בס"ד

עיבוד שפות טבעיות – דו"ח סיווג טקסטים

מגישים:

פורטנוי ליאור

אפראימוב אורן

מערכת הפעלה:

נבדק על כמה מערכות הפעלה: windows 7, windows 10 ו-Ubuntu 15.4.

התוכנית רצה XXX

סיווג ו"ערבוב" המסווג:

הסיווג שאנו בנינו, נעשה על רשימה של וקטורים מעורבבת, כך ששלב הtest ו-learn יעשו על שני המחלקות ולא רק על מחלקה אחד!

בנוסף, כדי לשמור על "הוגנות" ועל מנת שנוכל להשוות בין שני המסווגים, שמרנו על אותו סדר של ווקטורים, כך שנוכל באמת לראות את ההבדלים.

שאלה 1 א':

בחרו כמה עשרות (עד 50) מילים אשר עשויות לדעתכם לתפוס את ההבדלים בין ביקורות חיוביות לשליליות. תוכלו להעזר ברשימות מילים קיימות ברשת למטלות מסוג זה (שתמצאו בעצמכם). פרטו את הרשימה שבחרתם.

תשובה:

לינק לרשימת המילים (לחץ עליי)

בחירת המילים נעשתה על ידי חיפוש בגוגל, ללא התעמקות ברשימה זו או אחרת. למרות, שאחוזי הזיהוי לא מאוד משמעותיים, והיה ניתן לשפר בעזרת חיפוש רשימה טובה יותר, או על ידי הפונקציה SelectKBest, בחרנו להשאיר אותה על מנת לדעת אילו מילים טובות יותר וטובות פחות.

שאלה 1 ד':

יש להעריך את הביצועים של כל מסווג בעזרת ten-fold-cross-validation ולדווח את הדיוק(accuracy) **הממוצע** של כל ה-folds. פרטו את התוצאות שקבלתם ודונו בהם: האם הן כפי שציפיתם? איזה מסווג עבד טוב יותר מאחרים?

תשובה:

~~~~Question 1~~~~
SVM classifier - the accuracy is of is 0.585 it's take 6.6809824647848775 sec
Navie-Bayes classifier - the accuracy is of is 0.573 it's take 0.16147776422096527 sec
Decision-Tree classifier - the accuracy is of is 0.6145 it's take 0.43218742423175893 sec
KNN classifier - the accuracy is of is 0.565 it's take 0.7158598082232501 sec

כן, צפינו שמסווג SVM או Decision Tree ייתנו את הערך הכי טוב.
SVM היה צריך להכריע בין 2 מחלקות, וזה משימה די קלה עבורו, הוא צריך לקבל "פיצרים" טובים וכך הוא יידע להכריע בין המחלקות, אך בגלל שמילים פיצרים) שהבאנו לא היו ממש טובות, Decision Tree קיבל את התוצאה כי טובה.

שאלה 2 א':

בנו feature vectors מהטקסטים למשל בעזרת (CountVectorizer).
תשתמשו **בכל המילים של הטקסט** (לא מילון מוגבל), **ללא stop words** של אנגלית. המשיכו וחשבו שכיחויות (לא אינדיקציה בינארית) של כל המילים (השתמשו ב tf-idf) בעזרת TfidfTransformer. כמה מילים שונות ישנן בטקסטים (במילים אחרות, מה אורך ה feature vectors שנוצרו)?

תשובה:

נוצרו 22,878 מילים (טוקנים) שונים זה מזה.

שאלה 2 ג':

סווגו כעת את ה feature vectors בעזרת המסווגים מסעיף 1.ג.
העריכו את ביצועי המסווגים בעזרת ten-fold-cross-validation כפי שעשיתם ב 1.ד. האם תוצאות טובות יותר כעת? דונו בהבדלים מהסעיף הקודם (1.ד.).

תשובה:

~~~~Question 2~~~~

SVM classifier - the accuracy is of is 0.5235 it's take 35.734041024999286 min
Navie-Bayes classifier - the accuracy is of is 0.789 it's take 0.2093898359104287 min
Decision-Tree classifier - the accuracy is of is 0.675 it's take 2.973571964186552 min
KNN classifier - the accuracy is of is 0.6425 it's take 5.950003413644413 min

תשובה:

פרט למסווג SVM, תוצאות כל המסווגים ישתפרו ובחלק אף שיפור ניכר (Navie-Bayes).

