דו"ח תרגיל 3 – עיבוד שפות טבעיות

מגיש: שגיא גוילי, ת.ז. 203638804

בתרגיל זה התבקשתי לבצע סיווג טקסטים לשתי קטגוריות- אנגלית ואנגלית פשוטה. לצורך כך קיבלתי קבצי טקסט, ששימשו כקלט לתוכנית, שהכילו ערכים מויקיפדיה כאשר לכל ערך יש קובץ טקסט בשפה האנגלית ובשפה האנגלית הפשוטה. יחידת הסיווג הייתה כל משפט בטקסט, לכן בתחילה חילקתי את כל הטקסטים לשני קורפוסים שווי גודל, כלומר אותו מספר משפטים- לפי הדרישה שהייתה בהוראות- אם ערך אחד היה בעל יותר משפטים היינו מדללים אותו ובוחרים ממנו משפטים אקראיים. לבסוף איחדתי בין הקורפוסים וקיבלתי data set של משפטים מעורבבים משתי השפות.

בנוסף, הכנתי כבר מראש את הוקטור המחלקות לכל אחת מיחידות הסיווג, כלומר וקטור באורך ה-data set כאשר כל ערך הוא 0 אם המשפט הוא מאנגלית פשוטה ו-1 אם הוא מאנגלית.

שלב א'

בשלב הראשון ביצעתי סיווג על ידי שימוש ב-Bag of words. השתמשתי בספריית ה-scikit learn על מנת להפוך את הטקסטים ל-feature_vectors (שימוש ב-CountVectorized). מספר המילים השונות שקיבלתי בקורפוס האנגלית היה – 18403 ובאנגלית הפשוטה 11318 (מספר זה השתנה בכל הרצה כמובן שכן הייתי מגריל משפטים שונים בכל הרצה ומספר המילים בכל משפט הוא שונה, אך באופן כללי המספר נע סביב מספרים אלה)

לאחר מכן השתמשתי במתודות של הספרייה scikit learn על מנת לבצע את הסיווג על ידי כל אחד מהמסווגים; בכל מסווג ראשית חישבתי את ה-tf-idf של כל מילה על ידי TfidfTransformer ולאחר מכן השתמשתי באלגוריתם הסיווג המבוקש שגם לו היה מימוש ב-scikit learn (אימון על ה-feature_vectors ווקטור המחלקות)

לבסוף השתמשתי ב-cross_val_score על מנת לבצע ten-fold-cross-validation שיעריך את הביצועים של כל מסווג, ולהלן תוצאות ה-Accuracy:

Naïve Bayes: 69.19

KNN: 50.47

Logistic Regression :70.50

ניתן לראות שהסיווג נעשה בצורה יחסית טובה, בעיקר ב-Bayes ו-Logistic Regression , מבחינת ההשוואה לציפיות:

- מודל Naïve Bayes מתאים בדיוק לבעיה שלנו- סיווג בינארי, האם כל יחידת סיווג שייכת למחלקה או לא, במקרה שלנו בדיקה של האם שייכת לאנגלית פשוטה כן או לא, האם שייכת לאנגלית רגילה כן או לא. מפני שזה סיווג בינארי הרי שנוח היה להניח שסיכויי ההצלחה הם 50%, אבל התוצאות היו קצת אחרות ואף מפתיעות, כמעט 70% הצלחה מה שמלמד אותנו שאכן היה הבדל לא קטן בין השפה האנגלית הפשוטה לאנגלית הרגילה. בנוסף לכך, כפי שראינו בהרצאה מודל Naïve Bayes הוא אמנם פשוט אך לרוב מספק דיוק ברמה גבוהה למדי מה שעוד יותר מסביר את התוצאות שהתקבלו.
- במודל ה-KNN התוצאות היו הנמוכות ביותר בשלב זה, הדבר לא היה מפתיע מדי לטעמי, שכן במסווג זה תכונות לא רלוונטיות (במקרה שלנו תכונות שדומות מאוד בשתי השפות) יפגעו באופטימליות המסווג- כלומר יפגעו בדיוקים שלו, ואכן (כפי שראה בשלב הבחירה הידנית של תכונות) קיימות תכונות רבות שהן "לא טובות" שמאוד דומות בשתי השפות.
- במודל ה-Logistic Regression הציפייה שלי הייתה דומה מאוד למודל Naïve Bayes (בגלל הדימיון הרב בין השניים), וכן התוצאות בסופו של דבר היו מאוד דומות (סביב ה-70%).

