דו"ח תרגיל בית 3: מגיש : ניב ספיר - 308315944

אתחיל בדגשים חשובים :

- בניתי 2 מודלים כנדרש. האחד לפי 10 fold cross validation והשני לפי חלוקה לקבוצות אימון ומבחן לפי היחס הנדרש 7:3 .

-יצרתי שני וקטורים כנדרש . כל וקטור עבר אימון על כל אחד מהמודלים. כלומר וקטור BOW שיצרתי בעזרת TfidfVectorizer עבר אימון תחת 10 fold cross validation וגם תחת חלוקה לקבוצות אימון. כנ"ל גם לגבי הוקטור BOW שיצרתי בעצמי.

- לפי השלבים שרשמתי לעיל, קיבלתי 4 תוצאות כמו כן כל וקטור עבר 2 מודלי אימון. סה"כ 4 תוצאות. (2 רפורטים ועוד 2 תוצאות דיוק לכל וקטור).

-כמו כן בחרתי לממש קלאסס בשם chunk וגם קלאסס בשם classify כנדרש. בנוסף זהינו את זהות המין של הסופר לפי הפונקציות המוצעות בתרגיל.

-הוקטור שיצרתי בעצמי מחשב לכל מילה משקל לפי תדרויות של המילה מול סה"כ כל המילים בקורפוס. לאחר מכאן מיינתי את הווקטור מהמשקל הגדול לקטן . כלומר מי שקיבל משקל גדול זאת מילה אשר יש לה תדירות גבוה שנכתב ע"י גבר או אישה.

תוצאות : BOW

```
== BoW Classification ==

Model 10 folds:
Cross Validation Accuracy: 0.8030651340996169
              precision    recall  f1-score   support

        male       0.91      0.99      0.95      6527
      female       0.98      0.90      0.94      6527

    accuracy                           0.94     13054
   macro avg       0.95      0.94      0.94     13054
weighted avg       0.95      0.94      0.94     13054

Model split val 3:7 :
3:7 split Accuracy: 0.8955833546081184
              precision    recall  f1-score   support

        male       0.88      0.98      0.93      6527
      female       0.98      0.87      0.92      6527

    accuracy                           0.92     13054
   macro avg       0.93      0.92      0.92     13054
weighted avg       0.93      0.92      0.92     13054
```

MY VECTOR:

```
== Custom Feature Vector Classification ==

Model 10 folds:
Cross Validation Accuracy: 0.5969348659003831
              precision    recall  f1-score   support

        male       0.76      0.86      0.81      6527
      female       0.84      0.74      0.78      6527

    accuracy                           0.80     13054
   macro avg       0.80      0.80      0.80     13054
weighted avg       0.80      0.80      0.80     13054

Model split val 3:7 :
3:7 split Accuracy: 0.6752616798570334
              precision    recall  f1-score   support

        male       0.73      0.82      0.77      6527
      female       0.79      0.69      0.74      6527

    accuracy                           0.76     13054
   macro avg       0.76      0.76      0.76     13054
weighted avg       0.76      0.76      0.76     13054
```

תשובות :

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?

1) Precision ניתן להגדיר בשאלה הבאה: כמה תוצאות שנבחרו הן רלוונטיות? כלומר זה כל המסמכים שהמסווג אומר עליהם כן ובודק איזה חלק צדק מתוך כל המסמכים שאמר עליהם כן.
לעומת זאת recall ניתן להגדיר בשאלה הבאה: כמה תוצאות רלוונטיות נבחרו? כלומר איזה אחוז מהווים המסמכים שהמסווג אומר עליהם כן מתוך אלו שבאמת כן. כמובן שאנחנו יודעים עליהם שהם כבר חיובים(כל המסמכים שהמסווג אמר כן וצדק מתוך אלו שבאמת כן).
לכן ניתן להסיק כי אחוז ה recall אצל נשים יהיה גבוה יותר מאחוז הגברים. כמו כן הסקה זאת נובעת ישירות מכמות הכותבות אל מול הכותבים.. כאמור ישנם הרבה יותר כותבים ולכן כמעט כל מה שנבחר אצל נשים הוא רלוונטי לנו. לכן נקבל recall יותר גבוה.
בנוסף נבחין כי קל לזהות יותר כתיבת נשים אל מול כתיבת גברים ועל כך מעיד ה recall .

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

2) למה התוצאות בין שיטת האימון של פיצול ל cross validation לפעמים אינם דומות :
א) שימוש בכל הדאטה שלנו : כאשר יש לנו מעט מאוד נתונים, פיצולם לאימון ולסט מבחנים עשוי להשאיר אותנו עם ערכת מבחנים קטנה מאוד. לעומת זאת אם נשתמש ב cross validation אנו בונים K מודלים שונים, כך שנוכל לבצע תחזיות על כל הנתונים שלנו. וכך נמנע מ overfitting .
ב) כאשר אנו יוצרים חמישה מודלים שונים באמצעות אלגוריתם הלמידה שלנו ובודקים אותו בחמש ערכות מבחן שונות, אנו יכולים להיות בטוחים יותר בביצועי האלגוריתם שלנו. כאשר אנו מבצעים הערכה יחידה במערך המבחנים שלנו, אנו מקבלים רק תוצאה אחת. תוצאה זו עשויה להיות בגלל מקרה או מערך מבחן מוטה (bias) מסיבה כלשהי.
ג) כאשר אנו מבצעים פיצול אקראי של מבחן הנתונים שלנו, אנו מניחים שהדוגמאות שלנו אינן תלויות. זה אומר שידע/ראייה של מופע כלשהו לא יעזור לנו להבין מופעים אחרים. עם זאת, זה לא תמיד המקרה.
ד) והסיבה אולי הכי חשובה לביצוע cross validation היא קבלת פרמטרים אופטימליים תוך גדי הימנעות מ overfitting . בחלוקה של הקבוצות. במקרה שלנו 10 קבוצות אנחנו יכולים לבדוק על איזה קבוצה מקבלים את סט הפרמטרים הכי טובים. בנוסף מאוד נפוץ היום לעשות cross validation על הפרמטרים עצמם.

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

3) למעשה המודל שמתבסס על BOW קיבל אצלי אחוז דיוק גבוה יותר בגלל שהווקטור עצמו הוא וקטור תדירויות של המילים כאשר כל מילה מקבלת משקל לפי התדירות שלה, כלומר מילה שיש לה תדירות גבוה תקבל משקל גבוה .
לווקטור שאני יצרתי אומנם גם עשיתי וקטור תדירויות אך נתתי משקלים ביחס לתדירויות של כל מילה ולא ביחס לכל שאר המילים כמו ש TfidfVectorizer עושה.

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

4) כן. נשים לב שזה תלוי שפה. לדוגמא בעברית היינו מעדיפים להשתמש במילות דקדוק כי אפשר להבחין בין המין המיועד לפי מילות הדקדוק אך באנגלית נשים לב שהיינו מעדיפים להשתמש במילות תוכן כי במילות דקדוק אי אפשר לזהות אם זה גבר או אישה.

יכול מאוד להיות שישנם שפות שגם מילות תוכן וגם מילות דקדוק אינם עוזרות לקבלת ההחלטה. ולכן התשובה היא תלוית שפה.