עיבוד שפות טבעיות

תרגיל בית 3

אורון ורנר 300881612

התרגיל הוכן תחת מערכת ההפעלה WINDOWS 10, וגרסת python 3.8.

רקע:

מטרת התרגיל הינה סיווג טקסטים שונים, בעזרת חבילת scikit-learn. בתרגיל זה השתמשתי בשני מסווגים, Naïve Bayes ו-Logistic Regression, ביצעתי סיווג בעזרת בחירת מאפיינים אוטומטית וידנית, וחקרתי את התוצאות שהתקבלו.

גישה כללית:

תרגיל זה התבסס על לימוד עצמי של חבילת פייתון. עקב כך, היה קיים קושי רב במימוש הפעולות אותם רציתי לבצע, שכן הפער בין המחשבה והכיוון התיאורטי לבין המימוש בפועל היה ניכר.
עבדתי בשלבים קטנים, כשבכל פעם הוספתי נדבך ממטרות התרגיל לקוד.
הושם דגש על בנייה נכונה והבנה של קבצי המקור ותוכנם, שכן הם מהווים את הבסיס ממנו נגזרות תוצאות חישוב המסווגים.

שלב א – סיווג בעזרת Bag of Words:

ראשית נבצע סיווג עבור שני הכותבים בעלי כמות המשפטים הגדולה ביותר בארגנטינה.
עבורי, שני המשתמשים בעלי כמות המשתמשים הגדולה ביותר הינם: Wardenclyffe56 ו-Wild_Marker.
כמות המשפטים בינם השוותה לצורך ביצוע המשך התהליך.

בשימוש ב-CountVectorizer, כמות המילים עבור Wardenclyffe56 הינה: 21644, וכמות המילים עבור Wild_Marker הינה: 22809. כלומר כמות יחסית זהה, ולאחר חישוב כמות המילים הכללית לקורפוס, מתקבל הערך: 33432. כלומר ניתן לראות שכמות המילים המשותפת גדלה רק פי כ-1.5, ועל כן מילים רבות משותפות לשני הכותבים.

לכל שורה בטקסט ניתן סיווג בהתאם למחלקה אליה היא שייכת (דובר א' או ב').
בוצע ten-folds cross validation, הכולל גם ערבוב של השורות מהקורפוס הכללי (על מנת למנוע יצירה של "מבנה" כלשהו בו עלולים לקבל מספר רב של פעמים מצב בו המשפטים הראשונים תמיד שייכים לדובר א' והמשפטים הסופיים תמיד שייכים לדובר ב') ולאחר מכן נבדקו תוצאות הסיווג בין שני המסווגים:

בשימוש במסווג Naïve Bayes, התקבלה תוצאת דיוק ממוצע של 79.83%
בשימוש במסווג Logistic Regression, התקבלה תוצאת דיוק ממוצע של 79.03%

ניתן לראות כי תוצאות המסווגים עבור 2 הדוברים זהות כמעט לחלוטין.
התוצאה הנ"ל תואמת את המחשבה הראשונית בה הערכתי שהתוצאות בין המסווגים יהיו יחסית דומות בעקבות כמות המידע הדלה באופן יחסי על שני הדוברים. על כן, קשה לצפות שלמסווג אחד יהיה יתרון על האחר.
סביר שתוצאות הדיוק עצמן (כ-80%) מתאפשרות בעקבות סגנון כתיבה החוזר על עצמו עבור כל משתמש, ובכך בעזרת יצירת feature vectors מתאימים, מצליחים המסווגים להתאים את המשפט לדובר באופן יחסית טוב.

כעת נבצע סיווג עבור משתמשים שונים מ-5 ארצות שונות.
עבור כל מדינה, השתמשתי בקבצי המקור עבור המשתמשים שכתבו הכי הרבה משפטים, בדומה לתרגיל 2.
בכל קורפוס שכזה, ערבבתי את המשפטים על מנת שרצף המשפטים יהיה שייך למשתמשים שונים (על מנת להימנע מ"בלוק" משפטים ששייך למשתמש א', ואחריו "בלוק" השייך למשתמש ב' וכן הלאה).

לאחר הערבוב, כל 20 משפטים אוחדו לכדי משפט בודד, בהתאם להוראות התרגיל.
הנ"ל בוצע עבור כל קורפוס השייך לכל אחת מחמשת המדינות.