שלב ב'

בשלב זה ביצעתי סיווג על ידי שימוש ב-300 המילים הנפוצות באנגלית, על ידי שימוש בקובץ הקלט שצורף לתרגיל. התהליך ברובו היה דומה לשלב א', רק שהפעם לא היה לי צורך להשתמש ב-CountVectorized שכן בניתי את ה-feature_vectors בעצמי- עברתי על כל יחידת סיווג בקורפוס המאוחד (כלומר כל משפט) והכנתי לו וקטור באורך 300 שבכל תא יש ערך 1 אם המילה ה-i בקלט המילים הנפוצות נמצא במשפט ו-0 אחרת.

וקטור המחלקות (ה-1ים ו-0ים) התאים לוקטור שבניתי ואימנתי את ה-feature_vectors באותה הצורה בדיוק כמו בשלב א'.

להלן תוצאות ה-Accuracy:

Naïve Bayes:62.51

KNN:57.42

Logistic Regression:65.93

ניתן לראות שהסיווג הניב תוצאות פחות מדוייקות במודל Naïve Bayes ובמודל ה- Logistic Regression אבל יותר טוב ב-KNN.

- במודל Naïve Bayes התוצאות אמנם נמוכות יותר מאשר מה שראינו בשלב הקודם אבל עדיין גבוהות יחסית לעומת מה שהייתי מצפה, אני מניח שהשימוש במילים בלבד כתכונות אימון הן אלה שפגעו בתוצאות בסופו של דבר שכן ההבדלים בין שתי המחלקות נבעו יותר ממרכיבים אחרים (שעליהם ארחיב בשלב ג').
- במודל הLogistic Regression גם ניתן להבחין ירידה קלה בדיוק, אני מניח ששוב הסיבה לכך טמונה בשימוש במילים בלבד ללא תכונות סגנון אחרות שמבליטות יותר את ההבדלים בין שתי המחלקות.
- במודל ה-KNN, לעומת שני המודלים האחרים, ראינו עלייה בדיוק והדבר לא הפתיע יותר מדי שכן, כמו שכתבתי בשלב הקודם, הדיוק הנמוך שקיבלנו במודל זה נבע ככל הנראה בגלל ריבוי התכונות הלא רלוונטיות שעלו מהשימוש ב-Bag of words, לעומת זאת כאן השתמשנו בתכונה של נפוצות מילים במחלקה אחת לעומת נפוצות המילים במחלקה שנייה, ולמרות שהתוצאות עדיין נמוכות יחסית לשאר המודלים, ככל הנראה 300 המילים היו קצת יותר רלוונטיות ברובן- במיוחד בהשוואה לתכונות הרבה פחות רלוונטיות שעלו מה- Bag of words.

שלב ג'

בשלב זה ביצעתי סיווג על ידי שימוש במאפיינים שבחרתי באופן ידני, בחרתי סך הכל 18 מאפיינים שמתוכן 5 מאפייני תוכן ו-13 מאפייני סגנון. אפרט כעת את המאפיינים אותם בחרתי, אפרט את הסיבה שבחרתי בכל מאפיין ולבסוף אציין את ה-score שקיבל כל מאפיין על ידי השימוש ב-SelectKBest מה שבודק בעצם מכמת למספר את כמות העזרה של כל מאפיין בדיוק הסיווג- כמה שיותר גבוה ככה יותר הועיל בסיווג (ה-scores משתנים בכל ריצה של התוכנית, התוצאות שאציג כאן יהיו בריצה ספציפית אבל ייצגו בצורה מהימנה את ההשפעה על דיוק הסיווג שכן מנעד ה-scores בכל אחת מהריצות היה קטן מאוד)

