עיבוד שפות טבעיות סמסטר א' תשפ"א

תרגיל בית מספר 3: **סיווג טקסטים**

מועד הגשת התרגיל: 12.12.2020 בשעה 23:59

בתרגיל זה תתנסו בסיווג טקסטים. הסייוג יתבצע בעזרת חבילת פייתון שימושית ביותר : scikit-learn. חלק חשוב מהתרגיל הזה הוא היכולת לקרוא ולהבין את התיעוד של החבילה כדי שתוכלו להשתמש בה בצורה טובה. נושאים רלבנטיים לתרגיל אפשר למצוא בקישורים הבאים:

תיעוד כללי ודוגמאות : https://scikit-learn.org/stable/tutorial/index.html
סיווג טקסטים: https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html
הערכת ביצועים של מסווג: https://scikit-learn.org/stable/modules/cross_validation.html
בחירת features לסיווג: https://scikit-learn.org/stable/modules/feature_selection.html

בתרגיל זה עליכם להתמודד עם משימת סיווג בינארית, בכמה אופנים שונים.

במשימה זו תידרשו להפריד בין טקסט של ויקיפדיה באנגלית תקנית לבין טקסט של ויקיפדיה באנגלית פשוטה. הקלט יכלול קבצי טקסט שמבוססים על קבוצה זהה של ערכי ויקיפדיה באנגליות סטנדרטית ובאנגלית פשוטה. קבצי הקלט יהיו באותו פורמט של קבצי הקלט שקיבלתם בתרגיל 2. **יחידת הסיווג למשימה זו היא משפט בודד** (כלומר, יש לבנות feature vector לכל משפט). כדי ליצור משימת סיווג מאוזנת, השוו את מספר המשפטים בשתי הקבוצות, כך שעבור כל ערך ויקיפדיה יהיו ברשותכם מספר זהה של משפטים באנגלית תקנית ובאנגלית פשוטה (ע"י דגימה אקראית של משפטים מהערך ה"ארוך" בכמות ששווה למספר המשפטים בערך ה"קצר"). כך, לדוגמה, הערך banana באנגלית סטנדרטית כולל 346 משפטים, ובאנגלית פשוטה 95 משפטים. דגמו באקראי 95 משפטים מתוך הערך באנגלית הסטנדרטית ע"מ לאזן בין הערכים השונים. שימו לב שייתכן מצב שבו הכיוון יתהפך, כלומר יהיו יותר משפטים עבור ערך כלשהו באנגלית פשוטה ועליכם לתמוך בכך. אספו את כל המשפטים שקיבלתם מכל שפה בנפרד וערבבו אותם. כעת יהיה ברשותכם data set שיכלול שתי קבוצות שוות גודל של משפטים מעורבבים, אחת באנגלית סטנדרטית ואחת באנגלית פשוטה.

שלב א'

סיווג בעזרת Bag of Words:

1. צרו מהטקסטים feature vectors. ניתן להשתמש לצורך כך ב CountVectorizer שממומש ב scikit learn. חשבו את ערך ה tf-idf לכל מילה בטקסט. ניתן להשתמש ב TfidfTransformer שגם ממומש ב scikit learn. כמה מילים שונות מופיעות בכל אחד מהקורפוסים?
2. הגדירו לכל יחידת סיווג את המחלקה המתאימה: 1 (אנגלית סטנדרטית) או 0 (אנגלית פשוטה).
3. סווגו את הטקסטים בעזרת המסווגים הבאים (לשלושתם קיים מימוש ב scikit learn):
   - Naive Bayes (MultinomialNB)
   - KNN
   - Logistic Regression
4. העריכו את הביצועים של כל מסווג בעזרת ten-fold-cross-validation ודווחו על הדיוק (accuracy) הממוצע של כל ה folds. פרטו את התוצאות בדו"ח. הסבירו האם התוצאות תואמות את הציפיות שלכם, והציגו את ההבדלים בביצועים של המסווגים.

שלב ב'

סיווג על בסיס 300 המילים הנפוצות ביותר בשפה האנגלית: נתון קובץ עם 300 המילים הנפוצות ביותר באנגלית.

1. בנו feature vector באורך 300 לכל משפט. כל ערך בוקטור הוא אינדיקציה בוליאנית להופעת המילה במשפט, כך שעבור המילה ה i ברשימה, הערך ה i בוקטור יהיה 1 אם המילה נמצאת במשפט ו 0 אחרת.
   בשלב זה עליכם לייצר את הוקטור בעצמכם ולא באמצעות מבני נתונים מוכנים של Scikit-learn.
2. חזרו על סעיף 2-4 משלב א'.

