תרגיל בית 3- סיווג טקסטים

- ג'וואד מרג'יה (Jawad Margieh) 203686670
- יורי ימליאנוב (Yuri Emelianov) 316750504

כללי:
תרגיל בית נעשה במערכת הפעלה 64 bit Windows7/8
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 1.ד.:

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 שאנו מחזיקים ביד אבל במקרה שלנו לא הייתה התחשבות בנתונים(מילים של השיר, זמר..) כלומר קבלנו משהוא כללי לחלוטין.

- 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 1.ה.:

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 ש–SVM ו–BAISE NAIVE מראים תוצאות פחות טובות(לא הבדל כזה משמעותי אבל פחות טוב). שאר המסווגים נראים קצת יותר טוב אבל כמו שנאמר גם לא הבדל כזה משמעותי (תוספת של 5-10 אחוז).

סעיף 1.ו.:

פרטו אילו מילים הוספתם:

- לפי המלצה ניסינו להוסיף מילים שמופעות בשמות של השירים.
  לדוגמא: michelle של שיר Beatles.
  מילים שהוספנו: ‘speech’, ‘blue’,‘everybody’, ‘medley’, ‘nobody, ‘hello’, ‘come’, ‘michelle’, ‘long’

תוצאות:

SVM - accuracy: 0.79 (+/- 0.09)
Naïve Baise - accuracy: 0.76 (+/- 0.12)
Decision Tree - accuracy: 0.78 (+/- 0.18)
KNN - accuracy: 0.76 (+/- 0.18)

- האם יש שינוי משמעותי בתוצאות?
  מסקנה: כמעט ולא רואים הבדל בין התוצאות/ביצועים של כל אחד מהמסווגים בהשוואה סעיף 1.ד.
  התוצאות לא השתפרו בצורה שאפשר להגיד שיש שינוי משמעותי אז אנחנו חושבים שאין שינוי משמעותי

2. סיווג בעזרת:bag-of-words

<u>סעיף 2.א.:</u>

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

- **4598**

<u>סעיף 2.ג.:</u>

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 השתפרו בערך ב5%-10% בעיקר בשני מסווגים: Naïve Baise ו KNN. ב SVM ו Decesion Tree הם כמעט ונותרו ללא שינוי.
- שוב פעם התוצאות הם בינוניות ולא רואים שיפור משמעותי.

- הבדלים מהסעיף הקודם (1.ד.):

- אורך feature vectors הרבה יותר גדול, כל אחד מה– FV הוא באורך 4598 לעומת 50
- אין התייחסות ל Stop words, כלומר לא לקחנו מילים אלו ב FV שלנו.
- כנראה שני השינויים האלו לא ממש עזרו למשימת הסיווג (לא משמעותית) מהסעיף הקודם (1.ד)

3 .בחירת המילים המשמעותיות ביותר לסיווג:

סעיף 3.א.

המילים שהתקבלו מהפעלת KBest :

'alive','baby','believe','blue','body','bout','britney','cameras', 'cares','chorus','clear','confess','control','crazy','crowd','damn','darling','dont','feel','feels','felt','floor','god','hit','home','hot','im','instrumental','just','kinda','let','lets','like','love','make','moving','oh','old','ready','shouldn','shut','sing','tonight', 'touch','tryin','turn','wanna','watch','ya','yes',

- לא ניתן לדעת שום דבר על רשימת המילים וזאת מכיוון שכל שיר יכול לדבר על נושא שונה, שיר אחד יכול להיות עצוב ושיר אחר יכול להיות שמח או שיר אחד יכול להיות מדבר על בנאדם שהזמר אוהב ושיר אחר על נושא אחר לגמרי..
  מהמילים שהוספנו פגענו במילה אחת מרשימת ה Best50 ובהשוואה למילים הכי נפוצות אין אף מילה משותפת כי לפי הדרישה לא התייחסנו ל stop words שהן מילות הכי נפוצות בטקסטים אך ללא משמעות.
  בסופו של דבר המילים שקיבלנו הן ממש לא צפויות.

4 .סיווג בעזרת רשימת מילים מצומצמת

סעיף 4.ב

SVM - accuracy: 0.77 (+/- 0.08)
Naïve Baise - accuracy: 0.80 (+/- 0.10)
Decision Tree - accuracy: 0.85 (+/- 0.09)
KNN - accuracy: 0.78 (+/- 0.09)

עדיין ההבדלים הם לא כזה משמעותיים אבל יש לשים לב לתוצאה של ה מסווג **Decision Tree** אשר נתן תוצאה לא רעה בכלל יחסית לתוצאות של מסווגים אחרים בסעיף הזה ובסעיפים קודמים, וזאת מסיבה שהמילון הוא מצומצם ויותר איכותי, כלומר יותר משפיע.

5. סיווג לבחירתנו

בסעיף זה יצרנו מילון מילים באורך 30 כאשר מילים אלו מופיעות בהרבה שירי אהבה/רומנטיקה. את הרשימה אנחנו יצרנו.
ואז כמו בסעיפים קודמים, אלה הם ה FV שלנו.
תוצאות שקיבלנו:

SVM - accuracy: 0.77 (+/- 0.01)
Naïve Baise - accuracy: 0.77 (+/- 0.01)
Decision Tree - accuracy: 0.77 (+/- 0.01)
KNN - accuracy: 0.76 (+/- 0.04)

התוצאות הם פלוס מינוס טובות יחסית, כמובן נסינו מילונים שונים, ועבור זמרים שונים הגענו לתוצאות שונות.