כלל הקורפוסים הושוו מבחינת כמות המשפטים, לכל שורה בטקסט ניתן סיווג בהתאם למחלקה אליה היא שייכת (0-4, בהתאם למדינת המקור), והם אוחדו לכדי קורפוס אחד גדול.
בשימוש ב-CountVectorizer, כמות המילים השונות שהתקבלו עבור הקורפוס הכולל הינן 107,603.

גם כאן, בדומה לבדיקת הדוברים, בוצע ten-folds cross validation, הכולל גם ערבוב של השורות מהקורפוס הכללי (על מנת למנוע יצירה של "מבנה" כלשהו בו עלולים לקבל מספר רב של פעמים מצב בו המשפטים הראשונים תמיד שייכים למדינה א' והמשפטים הסופיים תמיד שייכים למדינה ב') ולאחר מכן נבדקו תוצאות הסיווג בין שני המסווגים:

בשימוש במסווג Naïve Bayes, התקבלה תוצאת דיוק ממוצע של 97.68%
בשימוש במסווג Logistic Regression, התקבלה תוצאת דיוק ממוצע של 95.42%

ניתן לראות כי אחוז ההצלחה גבוה מאוד והנ"ל הפתיע אותי בתחילה.
על מנת לוודא את נכונות התוצאות, ביצעתי מספר פעולות אשר היוו "בדיקת שפיות" עבורי (sanity check):

1. הרצתי את הבדיקות הנ"ל מספר פעמים, על מנת לוודא שהתוצאות חוזרות על עצמן. יש להזכיר כי קבוצות ה-Test וה-Train מוגרלות באופן אקראי בכל פעם עבור עשרת מקרי הבדיקה, ועל כן חזרתיות בתוצאות מוכיחה כי הנ"ל לא התקבל במקרה.
2. ניתחתי את התפלגות קבוצות ה-Test ווידאתי שלא קיימת הטיה עבור מדינות מסוימות. אכן, בקבוצת הבדיקה קיימת התפלגות יחסית זהה בין כמות המשפטים עבור כל אחת מהמדינות.
3. שאלתי בפורום הקורס 😊. רק כדי לוודא שאני לא מפספס משהו גדול נוסף.

כעת, משהגעתי למסקנה כי התוצאה הגיונית, ניתן להניח בניתוח אוחר כי כמות המידע הגדולה אותה אנו מחזיקים מאפשרת אימון מספיק חזק על מנת שתוצאותיו במהלך הבדיקה יהיו גבוהות מאוד.
ניתן לשער, כי במדינות שונות אשר כותבות בשפה שאינה שפת האם שלהם, קיימות טעויות זהות או ביטויים/שגיאות החוזרים על עצמם. על כן, בעת ניתוח משפטי הבדיקה, קל יחסית לזהות את סגנון הכתיבה ואת מקור המדינה אליה הוא משתייך.

עבור המסווגים, Naïve Bayes מניח שה- features הינם בלתי תלויים במהלך עבודתו. כיוון שהתוצאות המתקבלות עבורו מעט גבוהות יותר, אני מניח שהצלחתו נובעת מ-features במדינות שאכן אינם תלויים ב-features אחרים, למשל שגיאות כתיב, מילים נפוצות, טעות נפוצה וכו' ופחות במבנה תחבירי שלם או שימוש בסימנים מרובים (למשל שימוש שגוי ב-'s, 'll, 've), שכן אלו תלויים במילים לפניהם.
כלומר, לא מתבצע ניתוח trigram שיכול להסיק שקיימת צורה נפוצה המכילה ve ' word שכן מדובר ב-3 טוקנים שונים והמסווג מניח אי תלות.

שלב ב – סיווג בעזרת בחירת מאפיינים באופן ידני:

עבור בחירת המאפיינים הידניים, ניסיתי לחשוב על מאפיינים אשר עשויים להיות ייחודיים ומשפיעים הן עבור המשתמש הבודד, והן עבור מדינות שונות. כמובן, מכיוון ששלב א' ו-ג' עוסקים במילים, ניסיתי להימנע ככל האפשר מלחזור על השימוש במאפיינים אלה. זאת על מנת לצפות בשינוי בתוצאות וכן להבין את השפעתם של מאפיינים "רכים" יותר.

ראשית, בחרתי להתייחס לשני פרמטרים אשר לטעמי עשויים להציג הבדלים עבור שני הקורפוסים שלנו – אורך משפט, וכמות המילים במשפט.
לדעתי, מאפיינים אלה עשויים לעזור יחסית בקלות להציג הבדלים בין משתמשים ומדינות.
הרצת פעולות שני המסווגים על שני הקורפוסים, אכן איששה את ההנחה:
עבור זיהוי הדובר התקבל משני המסווגים אחוז הצלחה ממוצע של כ-58%.
עבור זיהוי מדינת התקבל משני המסווגים אחוז הצלחה ממוצע של כ-30%.

