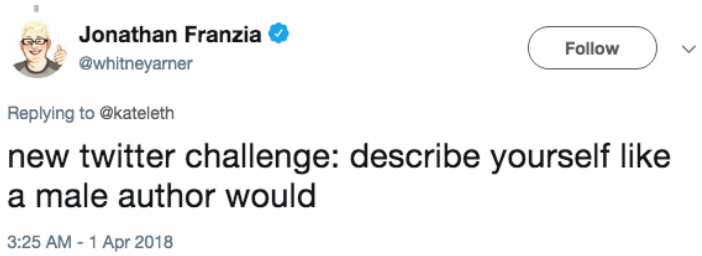
תרגיל 3

סיווג

מבוא

משימות סיווג נפוצות מאוד בתחום של עיבוד השפות, בעיקר כי בעיות שונות ומגוונות ניתנות לייצוג כבעיות סיווג. משימות סיווג רבות, כולל כאלו שאנשים מתקשים בהן מאוד, יכולות להתבצע באופן מעולה ע"י אמצעי למידת מכונה פשוטים. בתרגיל זה נתנסה בסיווג טקסט על-פי מגדר הכותבת, כלומר, נבנה תכנית שתאמן מסווגים שונים לסווג יחידות טקסט לשתי מחלקות: האחת טקסט שנכתב ע"י גברים, והשנייה טקסט שנכתב ע"י נשים. לצורך תרגיל זה נשתמש בקורפוס ה-BNC, ונניח מגדר בינארי (male ו-female). כמו כן, ניעזר באובייקטים מספריית scikit-learn.



שלב 1: הגדרת המחלקות

מרבית הטקסטים ב-BNC אינם מפרטים את מגדר הכותבת, אך מצוין בהם שם הכותבת. לכן, ניעזר בשדה זה על-מנת לקבוע את מגדר הכותבת של כל טקסט:

1. הוסיפו שדה למחלקה Sentence ובה שם ה-author, אם ישנו.
2. היעזרו ב-Gender Guesser בכדי לזהות את מגדר הכותבת. לעיתים יש יותר מכותבת אחת.
3. נרצה להשתמש בטקסטים שבהם מגדר הכותבת וודאי. לכן, נוסיף ל-Sentence שדה בשם gender ובו 3 תגיות: זכר, נקבה ולא ידוע (מתייחס למקרים בהם יש יותר מכותבת אחת ואלו בהם המגדר שונה, מגדר שמסומן unkown וכדומה).

שלב 2: חלוקה ליחידות סיווג

מאחר ומשפט הוא יחידת שיח קצרה יחסית, זוהי משימה סיווג קשה (מאוד!) לביצוע עבור קלטים כאלו. נרצה שיחידת הסיווג שלנו תכיל מספיק טקסט כדי לחלץ ממנו מאפיינים משמעותיים למסווג. לכן, נרצה ליצור

מקשות של 10 משפטים רציפים מתוך הקורפוס, ואלו ישמשו כיחידות הסיווג. עבור כל קלט, ככל שגודל יחידת הסיווג גדל, כך מספר יחידות הסיווג יורדות. לכן, מאחר וסיווג דורש כמות נאה של דוגמאות, לתרגיל זה לא מצורף קלט לדוגמא: אתן מתבקשות להוריד את קבצי ה-BNC (הוראות על ההרצה בהמשך הקובץ).

בשלב זה הוסיפו למחלקה Corpus רשימה המכילה מקשות כאלה (chunks של 10 משפטים רציפים) מתוך הקורפוס. אם מספר המשפטים בקובץ אינו כפולה שלמה של 10, ותרו על שארית המשפטים. ניתן ליצור מחלקה חדשה בשם Chunk או להוסיף שדה של כל ה-chunks ל-Corpus, המימוש עצמו תלוי בכן.

שלב 3: יצירת מחלקה לסיווג

בקובץ ה-Python המצורף לתרגיל כלול template לתרגיל, ובו מחלקה בשם Classify. מחלקה זו נבנית על סמך קורפוס, ולכן עליה להיות קשורה לאובייקט Corpus כלשהו. השתמשו במחלקה זו (שדות ומתודות) על מנת לבצע את המשימות הבאות.