אתחיל ממאפייני הסגנון:

1. **אורך המשפט-** יחידת הסיווג שלי היא משפט, לאחר בחינה של יחידות הסיווג בכל אחד מהקורפוסים הבחנתי שאורכי המשפטים באנגלית הרגילה ארוכים בהרבה מהמשפטים באנגלית הפשוטה- לכן זהו המאפיין הראשון שראיתי לנכון לסווג לפיו שכן ברוב המקרים משפטים קצרים הסתברות גבוהה יהיו שייכים לאנגלית הפשוטה ומשפטים ארוכים לאנגלית הרגילה. ה-score שקיבלתי היה בהתאם- 1002.369, מה שמראה שהבחירה הייתה טובה ושיפרה את דיוק הסיווג מאוד.
2. **אורך מילה ממוצע-** קורפוס האנגלית הפשוטה יכיל ברובו משפטים באנגלית יותר קלה ובהירה לקורא ולכן תמעיט במילים גבוהות ומסובכות יותר. למילים גבוהות בשפה האנגלית יש נטייה להיות ארוכות יותר (pedestrian, pavement, appellation ועוד..) לכן ראיתי לנכון לחשב עבור כל משפט את אורך המילה הממוצעת בו כך שמשפטים עם אורך מילה ממוצע קטן יסווגו לאנגלית הפשוטה ועם אורך ממוצע גדול לאנגלית הרגילה. ה-score שקיבלתי היה בהתאם- 741.948, ואכן מאפיין זה שיפר את דיוק הסיווג.
3. **מספר הפסיקים במשפט-** פסיק מפריד לי בין חלקים שונים במשפט, בתחילה הנחתי באופן טבעי שמספר הפסיקים במשפטים ארוכים יותר יהיו רבים יותר (יותר חלקים במשפט ולכן יותר פסיקים) אך אחרי מעבר על הערכים בשתי השפות הופתעתי לגלות שדווקא באנגלית הפשוטה קיימים פסיקים רבים בכל משפט על אף אורכם הקצר לרוב, לעיתים אף יותר מאשר מכמות הפסיקים במשפטים הארוכים באנגלית הרגילה- הסיבה לכך טמונה ככל הנראה בהעדפה שיש למילות קישור באנגלית הרגילה כמעבר בין חלק לחלק יותר מאשר בפסיקים. בכל אופן ראיתי לנכון לנסות להשתמש במאפיין זה ואכן ה-score היה 525.465 מה שמאשש את היות מספר הפסיקים למאפיין משמעותי, גם אם הוא מופיע יותר באנגלית הפשוטה וגם אם ההפך.
4. **כמות המספרים במשפט-** תדירות הופעת מספרים במשפט ממוצע בערכי ויקיפדיה היא לא גבוהה במיוחד- רוב הערכים אינם מכילים נוסחאות ותאריכים, במקסימום תאריכי לידה וגילאים או נקודות ציון משמעותיות בהיסטוריה של ערך מסוים. באופן טבעי הנטייה שלי לחשוב שכאשר משפט הוא ארוך יותר הסבירות להימצאות המספרים בו גבוהה יותר ולכן במשפטים של האנגלית הפשוטה יהיו פחות מספרים ולכן מאפיין זה ידייק לי יותר את הסיווג. ה-score שקיבל המאפיין הוא 113.264, מה שמראה שהכיוון מחשבה שלי היה יחסית נכון, אמנם לא כמו מאפיינים קודמים אבל בהחלט תורם לא מעט לדיוק הסיווג.
5. **מספר המילים שנגמרות ב-ing-** קורפוס האנגלית הפשוטה מאופיין במשלב לשוני נמוך, כלומר השפה לא עשירה והדקדוק לא גבוה מדי. באנגלית השימוש בסיומת ing הוא נפוץ מאוד בעיקר בזמני Present Progressive, Past progressive ולפעמים אף ב-Perfect. לכן חשבתי לנכון להשתמש בעובדה זו כמאפיין ולבדוק כמה מילים בכל משפט נגמרות ב-ing כאשר הרציונל שלי היה שמשפטים בהם יש יותר מילים שנגמרות ב-ing יסווגו לאנגלית הרגילה. למרבה ההפתעה ה-score היה 375.816, לא ציפיתי ל-outcome טוב כלכך ממאפיין כזה, שמאפיינים אחרים יהיו משמעותיים הרבה יותר, לבסוף מאפיין זה התברר כמועיל במיוחד.
6. **מספר המילים השונות במשפט-** כפי שכתבתי במאפיינים קודמים, האנגלית הפשוטה מאופיינת במשלב לשוני נמוך, אוצר מילים קטן יחסית לאנגלית הרגילה ודקדוק לא גבוה

