# שיעורי בית מספר 3

# בעיבוד שפות טבעיות

**מגישים:**

**אילן גודיק – 316315332**

**צ'רלי מובאריכי – 316278118**

**שאלה מספר 1**

בשלב זה עלינו לבחור מאגר מילים, שלפי הכלה בינארית שלהם בטקסט, על המסווג יהיה להחליט האם ביקורת של סרט היא חיובית או שלילית.

אנחנו עברנו על מדגם די גדול של ביקורות, גם חיוביות וגם שליליות, ואספנו מילים שלדעתינו הן המכריעות האם הביקורת חיובית או שלילית.

סך הכל אספנו 210 מילים.

שמנו לב, שישנם מילים רבות מאותו השורש, עם נטיות שונות, כגון:

worse, worst, beautiful, beautifully, disappoint, disappointing, disappointment, surprise, surprised, surprising, tense, intense, …

ועוד רבים אחרים.

מה שהחלטנו לעשות, זה לכלול רק את השורשים, שהן התחיליות המכלילות את המילים שהן נטיות אחת של השנייה.

לאחר ביצוע הקטנת המילים לשורשים, נותרנו עם 161 שורים ומילים.

לשם יצירת הFeature Vector של טקסט, בדקנו האם קיים טוקן בטקסט המכיל כל שורש בהתאם.

תוצאות עם כל ה161 מילים שבחרנו: (הBaseline שלנו)

- SVM: 0.79 (+- 0.05)
- Naive Bayes: 0.81 (+- 0.05)
- DecisionTree: 0.69 (+- 0.06)
- KNN: 0.67 (+- 0.08)

והFeatures הם:

10, absent, anti, applauded, applause, avoid, away, awful, bad, beautiful, beauty, best, better, bother, brillian, but, can, care, charm, classic, confusing, contrary, crap, cried, cry, cute, delightful, despite, different, disappoint, emotion, enjoy, entertain, especially, even, ever, excellen, exceptional, extremely, fail, fake, fan, fantastic, favorite, fear, feel, felt, fine, fool, forgettable, forgotten, fun, genius, good, gorgeous, great, hate, highly, hilarious, horribl, however, humor, humorous, impress, incredible, insufferable, insulting, intense, intriguing, irritating, juicy, lack, laugh, liar, like, long, love, low, magnificent, masterpiece, meaningless, memorable, mild, minus, money, mood, most, move, moving, must, n't, negative, never, nice, not, nothing, off, other, over, pale, perfect, performance, pleasant, pleasure, plot, plus, pointless, poor, positive, powerful, pretty, problem, professional, quality, quite, rare, real, recommend, ridiculously, ripped, ruin, serious, shame, should, since, sold, strength, strong, stunning, stupid, sucks, superb, superior, surpris, talent, tear, tedious, ten, tense, terrible, thank, thumbs, touching, tough, truly, unbelievable, unique, unsatisfied, very, waste, watch, well, well-made, wonderful, wors, worth, wrong, yet, predictable, miscast, baffl