לדעתנו, עודף פיצרים, פגע במסווג SVM שנדרש לו הרבה יותר זמן (הבדלים ניכרים מכמה שניות לעשות דקות), היות שרוב הפיצרים שנבחרו היו לא מועילים ולכן המסווג נפגע.

בשאר המסווגים, עודף פיצרים הועיל ונתן תוצאות הרבה יותר טוב, כנראה מדרך החישוב שלהם, שכן השיפורים הרבה יותר גבוהים מקודם.

שאלה 3 א':

השתמשו ב SelectKBest על מנת לבחור את 50 המילים בעלות התרומה **הגבוהה ביותר** לסיווג. התרשמו מרשימת המילים והשוו אותה לרשימה שבחרתם באופן ידני בסעיף 1.א. האם כל המילים שקבלתם כעת הן צפויות? פרטו את התוצאות (50 המילים).

תשובה:

']amazing', 'annoying', 'avoid', 'awful', 'bad', 'badly', 'beautiful', 'best', 'boring', 'brilliant', 'effects', 'excellent', 'great', 'highly', 'hitchcock', 'horrible', 'hour', 'idea,'
' just', 'lame', 'life', 'like', 'lives', 'looks', 'love', 'loved', 'make', 'masterpiece', 'minutes', 'money', 'mother', 'perfect', 'performance', 'plot', 'poor', 'poorly', 'portman' ,'
ridiculous', 'script', 'strong', 'stupid', 'superb', 'terrible', 'thing', 'war', 'waste', 'wasted', 'wonderful', 'worse', 'worst['

רוב המילים היו צפויים, לאדם אשר מבין ויודע איך נראים ביקורות.

אך, יש מילים כמו, hour או 'hitchcock, שלא ממש היו צפויים כי האחד זה שם של סרט (כנראה היה הרבה ביקורות על הסרט) ואילו השני לא מסמן עם דעתו היא חיובית או שלילית וזה רק מסמן זמן (כמו המילים juse או make)

לכן, אפשר להגיד שאת רוב המילים היו צפוי לדעת שייבחרו שכן הן מסמנים ביקורת "חיובית" או "שלילית", אך יש כמה מילים שהופיעו בגלל שם הסרט, שחקנים או כי הם נפוצות בביקורות האלו שלא ממש תרמו לסיווג!

<u>שאלה 4 ב':</u>

חיזרו על סעיף 2 (**עם כל תת-סעיפיו**) כאשר כעת אתם משתמשים ברשימת מילים סגורה ל CountVectorizer (יש לו אופציה לקבל מילון, ראו תיעוד).
השתמשו ברשימת 50 המילים בעלות התרומה הגבוהה ביותר לסיווג מסעיף 3.א.
השווה את התוצאות למספרים שקבלתם עבור 2.ג. (עם bag-of-words). האם ההבדלים משמעותיים?

<u>תשובה:</u>

~~~~Question 4~~~~
SVM classifier - the accuracy is of is 0.7955 it's take 9.041446205526427 sec
Navie-Bayes classifier - the accuracy is of is 0.8055 it's take 0.7730379447352789 sec
Decision-Tree classifier - the accuracy is of is 0.6865 it's take 1.0152127963269777 sec
KNN classifier - the accuracy is of is 0.7495 it's take 2.133340372447492 sec

ההבדלים ניכרים!!
בסיווג הנ"ל, הכנסנו את המילים הכי טובות לסיווג (למרות שיש כמה שאפשר לוותר), שכן סיווג טוב נעשה על ידי בחירת פיצרים טובים, כך שמסווג יכול להפריד באופן חד משמעי בין 2 מחלקות!
למרות שמסווג SVM ו-Navie-Bayes הגעו לדיוק של כ-80% שני המסווגים האחרים היו מעט רחוקים יותר, כנראה מאופן הסיווג שלהם, כי המילים אשר ניתנו כן, היו מאוד טובות ואפיינו היטב את ההבדל בין שני המחלקות (הביקורות).
~~~~

בס"ד