שחר טפלר

עיבוד שפות טבעיות – תרגיל בית 3

חלק 2 – חלוקה ליחידות סיווג

יצרתי מחלקה חדשה בשם Chunk, שמכילה את השדות:

- gender – המגדר של הכותב/ת עבור המשפטים
- sentences – רשימה של אובייקטים של Sentence

יצרתי גם מתודה ב-Corpus שעוברת על כל המשפטים ומחלקת אותם ל-chunks בגודל מוגדר של CHUNK_SIZE (במקרה שלנו, 10).

חלק 3 – יצירת מחלקה לסיווג

הוספתי למחלקה Classify את השדה corpus כך שהם יהיו מקושרים. בנוסף, המחלקה מכילה את כל ה-chunks שיש להם מגדר מוגדר ע"י סינון.

חלק 4 – סיווג

על מנת ליצור ווקטור BoW בחרתי להתשמש ב-Tfidf.

TF-IDF מאוד דומה ל-BoW אבל ההבדל העיקרי הוא שהערכים הם לפי תדירות המילה, אבל ב-Inverse. כלומר, ככל שמילה היא יותר נדירה, תהיה לה חשיבות גדולה יותר. זה מעניק יתרון על פני BoW מכיוון שבשיטה זו, אנחנו נותנים פחות חשיבות למילים שאין להן באמת משמעות והן חלק תחבירי של המשפט (stop-words), ומזהים מילים ייחודיות שיעניקו לנו יותר מידע עבור משימת הסיווג.

את ווקטור המאפיינים יצרתי ע"פ המאפיינים הבאים:

- מספר הטוקנים הממוצע למשפט
- מספר סימני הפיסוק הממוצע למשפט
- אורך מילה ממוצע במשפט
- יחס המילים הייחודיות (amount of unique / amount of words)

שאלות:

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?

   כן, היו הבדלים אבל הם לא היו מאוד גדולים. ניתן להסיק מזה שקשה להבדיל בין כתיבה של נשים לכתיבה של גברים.

2. האם תוצאות הסיווג הרגיל דומות לתוצאות ה-validation cross? בין אם כן ובין אם לאו, נסו לשער מדוע.

התוצאות ב-BoW לא דומות בין הסיווג הרגיל ובין ה-validation cross, התוצאה של הסיווג הרגיל גבוהה יותר. אני משערת שזה קרה בגלל שהיה מצב של overfitting בסיווג הרגיל, וב-cross validation זה לא קרה כי השיטה עצמה באה למנוע את זה.

בווקטור הפיצ'רים שאנחנו בנינו אין הבדל כמעט בכלל, וזאת מכיוון שהתוצאות נמוכות ואף מודל לא הגיע למצב של overfitting (שניהם ב-underfitting).

3. איזה משני המודלים (זה המתבסס על BoW וזה שמתבסס על התכונות שאתן הגדרתן) הפיק דיוק גבוה יותר? מדוע?

מודל ה-BoW הפיק דיוק יותר גבוה מהמודל שאני הגדרתי. כנראה זה קרה מכיוון שבחרתי פיצ'רים שפחות מתאימים למשימת הסיווג של מגדר הכותבת, או שהייתי יכולה להשתמש ביותר פיצ'רים.

בנוסף, בניגוד לווקטור הפיצ'רים, מודל ה-BoW מאפשר לנו לתת ציון של חשיבות לכל טוקן, שגם לזה יש חלק בסיווג המגדר.

4. האם למשימה כזאת עדיף להשתמש כתכונות במילות תוכן או במילים דקדוקיות (פונקציונליות)? מדוע?

למשימת סיווג מגדר הכותבת, נעדיף להשתמש במילים דקדוקיות מכיוון שאנחנו רוצים לסווג לפי <u>סגנון הכתיבה</u> – זה מה ש"מפריד" בצורה הטובה ביותר בין כתיבה של נשים לבין גברים (זה גם נאמר בהרצאה). לעומת זאת, התוכן יכול להיות מאוד דומה ואף כמעט זהה, כי מדובר בטקסט של נושא מסוים, שמצריך שימוש במילים מסוימות.

שחר טפלר

תוצאות ריצה:

```
Before Down-sampling:
Female: 18978   Male: 40624
After Down-sampling:
Female: 18978   Male: 18978
== BoW Classification ==
Cross Validation Accuracy: 78.805%
              precision    recall  f1-score   support

      female       0.86      0.94      0.90      5708
        male       0.94      0.85      0.89      5679

    accuracy                           0.90     11387
   macro avg       0.90      0.90      0.90     11387
weighted avg       0.90      0.90      0.90     11387


== Custom Feature Vector Classification ==
Cross Validation Accuracy: 57.288
              precision    recall  f1-score   support

      female       0.59      0.59      0.59      5708
        male       0.58      0.58      0.58      5679

    accuracy                           0.58     11387
   macro avg       0.58      0.58      0.58     11387
weighted avg       0.58      0.58      0.58     11387
```