## תרגיל בית 3- סיווג טקסטים

* ג’וואד מרג’יה (Jawad Margieh) 203686670
* יורי ימליאנוב (Yuri Emelianov) 316750504

**כללי:**

תרגיל בית נעשה במערכת הפעלה Windows7/8 64 bit

השתמשנו ב PyCharm כ IDE

)ד”א הרצנו את הקוד גם על מאקבוק עם מערכת הפעלה OS X El Capitan)

הקוד מתועד בצורה מלאה.

**1. סיווג בעזרת בחירת features באופן ידני:**

**סעיף 1א:**

מילים קבלנו כקלט לתרגיל בית:

'the','of','and','to','a','in','for','is','on','that','by','with','I','or','not','you','be','are','this','at','it','as','from','your','have','was','an','will','all','can',

'more','has','Home','In','we','one','but','about','which','New','do',

'their','our','Search','other','they','up','my','time','out'

**סעיף 1ד:**

SVM - accuracy: 0.79 (+/- 0.12)

Naïve Baise - accuracy: 0.77 (+/- 0.12)

Decision Tree - accuracy: 0.79 (+/- 0.16)

KNN - accuracy: 0.74 (+/- 0.14)

פירוט של התוצאות שקבלתם:

האם הן כפי שציפיתם?

* התוצאות היו פחות או יותר צפויות ואחוז סיווג לא יותר מידי גבוה בגלל שמילים שקבלנו הם מילים הנפוצות באותה רמה לכל שיר וגם לאותו מסווג אז חשבנו שפחות או יותר נקבל בכל מסווג אותו סיווג. חשוב לציין שאם בוחרים מילים כfeatures למסווג אז צריך להתחשב בנתונים וכאן לא הייתה התחשבות כלומר קבלנו משהוא כללי.

איזה מסווג עבד טוב יותר מאחרים?

* כל מסווג פחות או יותר מגיע לאותה תוצאה ומריצה לריצה רואים שינוי קל ולא משמעותי(ניסינו להריץ הרבה פעמים), ניתן לראות מהתוצאות שאין מסווג שמנצח מהריצות שראינו בדרך כלל Decision Tree נותנת ממש קצת אבל תוצאה יותר טובה וכל השאר פחות או יותר מראים אותו דבר.

**סעיף 1ה:**

Accumulated SVM - accuracy: 0.72 (+/- 0.11)

Accumulated Naïve Baise - accuracy: 0.66 (+/- 0.10)

Accumulated Decision Tree - accuracy: 0.80 (+/- 0.12)

Accumulated KNN - accuracy: 0.81 (+/- 0.10)

האם יש שינוי לטובה/לרעה בתוצאות?

לפי מה שקבלנו רואים שSVM וNAIVE BAISE מראים תוצאות פחות טובות(לא משמעותי אבל פחות טובות). שאר המסווגים נראים קצת יותר טוב אבל כמו שנאמר לא משמעותי(תוספת בין 5-10 אחוז)

**סעיף 1ו:**

פרטו אילו מילים הוספתם:

* לפי המלצה ניסינו להוסיף מילים שמופעים בשמות של השירים.

אם למשל מילה מופיע אז הוספנו ,דוגמא: Michell של שיר Beatles

האם יש שינוי משמעותי בתוצאות?

* התוצאות לא השתפרו בצורה שאפשר להגיד שיש שינוי משמעותי אז אנחנו חושבים שאין שינוי משמעותי

**2. סיווג בעזרת bag-of-words:**

**סעיף 2א:**

כמה מילים שונות ישנן בטקסטים (במילים אחרות, מה אורך ה- feature vectors שנוצרו)?

* **4598**

**סעיף 2ג:**

האם תוצאות טובות יותר כעת?

SVM - accuracy: 0.79 (+/- 0.10)

Naïve Baise - accuracy: 0.82 (+/- 0.08)

Decision Tree - accuracy: 0.81 (+/- 0.07)

KNN - accuracy: 0.82 (+/- 0.10)

* ניתן לראות שתוצאות קצת השתפרו בערך ב10% אצל המסווגים שלפני זה היות קצת פחות טובים משאר: Naïve Baise וKNN. עוד הפעם הסיווגים כאן נראים קצת פחות טובים אבל לא בצורה דראסטית

הבדלים מהסעיף הקודם (1.ד):

* הfeature vector הרבה יותר גדול (50 VS 4598)
* אין STOP WORD של אנגלית

אז כנראה שני הדברים האלו עזרו קצת לשפר (לא משמעותי) את הסעיף הקודם(1.ד)

**3. בחירת המילים המשמעותיות ביותר לסיווג:**