שלב 4: סיווג

עקבו אחר השלבים הבאים:

1. הורידו את קורפוס ה-BNC. צרו תיקייה משלכן ובה 1000 קבצים כלשהם ממנו. על תיקייה זו תבצעו את הבדיקות שלכן. הקלט לתרגיל יהיה תיקייה ובה קבצי XML מתוך קורפוס זה.
2. קראו את הקבצים בתיקיות לתוך אובייקט Corpus שיצרתן וקישרתן לאובייקט Classify.
3. הכניסו את כל יחידות הסיווג שלהן gender וודאי (male או female , ללא unkown) ל-Classify בכל דרך שנוחה לכן.
4. על-מנת לסווג באופן מיטבי, נרצה שהמחלקות תהיינה מאוזנות. הדפיסו לקובץ הפלט את מספר הפריטים בכל מחלקה, עשו down-sampling (רנדומלי) למחלקה הגדולה יותר והדפיסו את מספר הפריטים בכל מחלקה לאחר השינוי. אופן הפלט מפורט בהמשך.
5. 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.
6. צרו גם וקטור משלכן, שהתכונות בו משקפות מאפייני סגנון ותוכן.
7. השתמשו במסווג KNearestNeighbors על מנת לסווג. סווגו ב-2 דרכים: האחת 10-fold Cross Validation והשנייה סיווג רגיל: חלוקה לקבוצת אימון וקבוצת בדיקה (יחס של 3:7), אמנו ובחנו את המודל. הוסיפו לקובץ הפלט דו"ח המפרט על תוצאות האימון.
8. אמנו את שני מודלים: האחד לפי וקטור ה-BoW והשני לפי הוקטור שבניתן.
9. צרפו לדו"ח את התוצאות שקיבלתן.

על הפלט (בקובץ הפלט) להיראות כך:

Before Down-sampling:
Female: <num of chunks written by females> Male: <num of chunks written by males>

After Down-sampling:
Female: <num of chunks written by females> Male: <num of chunks written by males>

== BoW Classification ==
Cross Validation Accuracy: <cross-val accuracy, in percentage, 3 decimal digits>
<Classification report>

== Custom Feature Vector Classification ==
Cross Validation Accuracy: <cross-val accuracy, in percentage, 3 decimal digits>
<Classification report>

ענו בדו"ח על השאלות הבאות:

1. האם היו הבדלים ב-percision ו-recall בין המחלקות? אם כן, מה ניתן להסיק מהם עבור כתיבה של נשים אל מול כתיבה של גברים?
2. האם תוצאות הסיווג הרגיל דומות לתוצאות ה-cross validation? בין אם כן ובין אם לאו, נסו לשער מדוע.
3. איזה משני המודלים (זה המתבסס על BoW וזה שמתבסס על התכונות שאתן הגדרתן) הפיק דיוק גבוה יותר? מדוע?
4. האם למשימה כזאת עדיף להשתמש כתכונות במילות תוכן או במילים דקדוקיות (פונקציונליות)? מדוע?

אופן הגשה

1. לתרגיל מצורף קובץ Python בשם **<surname>_<first_name>_ex3.py**. עליכן לממש את התכנית (בגרסה 3.5 ומעלה) ולשנות את שם הקובץ בהתאם (לדוגמא: zeira_yuli_ex3.py). **אין להשתמש בספריות nltk, spacy או transformers**. על הקובץ לרוץ תחת הפקודה:
python <name_ex3.py> <xml_dir> <output_file>
(כמצוין בקובץ עצמו). **לקובץ הפלט יש לכתוב בקידוד utf-8**. הקפידו שזמן הריצה של התכנית לא יעלה על 10 דקות.
2. קובץ PDF בשם **<surname>_<first_name>_report3.pdf** (לדוגמא: zeira_yuli_report3.pdf) ובו דו"ח המפרט על הקוד שכתבתן ועל ההחלטות שקיבלתן במהלך העבודה. על הדו"ח להיות באורך שני עמודים לכל היותר.

*** יש להגיש את שני הקבצים הללו בנפרד, ולא כקובץ zip (ודומיו) ***

יש להקפיד על עבודה עצמית, צוות הקורס יתייחס בחומרה להעתקות או שיתופי קוד.
ניתן לשאול שאלות על התרגיל בפורום הייעודי לכך במודל.
יש להגיש את התרגיל עד לתאריך 30.04.22 בשעה 23:59.

בהצלחה!