אמנם הדבר אינו נראה מרשים במיוחד, אך יש לזכור כי מדובר ב-feature vector המכיל **2 פיצ'רים בלבד.**
כלומר, איכותם בהחלט מגיעה לידי ביטוי.

עם זאת, ברור שניתן להשתפר ולכן עברתי לבדוק שימוש בסימני פיסוק וסמלים. בכתיבה "אינטרנטית" ידוע שקיים שימוש רב ולעיתים מוגזם בסימני פיסוק או סמלים, ואופן השימוש בהם עשוי להצביע על דובר או מדינה מסוימים.
בהוספת 7 סימני פיסוק וסמלים נפוצים, אחוז ההצלחה עבור זיהוי הדובר נותר כשהיה, אולם עבור זיהוי המדינה על אחוז ההצלחה לכ-35%.
ניתן להניח כי בעבור כמות המידע הגדולה אשר אנו מחזיקים בקורפוס המדינות – בדיקת השימוש בסמלים וסימני פיסוק אכן עזרה במקצת, אולם עבור זיהוי הדוברים, רוב המשפטים אינם מורכבים מסימני פיסוק או סמלים (אלא ממילים) ועל כן תרומתם זניחה.
לאחר בדיקת סימני פיסוק וסמלים שונים, החלטתי לזנוח את רובם מלבד סימן הגרש, שכן תרומתם כמעט ולא הורגשה.

כעת החלטתי לנסות שילובים בסיסיים של מילים באנגלית (כיוון שאנו לא בודקים n-grams). ראשית בדקתי את הרצף 'I am' אשר הצליח באופן מפתיע לייצר שיפור עבור זיהוי המדינה והעלה את אחוז ההצלחה לכ-38%.
הוספת הרצף 'you're' עזר גם הוא מעט, אולם התוספת המפתיעה הגיע בעזרת הרצף '...' אשר העלה את אחוז ההצלחה בזיהוי המדינה לכ-40% תחת המסווג naïve bayes ולכ-42% תחת LR.

הוספת רצפי סימנים נוספים כמו '!!!' או '???' לא הפיקו תועלת.
כמו כן, רצף הסימנים גם לא עזר עבור זיהוי הדובר ולכן אחוז ההצלחה במבחנים אלו נשאר על כ-57-58%.

לסיום ניסיתי לחשוב על תווים או מילים שיכולים לעזור בשימוש הזיהוי בין משתמשים ובין מדינות וניסיתי שילובים כמו I'm, 's, 're, 'll, like וכן הלאה, אולם אלו לא כמעט ולא עזרו בזיהוי הדובר או בזיהוי המדינה.

לבסוף, צמצמתי את כמות המאפיינים עבור מאפיינים שתרומתם הייתה אפסית או זניחה, והשארתי רק את המאפיינים שהשפעתם הייתה ניתנת לזיהוי.

עבור מאפיינים אלו התקבלו התוצאות הבאות:

עבור זיהוי הדובר:

בשימוש במסווג Naïve Bayes, התקבלה תוצאת דיוק ממוצע של 57.08%
בשימוש במסווג Logistic Regression, התקבלה תוצאת דיוק ממוצע של 58.62%

עבור זיהוי מדינת המקור:

בשימוש במסווג Naïve Bayes, התקבלה תוצאת דיוק ממוצע של 40.30%
בשימוש במסווג Logistic Regression, התקבלה תוצאת דיוק ממוצע של 42.31%

שלב ג – סיווג בעזרת בחירת המילים המשמעותיות ביותר:

