תרגיל בית 3 – סיווג טקסטים

חיים שאללשוילי 200832780

צחי כפיר 200681476

יש להשתמש בקובץ words.txt ולשים אותו בתיקיית ה-input. קובץ זה מכיל רשימת מילים המרכיבות את שק המילים שבו המסווג ישתמש בכדי לייצג את הביקורות. שק המילים הנ"ל הוא שק המילים המניב את התוצאות הטובות ביותר ומתואר בסעיף 3.

# 1.

תחילה בחרנו אוסף מילים שנראו מועילות באופן ידני והערכנו את ביצועי המסווג בשיטת ten-fold-cross-validation תוך כדי בחינת המדדים precision, recall, balanced f-score ו- accuracy.

הגדרנו את המטלה כמטלת סיווג של ביקורות חיוביות.

השווינו את תוצאות המדדים לביצועי המסווג המייצג מילים על פי bagOfWords של **כל** המילים בקורפוס. למרות שרשימת המילים הנ"ל מוסיפה רעש רב לתוצאות. תוצאות המדדים יצאו גבוהות משמעותית מתוצאות ההרצה עם שק המילים שבחרנו באופן ידני.

תוצאות המדדים עבור שק המילים המכיל את כל המילים בקורפוס:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **/** | **Precision** | **recall** | **accuracy** | **fscore** |
| **all the words**  **(size = 50620)** | 0.8798 | 0.6839 | 0.795 | 0.7691 |

לכן החלטנו לבחור רשימת מילים באופן אוטומטי שתשפר את תוצאות המדדים הנ"ל ותסווג בצורה טובה יותר ביקורות חדשים כפי שנראה בהמשך הדו"ח.

שק המילים שנבחר הוא איחוד כל המילים ה:

* המופיעות אך ורק בביקורות החיוביות ומתפרסות על יותר משתי ביקורות
* המופיעות אך ורק בביקורת השליליות ומתפרסות על יותר משתי ביקורות
* המופיעות גם בביקורות החיוביות וגם בשליליות, והפרש ההופעות גדול מ-1 ( ברוב המקרים המילים מופיעות פעם אחת ולכן הן לא רלוונטיות, בחירה זו לא מכניסה אותם לשק המילים)

.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **/** | **Precision** | **recall** | **accuracy** | **fscore** |
| **all the words**  **(size = 50620)** | 0.8798 | 0.6839 | 0.795 | 0.7691 |
| **AutoFindAlgo 1**  **(size = ~22500)** | 0.8985 | 0.737 | 0.8265 | 0.8092 |

כפי שניתן לראות כל תוצאות המדדים עלו. תוצאה זו אינה מפתיעה שכן התאמנו את שק המילים לביקורות הללו והורדנו מספר מילים רב שכמעט לא רלוונטי לחישוב ורק גורם לרעש (כמעט חצי מין המילים).

נעריך את תוצאות המסווג החדש על פי ביקורות חדשות בהמשך הדוח.

# 2.

בשלב זה מצאנו את 300 המילים השכיחות ביותר בכל קורפוס האימון, הסרנו אותן משק המילים שיצרנו ואמנו את המסווג שוב.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **/** | **Precision** | **recall** | **accuracy** | **fscore** |
| **Auto finding Algo 1**  **Without common 300**  **(size=~22200)** | 0.9095 | 0.729 | 0.828 | 0.809 |

כפי שניתן להבחין השינויים לא משמעותיים שכן היחס בין כמות המילים שהוסרו מכלל המילים זניח.

תוצאת מדד ה-precision עלתה במעט ותוצאת מדד ה-recall ירדה במעט – כאשר המערכת החליטה כי ביקורת היא חיוביות , ההסתברות שאכן היא חיובית גדלה אך המערכת סיווגה פחות טקסטים חיוביים כחיוביים.