שלב ג'

סיווג בעזרת בחירת מאפיינים (features) באופן ידני:

1. חשבו על מאפיינים שיכולים להיות משמעותיים לסיווג. מאפיינים אפשריים הם אורך משפט, סימני פיסוק, צירופי מילים מסוימים, אורך מילה וכן הלאה. נסו להבחין אלה מהמאפיינים קשורים לסגנון הכתיבה ואילו קשורים יותר לתוכן. נסו לחשוב אילו מאפיינים יהיו שימושיים למשימת הסיווג הזו. יש לכלול לפחות 10 מאפיינים שונים שקשורים לסגנון, ועוד לפחות 5 שקשורים לתוכן.
2. לכל יחידת סיווג הגדירו feature vector: וקטור של מספרים (שאורכו כמספר המאפיינים) אשר מייצג את המאפיינים שבחרתם למטרת הסיווג.
3. החבילה scikit-learn מספקת כלים לביצוע ניתוח הקלט, על מנת לאמוד את התרומה היחסית של כל מאפיין למשימת הסיווג. השתמשו ב SelectKBest על מנת לזהות את המאפיינים בעלי התרומה הגבוהה ביותר לסיווג ופרטו את התוצאות בדו"ח. ניתן להשתמש בניתוח הזה באופן איטרטיבי ע"מ לעדכן את המאפיינים עד שתגיעו לדיוק טוב ככל שתוכלו. תעדו את תהליך בחירת המאפיינים שעשיתם עד לגרסה הסופית שבה השתמשתם לסיווג.
4. חזרו על סעיפים 2-4 משלב א'.

הוראות הגשה:

- הפעילו את המסווגים עם ערכי ברירת המחדל של הפרמטרים – כלומר יצרו אותם בשימוש ב constructor הדיפולטי - ללא פרמטרים.
- הגישו את הקבצים הבאים:
  - קובץ קוד בשם hw3.py
  - אם דרושים לכם קבצים נוספים על מנת שהקוד יהיה מסודר, אתם יכולים להוסיף אותם – רק שימו לב שהריצה תקינה ושלא חסר דבר.
  - דוח בפורמט PDF בשם hw3Report.pdf.

- על הקוד להיות מופעל משורת הפקודה בעזרת הפקודה:

```
python hw3.py <input_dir_path> <top_300_path> <output_path>
```

באשר:

- input_dir_1 הוא הנתיב המלא לתיקייה שבה נמצאים קבצי הקלט למשימת הסיווג
- top_300_path הוא הנתיב המלא ושם הקובץ שמכיל את רשימת 300 המילים הנפוצות באנגלית.
- output_path הוא השם והנתיב המלא של קובץ פלט שיכיל את תוצאות הסיווגים, על פי הפורמט המוגדר בהמשך.

על התכנית לכתוב לקובץ הפלט את התוצאות של שלבים א' עד ג' – כלומר את הדיוק של המסווגים השונים, עבור שתי משימות הסיווג, בכל אחת מהדרכים שהוגדרו בתרגיל. את הדיוק הדפיסו באחוזים, בדיוק של 2 ספרות אחרי הנקודה העשרונית. הפלט צריך להיראות כך:

Phase1 (Bag of Words):
Naïve Bayes: <accuracy>
KNN: <accurach>
Logistic Regression: <accuracy>

Phase2 (300 most frequent words):
Naïve Bayes: <accuracy>
KNN: <accurach>
Logistic Regression: <accuracy>

Phase3 (My featurs):
Naïve Bayes: <accuracy>
KNN: <accurach>
Logistic Regression: <accuracy>:

- הדו"ח שאתם מגישים צריך לכלול תשובות לשאלות שמופיעות בסעיפים הבאים:
  - שלב א' סעיף 1.
  - שלב ג' סעיף 1.

בנוסף פרטו עבור כל אחד מהשלבים את התשובות שלכם לשאלות שמופיעות בשלב א' סעיף 4.

יש להקפיד על עבודה עצמאית. צוות הקורס יתייחס בחומרה להעתקות או שיתופי קוד.
תאריך הגשה: 12.12.2020 , עד השעה 23:59.
שאלות על התרגיל אפשר לשאול בפורום תרגילי בית.

בהצלחה!