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 הדובר. כלומר, נבנה תכנית שתאמן מסווגים שונים לסווג יחידות טקסט לשני דוברים. לצורך תרגיל זה נשתמש בקורפוס ה-כנסת שבניתם בתרגילים הקודמים. כמו כן, ניעזר באובייקטים מספריית scikit-learn.

שלב 1: הגדרת המחלקות

בתרגיל זה נשתמש בקורפוס הכנסת שיצרתם בתרגילים הקודמים. מי שקיבלו הערות משמעותיות על הקורפוס שלהם מצופים לתקן אותן עבור תרגיל זה. להזכירכם, בקובץ הjsonl של הקורפוס שלנו יש שדה המציין את שם הדובר. מצאו את שני הדוברים עם המספר הכי גדול של משפטים בקורפוס שלכם. דוברים אלו יהוו את המחלקות שלכם לסיווג באופן הבא:

1. משימת סיווג בינארית binary classification: עליכם לחלק את הדאטה לשתי מחלקות - כל אחד משני הדוברים מהווה מחלקה בפני עצמו. עבור כל מחלקה השתמשו בכל הרשומות של הדובר המתאים ורק בהם. יחידת הסיווג במשימה זו תהיה משפט אחד. כתבו בדו"ח מי הם שני הדוברים שיהוו את המחלקות שלכם.
2. משימת סיווג מרובת מחלקות multi-class classification: במשימה זו יהיו לכם 3 מחלקות: אחת עבור כל אחד משני הדוברים שמצאתם, כמו במשימה 1, והשלישית תהיה מחלקה של "אחר" בה יהיו כל הרשומות הנותרות של שאר הדוברים. יחידת הסיווג במשימה זו תהיה משפט אחד.

הערה: אותו דובר בקורפוס שלכם יכול להופיע במספר דרכים שונות. למשל "בנימין גנץ" ו"בני גנץ", או הבדלים שנובעים מטעויות בחילוץ ונקיון השמות. עבור **בחירת המחלקות**, אינכם נדרשים לאחד את השמות של אותו דובר. בחירת המחלקות תתבצע רק על פי מחרוזת השם, וכל מחרוזת שונה תחשב לשם דובר אחר. יחד עם זאת, על מנת למקסם את כמות המשפטים לאימון ולהמנע מoverfitting, **לאחר** בחירת שתי המחלקות הראשיות (שני הדוברים בעלי מספר המשפטים הרב ביותר), יש לנסות לכלול כמה שיותר משפטים של שני דוברים אלו תחת המחלקה שלהם, גם אם מחרוזת השם שלהם כתובה בדרכים שונות בקורפוס. הסבירו בדו"ח איך התמודדתם עם דרישה זו.

שלב 2: איזון המחלקות

על-מנת לסווג באופן מיטבי, נרצה שהמחלקות תהיינה מאוזנות. לשם כך, עבור כל אחת מהמשימות עשו down-sampling (רנדומלי) למחלקות הגדולות. כלומר, בחרו באופן רנדומלי פריטים מהמחלקה/ות הגדולה/ות כמספר הפריטים במחלקה הקטנה וזרקו את יתר הפריטים במחלקה, כך שיתקבלו שתי מחלקות באותו הגודל עבור המשימה הבינארית ושלוש מחלקות באותו הגודל עבור המשימה רבת המחלקות.

כתבו בדו"ח מה היה מספר הפריטים בכל מחלקה לפני ואחרי הdown-sampling שביצעתם.

שלב 3: יצירת וקטור מאפיינים (feature vector)

עליכם ליצור שני וקטורים שונים באופן הבא:

1. Bag of Words: עבור כל משפט 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, סימני פיסוק, צירופי מילים מסויימים וכיו"ב. הנכם מוזמנים להשתמש כתכונות גם בעמודות אחרות בדאטה, מלבד עמודת הטקסט. **בוקטור זה אסור לכם להשתמש בוקטור BoW**.