להלן רשימת מאה המילים המשמעותיות ביותר לזיהוי בין דוברים:

```
[['ai' 'also' 'armor' 'army' 'battle']
 ['because' 'big' 'bonus' 'but' 'care']
 ['china' 'combat' 'cost' 'de' 'denuvo']
 ['divisions' 'dlc' 'don' 'due' 'eu4']
 ['france' 'fuck' 'game' 'gems' 'germany']
 ['god' 'gods' 'gonna' 'he' 'her']
 ['him' 'his' 'hoi3' 'hoi4' 'idk']
 ['imagine' 'infantry' 'instead' 'into' 'its']
 ['japan' 'know' 'life' 'like' 'll']
 ['lol' 'lot' 'me' 'men' 'might']
 ['my' 'myself' 'naval' 'nobody' 'not']
 ['ok' 'op' 'or' 'paradox' 'people']
 ['person' 'planes' 'plus' 'post' 'probably']
 ['province' 'provinces' 'que' 'regular' 'she']
 ['ships' 'skin' 'smite' 'solo' 'someone']
 ['sorry' 'stuff' 'stupid' 'sub' 'system']
 ['teammates' 'tech' 'the' 'they' 'thing']
 ['this' 'tho' 'trade' 'troops' 'uk']
 ['ult' 'unit' 'units' 'usually' 'video']
 ['want' 'war' 'white' 'women' 'ymir']]
```

בניתוח המילים הנ"ל, קשה לראות דפוס מסוים, אולם ניתן לשים לב כי חלק מהמילים קשורות למשחקי מחשב ומחשבים באופן כזה או אחר. כמו כן, חלק מהמילים קשורות לשיח וכתיבה ברשת.
על כן ייתכן שמילים אלו נפוצות יותר אצל דוברים מסוימים, ומכאן שעוזרות בזיהוי הדובר.

להלן רשימת מאה המילים המשמעותיות ביותר לזיהוי מדינות המקור:

```
[['aap' 'ai' 'albania' 'albanian' 'albanians']
 ['and' 'anime' 'are' 'argentina' 'argentine']
 ['army' 'as' 'atx' 'baccano' 'bjp']
 ['but' 'by' 'cannot' 'cant' 'chakra']
 ['chromecast' 'cpu' 'cypriot' 'cypriots' 'cyprus']
 ['de' 'didnt' 'doesnt' 'dont' 'durarara']
 ['el' 'es' 'eu' 'firo' 'game']
 ['games' 'georgia' 'georgian' 'gpu' 'greece']
 ['greek' 'he' 'idk' 'im' 'in']
 ['india' 'indian' 'isaac' 'islam' 'island']
 ['it' 'itachi' 'kakashi' 'konoha' 'la']
 ['ladd' 'lol' 'malta' 'maltese' 'mate']
 ['minato' 'modi' 'ms' 'muhammad' 'muslim']
 ['muslims' 'my' 'narita' 'naruto' 'novels']
 ['obito' 'of' 'okay' 'psu' 'que']
 ['ram' 're' 'religion' 'russia' 'ryzen']
 ['sakura' 'sasuke' 'seattle' 'sharingan' 'shit']
 ['spanish' 'taiwan' 'thanks' 'thats' 'the']
 ['their' 'them' 'to' 'turkey' 'turkish']
 ['uchiha' 'ukraine' 'us' 'which' 'yeah']]
```

בניתוח המילים מקורפוס המדינות, ניתן לראות הבדל מהותי – כמות מילים רבה שייכת למדינות או מקומות. ייתכן שדוברים מאותה מדינה נוהגים לציין את המדינה שלהם או עיר מסוימת במשפט, כחלק משיחה (אולי כמתן דוגמא "במדינה שלי, ***, אנחנו בדר"כ נוהגים...") או שפשוט רוצים להזכיר או להביע עמדה הקשורה למקום מסוים.
בנוסף, ניתן לזהות אלמנטים דתיים – איסלאם, מוחמד, מוסלמים. גם כאן, ייתכן כי דעות רבות (קדומות או נוכחיות) משלבות את האלמנטים הנ"ל בתוך משפטי השיחה הרגילים ועל כן המילים הללו עוזרות לזהות את מדינת המקור.

על אף שהמאפיינים שבחרתי בשלב ב' התמקדו פחות במילים ספציפיות - בשני המקרים, המילים שנבחרו בשלב זה הפתיעו אותי ב"חשיבותם" (כפי שחושבה ע"י הקוד), והם אינם דומים למאפיינים אותם בחרתי או שהייתי חושב לתת להם משקל רב באופן יחסי.

להלן תוצאות המסווגים שהתקבלו בעזרת השימוש במאה המילים המשמעותיות ביותר:

עבור זיהוי הדובר התקבל אחוז הצלחה ממוצע של כ-62.72% עבור מסווג Naïve Bayes
וכ-63.50% עבור מסווג Logistic Regression.

עבור זיהוי מדינת המקור, התקבל אחוז הצלחה ממוצע של כ-49.67% עבור מסווג Naïve Bayes
וכ-55.33% עבור מסווג Logistic Regression.