תוצאת מדד ה-accuracy עלתה במעט ותוצאת מדד ה-f-score ירדה במעט – המערכת צדקה ביותר מקרים.

# 3.

בשלב זה ניסינו לשפר את ביצועי המסווג על ידי שימוש באיחוד מספר אלגוריתמים אלטרנטיביים לבחירת שק המילים:

* One – כלל המילים המופיעות אך ורק בביקורות החיוביות ומתפרסות על יותר משתי ביקורות איחוד עם כלל המילים המופיעות אך ורק בביקורת השליליות ומתפרסות על יותר משתי ביקורות
* Two- כלל המילים המופיעות גם בביקורות החיוביות וגם בשליליות והפרש ההופעות גדול מ-1

Three – X המילים המופיעות גם בביקורות החיוביות וגם בשליליות ויחס הפריסה שלהם גדול ביותר. ( X מספר המילים שאנו בוחרים)

כאשר יחס פריסה עבור מילה מוגדר להיות היחס בין מספר הטקטסים החיובים בהם מופיע המילה ולמספר הטקטסים השלילים בהם מופיע המילה

ניתן להתרשם מהתוצאות של כל אלגוריתם:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **/** | **Precision** | **recall** | **accuracy** | **fscore** |
| **all the words** | 0.8798 | 0.6839 | 0.795 | 0.7691 |
| **one & two** | 0.8985 | 0.737 | 0.8265 | 0.8092 |
| **one & two & three 500** | 0.8985 | 0.737 | 0.8265 | 0.8092 |
| **Three 500** | 0.8841 | 0.796 | 0.8455 | 0.8368 |
| **three 5000** | 0.9284 | 0.8029 | 0.8705 | 0.8608 |
| **three 2000** | 0.9334 | 0.856 | 0.8975 | 0.8927 |
| **one & three** | 0.9978 | 0.899 | 0.9485 | 0.9455 |
| **one & three 2000**  **(size = 8054)** | 0.9902 | 0.909 | 0.95 | 0.9476 |

כפי שניתן לראות האלגוריתם one איחוד עם three 2000 הניב את התוצאות הטובות ביותר.

אלגוריתם זה לא בוחר אף אחת מהמילים ברשימת 300 המילים הנפוצות ביותר בקורפוס ולכן לא נמצא שינוי בתוצאות כאשר מורידים את 300 המילים הנפוצות ביותר.

כמו כן ניתן להבחין כי מדד ה-accuracy ומדד ה-f-score מניבים תוצאות קרובות מאד שכן גם מדדי הprecision וה-recall עלו יחדיו.

# בדיקת איחוד המסווג על ביקורות חדשות

בשלב זה לקחנו 100 ביקורות חיוביות ו-100 ביקורות שליליות שלא מופיעות בתיקיית ה-training ולכן המילים שלהם לא נכללו כאשר בחרנו את שק המילים.

כעת ניתן לבחון את המסווג ואת שיטת בחירת שק המילים על 200 הביקורות החדשות.

להלן התוצאות:

|  |  |  |
| --- | --- | --- |
| **/** | **100 Pos test** | **100 Neg test** |
| **all the words** | 66 | 89 |
| **one & two** | 72 | 89 |
| **one & two no 300** | 72 | 89 |
| **one & three 2000** | 94 | 99 |

ניתן לראות כי המסווג העובד עם שק מילים המכיל את כל המילים בקורפוס מחזיר את התוצאות הגרועות ביותר: מתוך 100 תגובות חיוביות המסווג הנ"ל מסווג רק 66 ביקורות כחיוביות.

האלגוריתמים מסעיפים א' וב' מחזירים תוצאות מעט יותר טובות.

האלגוריתם שנבחר בסעיף ג' מחזיר את התוצאות הטובות ביותר :

* מתוך 100 ביקורות שליליות המסווג מסווג 99 כשליליות
* מתוך 100 ביקורות חיוביות המסווג מסווג 94 כחיוביות