שלב 4: אימון

1. על מנת לסווג את שני סוגי וקטורי המאפיינים שלכם, אמנו שני סוגי מסווגים:
   i. KNearestNeighbors
   ii. LogisticRegression
2. העריכו את דיוק המסווגים ע"י 5-fold Cross Validation
3. הוסיפו לדו"ח classification_report המפרט את תוצאות ההערכה עבור כל משימה, עבור כל מסווג ועבור כל וקטור מאפיינים.

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

שלב 5: סיווג

לתרגיל מצורף קובץ בשם knesset_sentences.txt, המכיל בכל שורה משפט מתוך טקסטים של הכנסת. עליכם לסווג כל משפט לאחת המחלקות: הדובר הראשון שלכם (זה שהיה בעל מספר המשפטים הכי גדול בקורפוס כפי שמצאתם בשלב 1), הדובר השני (בעל מספר המשפטים השני בגודלו בקורפוס), או "אחר", בעזרת אחד מהמודלים שאימנתם, לבחירתכם. עליכם לכתוב את הסיווגים לקובץ בשם classification_results.txt.

כל שורה בקובץ תתייחס למשפט שבאותה שורה בקובץ המקורי, ותכיל רק את תוצאת הסיווג "first", "second", "other". ללא שורות רווח.

למשל:

first

first

other

second

…

הערות:

1. שימו לב, שבקובץ הקלט בשלב 5 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)
```

שאלות

ענו בדו"ח על השאלות הבאות:

1. מה הם האתגרים שיכולים להיווצר בשימוש במחלקה "אחר" במשימת הסיווג?
2. נניח שאתם משתתפים בתחרות מודלים לחיזוי בינארי שבה אם המודל שלכם יחזה נכון את כל הדוגמאות של הדובר הראשון, תקבלו פרס כספי גדול, ואם המודל שלכם יטעה על אפילו דוגמה אחת של הדובר הראשון תקבלו קנס כספי גבוה.

   מבין המדדים המופיעים בclassification report, איזה מדד תרצו למקסם? איזה מהמודלים שאימנתם תבחרו למטרה זו? הסבירו.
3. ענו שוב על 1 כאשר שינו את החוקים בתחרות וכעת אם המודל שלכם יסווג נכון את **כל** הדוגמאות של **שני** הדוברים תקבלו פרס כספי גבוה, אבל אם המודל שלכם יסווג אפילו דוגמה אחת בצורה לא נכונה, תקבלו קנס כספי גבוה.
4. הסבירו מה היתרונות והחסרונות של שיטת הcross validation על פני חלוקה פשוטה למחלקת אימון ובדיקה. איזו משיטות ההערכה אמינה יותר לדעתכם?
5. הסבירו מהם היתרונות והחסרונות של שני סוגי המסווגים KNN, LogisticRegression בהם השתמשתם. האם לדעתכם אחד מהם עדיף על פני השני, עבור משימות הסיווג שבתרגיל?
6. יחידת הסיווג בתרגיל היא משפט אחד. אם במקום זאת, היינו מחליטים על יחידת סיווג שמאחדת יחד מספר משפטים מאותה מחלקה, כך שכל דוגמה לסיווג הייתה מקבץ של משפטים. מה היו היתרונות והחסרונות בכך? התייחסו בתשובתכם ליחידות סיווג של 2, 5, 10, 100 משפטים.
7. איזה גודל של יחידת סיווג עדיף לדעתכם (1, 2, 5, 10, 100) במשימות שלנו? הסבירו.

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_speaker_classification.py** המכיל את כל הקוד הנדרש כדי לממש את שלבים 1-5.

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

- הפלט יהיה קובץ הסיווגים כפי שתואר בשלב 5 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_speaker_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 5.

c. קובץ PDF בשם **<id1>_<id2>_hw3_report.pdf** ובו דו"ח המפרט על הקוד, על ההחלטות שקיבלתם במהלך העבודה על התרגיל, הcalssification reports, מענה על השאלות וכל מה שהתבקשתם לפרט לאורך התרגיל. <u>אל תשכחו לציין בתחילת הדו"ח את שמותיכם ותעודות הזהות שלכם **בעברית**.</u>

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 02.01.25 בשעה 23:59.

בהצלחה!