במיוחד. משום כך חשבתי שבקורפוס זה מספר המילים השונות בכל משפט יהיה נמוך יחסית (הרבה שימושים ב-because,a,is וכדומה) מה שיבדיל אותם מהמשפטים באנגלית הרגילה בהן אוצר המילים הוא עשיר ומגוון יותר וכן המשפטים עצמם ארוכים יותר. להפתעתי מאפיין זה התגלה כפחות תורם לדיוק המסווג, עם score של 92.466- ככל הנראה מפני שבשונה ממה שחשבתי, למרות השוני באוצר המילים והאורכים השונים של המשפטים בשני הקורפוסים- עדיין השוני בין כל מילה במשפט הוא זהה יחסית בשני הקורפוסים.

7. **<u>מספר הפותחים והסוגרים-</u>** במהלך המעבר על הערכים השונים בשתי השפות שמתי לב לנפוצות קצת יותר גדולה של פותחים וסוגרים באנגלית הרגילה, לא הצלחתי כלכך להבין למה הדבר קורה אבל לדעתי הרבה פעמים בין פותחים וסגרים נוטים להוסיף מידע, לא תמיד חשוב מדי, על כל מיני נושאים שונים שדוברו עליהם בחלק האחרון של המשפט. באנגלית הפשוטה שמתי לב שנמנעים מכך יותר ויש יותר דבקות בעיקר בלי להרחיב יותר מדי, לעומת זאת באנגלית הרגילה יש יותר הרחבות מהסוג הזה וככל הנראה מפה נובע השוני. הפותחים (מכל הסוגים; מרובעים, עגולים ומסולסלים) קיבלו score של 9.482 והסוגרים score של 14.269. הופתעתי לגלות שהדברים לא בדיוק כפי שציפיתי ובסופו של דבר אחוז הסוגרים ופותחים בכל אחד מהקורפוסים כנראה די שווה.

במאפיינים הבאים התחלתי לבדוק נפוצות של מילות קישור במשפטים של כל אחד מהקורפוסים:

- and – נפוצה מאוד בשפה האנגלית, מקום רביעי במילים הנפוצות בשפה האנגלית, עם זאת היא מילה במשלב לשוני די נמוך וכאשר האוצר מילים קטן יותר יש נטייה להשתמש בה פעמים רבות יותר- לכן המחשבה שלי הייתה שהנפוצות שלה באנגלית הפשוטה תהיה גדולה הרבה יותר מאשר באנגלית הרגילה והעשירה יותר באוצר המילים שלה. ה-score שהתקבל הוא 377.839, מה שאישש את מה שחשבתי ומילה זו אכן תרמה מאוד לדיוק הסיווג.
- of- המילה השנייה הנפוצה ביותר בשפה האנגלית, בדומה ל-and גם בה יש שימוש רב יותר כאשר המשלב הלשוני של הטקסט נמוך יותר, ואכן בדומה ל-and ה-score שלה היה גבוה מאוד ואף יותר, 570.274.
- the – המילה הנפוצה ביותר בשפה האנגלית, משמעותית בסיווג מאותן סיבות כמו של of ו-and ואכן קיבלה score גבוה של 481.421.
- in- המילה ה-6 בנפוצות שלה בשפה האנגלית, המחשבה שלי בתחילה הייתה שלא יהיה פער גדול מדי בנפוצות שלה בין שני הקורפוסים שכן יש בה שימוש ברוב משלבי הלשון, אך הופתעתי לגלות שה-score שלה הוא 311.696, מה שגבוה אפילו יותר מ-a שקיבלה score נמוך יותר ולכן העדפתי את in.
- with- המילה ה-16 בנפוצות שלה בשפה האנגלית. מילה זו הפתיעה אותי מאוד שכן הרציונל שלי שמילים עם הנפוצות הגבוהה ביותר תהינה נפוצות יותר באנגלית פשוטה ולכן יהיו מאפיינים טובים יותר לסיווג. מילה זו היא עם נפוצות נמוכה יחסית אך קיבלה score של 183.875 שהיה גבוה יותר מ-a ולכן חשבתי לנכון לבחור במילה זו גם כן.

