד השפות, בעיקר כי בעיות רבות 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 הjsonl של הקורפוס שלנו יש שדה המציין אם המשפט נלקח מפרוטוקול של וועדה או של מליאה. השתמשו בשדה זה כדי להגדיר את המחלקה של כל משפט.

שלב 2: חלוקה ליחידות סיווג

מאחר ומשפט הוא יחידת שיח קצרה יחסית, משימת הסיווג עבור קלטים כאלו קשה לביצוע. 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, בחרו באופן רנדומלי פריטים מהמחלקה הגדולה כמספר הפריטים במחלקה הקטנה וזרקו את יתר הפריטים במחלקה, כך שיתקבלו שתי מחלקות באותו הגודל.

<u>כתבו 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 בשדות 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. LogisticRegression
2. העריכו את דיוק המסווגים ב-2 דרכים:
   i. 5-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.

הערה: למסווגים השונים יש פרמטרים שונים שאתם יכולים לקנפג או להשאיר את ברירות המחדל, לפי בחירתכם. הסבירו את החלטותיכם בדו"ח.

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. בחרו 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 הן ישתנו בכל ריצה. לשם כך, הוסיפו בתחילת הקוד:

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

שלב 7: בחנו את גודל הchunks

התנסו בגדלים שונים של יחידות סיווג (chunks), בחנו אותם והסבירו:

1. מה לדעתכם מספר המשפטים האידיאלי למשימת הסיווג בתרגיל זה?
2. פרטו את היתרונות והחסרונות ליצירת יחידות הסיווג. מה יהיו ההשלכות אם נגדיל ואם נקטין אותן באופן משמעותי?

שאלות

ענו בדו"ח על השאלות הבאות:

1. נניח שאתם משתתפים בתחרות מודלים לחיזוי שבה אם המודל שלכם יחזה נכון את כל הדוגמאות מסוג committee, תקבלו פרס כספי גדול, ואם המודל שלכם יטעה על אפילו דוגמה אחת מסוג committee תקבלו קנס כספי גבוה.
מבין המדדים המופיעים בclassification report, איזה מדד תרצו למקסם? איזה מהמודלים שאימנתם תבחרו למטרה זו? הסבירו.
2. ענו שוב על 1 כאשר שינו את החוקים בתחרות וכעת אם המודל שלכם יסווג נכון את כל הדוגמאות (plenary וcommittee) תקבלו פרס כספי גבוה, אבל אם המודל שלכם יסווג אפילו דוגמה אחת בצורה לא נכונה, תקבלו קנס כספי גבוה.
3. האם תוצאות הסיווג בחלוקת אימון-בדיקה של 10%-90% דומות לתוצאות ה-cross validation? בין אם כן ובין אם לאו, נסו לשער מדוע. איזו משיטות ההערכה אמינה יותר לדעתכם?
4. הסבירו מהם היתרונות והחסרונות של שני סוגי המסווגים KNN, LogisticRegression בהם השתמשתם. האם לדעתכם אחד מהם עדיף על פני השני, עבור משימת הסיווג שבתרגיל?

הערות כלליות

1. על הקוד שלכם להיות מסוגל להתמודד עם שגיאות בכל שלב בתהליך ולא לקרוס. השתמשו ב-Try Except blocks לפי הצורך.
2. שימו לב, בבדיקת תרגילי הבית בקורס ניתן משקל גדול מהניקוד הן על **הדו"ח**, ההסברים והידע שהפגנתם בחומר הנלמד והן על **הקוד**, אופן המימוש, יעילותו, קריאותו ועמידותו. בפרט, הרבה מהבדיקות הן אוטומטיות ולכן עליכם להקפיד על קוד תקין שרץ ללא שגיאות ועל עמידה **מדוייקת** בפלט הנדרש וביתר ההנחיות.
3. ניתן לשאול שאלות על התרגיל בפורום המיועד במודל. למעט מקרים אישיים מיוחדים, אין לשלוח שאלות הקשורות לתרגיל הבית במייל.
4. על אחריותכם לעקוב אחר הודעות הקורס במודל (בלוח הודעות ובפורום) ולהיות מעודכנים במידה ויהיו שינויים בהנחיות.

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.

- למען הסר ספק, json היא ספרייה סטנדרטית של python.
- מומלץ להשתמש עבור כל פרוייקט בסביבה וירטואלית virtual environment חדשה משלו על מנת להיות בטוחים שאתם משתמשים רק בספריות מותרות ולמנוע קונפליקטים עם ספריות קודמות שהתקנתם בעבר. ראו מצגת על כך במודל.

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 1-7.

i. - הקלט לקובץ יהיה נתיב לקובץ הקורפוס, נתיב לקובץ משפטים לסיווג, נתיב לתיקיית פלט

- הפלט יהיה קובץ הסיווגים כפי שתואר בשלב 6 שמור בתיקייה שהתקבלה בקלט. בשלב הגשת התרגיל על הקובץ לא להדפיס שום דבר למסך. נטרלו את הדפסת הclassification reports לקראת ההגשה.

ii. על הקובץ לרוץ תחת הפקודה (ללא הסימונים <>):

```
python knesset_protocol_classification.py <path/to/corpus_file.jsonl> <path/to/sentences_texts_file.txt> <path/to/output_dir>
```

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 **בעברית**.

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 1.7.24 בשעה 23:59.

בהצלחה!