מאפייני התוכן:

את בחירת מאפייני התוכן היה לי יותר קשה לאבחן; הקלט שלי הוא ערכים זהים בשתי שפות מאוד מאוד דומות והתוכנית שלי ממומשת בצורה כזאת שמספר המשפטים מאותו הערך בכל שפה הוא שווה. לכן ציפיתי שלא יהיו הבדלים גדולים מדי בין השפות בכל הקשור לתוכן כל משפט, כלומר יהיה לי קשה לסווג משפט לפי התוכן שלו כי תנאיי ההתחלה שלי מאוד דומים בשני הקורפוסים. לכן החלטתי ל"תקוף" את הבעיה בצורה שונה- השוואת מספר ההופעות של מילות תוכן שונות בין שני הקורפוסים, כלומר חיפשתי מילים בקורפוס האנגלית הפשוטה שיש להן הופעות רבות והשוואתי את מספר ההופעות שלהן בקורפוס האנגלית הרגילה- בצורה כזאת הצלחתי למצוא מילים שיבדילו לי בין שתי השפות בצורה מיטבית.

1. coffee- הופיעה בשפה האנגלית הרגילה 69 פעמים והאנגלית הפשוטה 107 פעמים, בגלל הנפוצות הרבה שלה באופן כללי וספציפית יותר באנגלית בפשוטה חשבתי לנכון להשתמש במאפיין זה. ה-score של מאפיין זה הוא 4.389 כך שלמרות הפער הלא קטן בהופעות, עדיין מאפיין זה לא עזר בדיוק ואולי אף פגע בו.
2. system- הופיעה בשפה האנגלית הרגילה 149 פעמים והאנגלית הפשוטה 50 פעמים, בגלל הנפוצות הרבה שלה באופן כללי וספציפית באנגלית הרגילה היא הופיעה פי 3 מאשר באנגלית הפשוטה, חשבתי לנכון להשתמש במאפיין זה. ה-score של מאפיין זה הוא 25.292 כך שכפי שציפיתי מאפיין זה עזר יחסית לדיוק המסווג.
3. basketball- הופיעה בשפה האנגלית הרגילה 106 פעמים והאנגלית הפשוטה 40 פעמים, בגלל הנפוצות הרבה שלה באופן כללי וספציפית באנגלית הרגילה היא הופיעה כמעט פי 3 מאשר באנגלית הפשוטה, חשבתי לנכון להשתמש במאפיין זה. ה-score של מאפיין זה הוא 23.132 כך שכפי שציפיתי מאפיין זה עזר יחסית לדיוק המסווג.
4. Leonardo- הופיעה בשפה האנגלית הרגילה 196 פעמים והאנגלית הפשוטה 125 פעמים, הפער בין מספר ההופעות בשני הקורפוסים לא היה גדול מדי אך בגלל הנפוצות הרבה שלה באופן כללי חשבתי שמאפיין כזה יוכל לעזור לי בדיוק המסווג. ה-score של מאפיין זה היה גבוה יחסית בסופו של דבר- 12.567 וזה היה די מפתיע (coffee דומה מאוד במספר ההופעות שלה בשני הקורפוסים והניבה תוצאות גרועות בהרבה). היה קשה להסביר תופעה זו אבל בסופו של דבר בחרתי להשתמש במילה זו כמאפיין. (הערה- בגלל שגיאה לא מוסברת בבדיקה של הימצאות כל מילה במשפט אז עשיתי lower לכל משפט ולכן בקוד חיפשתי leonardo עם אות קטנה)
5. people- הופיעה בשפה האנגלית הרגילה 102 פעמים והאנגלית הפשוטה 279 פעמים, בגלל הנפוצות הרבה שלה באופן כללי וספציפית באנגלית הפשוטה היא הופיעה כמעט פי 3 מאשר באנגלית הרגילה, חשבתי לנכון להשתמש במאפיין זה. ה-score של מאפיין זה הוא 70.330, הגבוה ביותר בפער משאר מאפייני התוכן ולכן הבחירה בו הייתה קלה וטובה.

לסיכום, ניתן לראות שעבור הערכי הקלט שלנו בתוכנית מאפייני הסגנון נתנו את הטון יותר מאשר ערכי התוכן- כאשר ה-score המקסימלי של מאפייני התוכן היה 70, בעוד שה-score הממוצע במאפייני הסגנון היה סביב ה400-500. מכך ניתן להכריע שההבדלים העיקריים בין שתי המחלוקות הם בעיקר סגנוניים- משפטים ארוכים יותר, משלב לשוני שונה, מילים ארוכות יותר ושאר מאפייני הסגנון האחרים. כאמור הדבר לא הפתיע במיוחד שכן לכל ערך קלט היה תואם בשפה השנייה שהכיל את אותו התוכן בתוספת של הרחבות במקסימום (וגם זה צומצם עקב השוואת כמות המשפטים מכל ערך) והתוצאות אכן תאמו לציפייה שלי.

ההחלטה שלי הייתה בסופו של דבר לאמן על סמך 15 המאפיינים הטובים ביותר (k=15) שכן בתרגיל התבקשנו למצוא לפחות 15 מאפיינים אז ראיתי לנכון לאמן על סמך 15 מאפיינים כדי שיתאים למבוקש.

לאחר ש-SelectKBest בחר את 15 המאפיינים בטובים ביותר מבין ה-18, אימנתי וסיווגתי באותה הצורה כמו בשלבים א' ו-ב'. להלן תוצאות ה-Accuracy:

Naïve Bayes:62.95

KNN:64.46

Logistic Regression:69.20

כפי שניתן לראות מודל Naïve Bayes ומודל הLogistic Regression שומרים על הדיוק של בין 60-70% כמו בשני השלבים הראשונים. אפשר להסביר זאת ככל הנראה מהסיבה שהמאפיינים שבחרתי מאוד דומים למאפיינים שבהם השתמשו גם בשני השלבים הראשונים, בעיקר בשלב השני בו היה שימוש אך ורק במילים ופחות בסגנון של הטקסט.

את ההבדל העיקרי והמשמעותי ביותר ניתן לראות במודל ה-KNN, שעלה מטווח ה-50-60% לטווח ה-60-70%. ההבדל נובע ככל הנראה משיטת החישוב במודל זה: כפי שציינתי בשלבים הקודמים, מודל זה מאוד מושפע ממאפיינים פחות טובים (כלומר שלא תורמים לשיפור הדיוק) ולכן הדיוק שלו

ירד מאוד אם הוא יאומן בהרבה מאוד מאפיינים שכאלה. במקרה של בחירת המאפיינים הידנית אני מקצה מספר מאוד מוגבל של מאפיינים, 15, שגם מתוכם עשינו כבר ברירה של מאפיינים שתורמים מאוד לדיוק (אחרי בדיקת ה-score של כל אחד). מהסיבה הזו ככל הנראה אחוז הדיוק במודל זה עלה ביחס לשלבים הקודמים.