## Genetic Algorithm Feature Selection

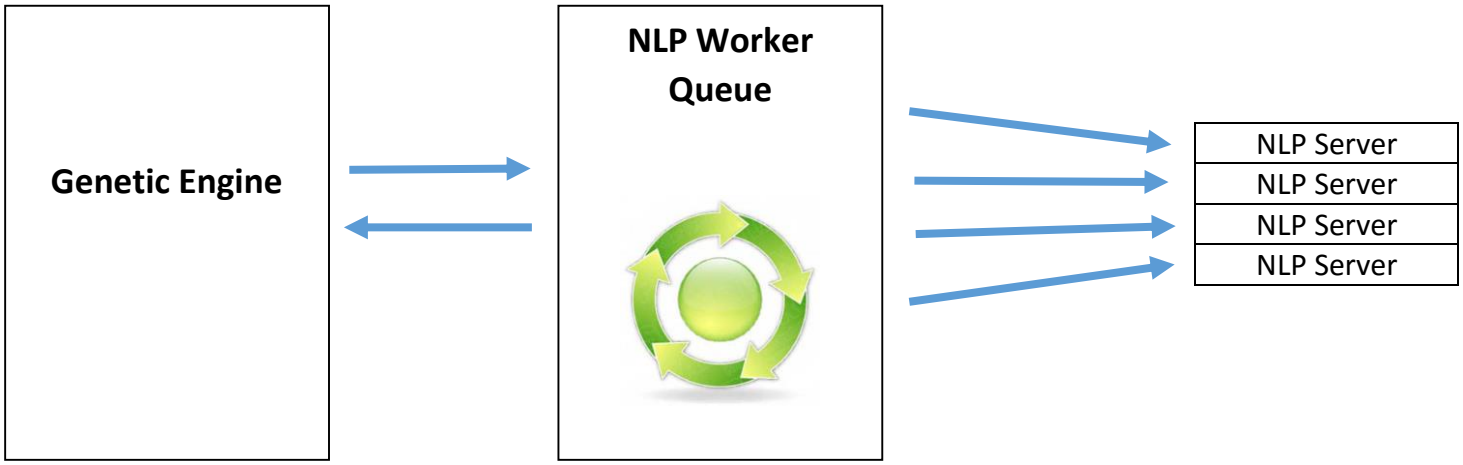
עלינו לספק 50 Features.

לשם כך, פתרנו את הבעייה של Feature Selection בעזרת מנוע האלגוריתם הגנטי שפיתחנו במעבדה בבינה מלאכותית, בדרך הבאה:

הבעייה היא מציאת תת קבוצה של Features בגודל 50 הממקסמת את אחוזי ההצלחה של הClassifier.

ייצגנו כל גן ע"י מחרוזת בינארית המייצגת תת קבוצה מסויימת של Features. יש את כל הפעולות הגנטיות של One point crossover וMutation, ובנוסף יש מנגנונים של Local optima detection ע"י זיהוי מתי יש לנו יותר מדי גנים דומים באוכלוסיה ע"י חישוב ההמוצע של מרחק ההאמינג בין כל הגנים, וישנה התמודדות עם המינימום הלוקאלי בדרכים שונות.

לשם התקשורת בין האלגוריתם הגנטי למסווגים בPython, ולשם ביצועים טובים, יצרנו מספר שרתי Python שרצים <u>במקביל</u> (כך אין לנו צורך להסתמך על המקביליות בScikit, שהייתה בעייתית), ויצרנו NLP Worker Queue המתחזק את השרתים הפנויים בכל רגע בתוכנית. כאשר ברצונינו לקבל את אחוזי ההצלחה עבור תת קבוצה מסויימת, אנו שולפים Process של שרת מהתור, שולחים לו בקשה, מקבלים תשובה, ומחזירים אותו לתור.



התוצאות כאשר עשינו אופטימיזציה עבור Naïve Bayes:

**התשובה ל1 ד':**

- SVM: 0.80 (+- 0.04)

- Naive Bayes: 0.83 (+- 0.05)

- DecisionTree: 0.71 (+- 0.04)

- KNN: 0.68 (+- 0.07)

התוצאות של המסווגים הפתיעו אותנו, מכיוון שבאופן גורף הNaive Bayes מצליח לסווג טוב יותר, וזהו אלגוריתם סיווג כ"כ טריוויאלי ביחס לאחרים, ובמיוחד לSVM, שציפינו לתוצאות מובילות בו מאשר באחרים. בKNN ציפינו לתוצאות נמוכות יותר מSVM, וכן גם בDecision Tree. הבחנו שפחות Features, אך יותר איכותיים, נותנים תוצאות טובות יותר בDecision Tree.

כלומר, הצלחנו להגיע לשיפור עד ל83% של הצלחה עם שימוש רק ב50 Features, והם:

**התשובה ל1 א':**

applause, avoid, bad, beautiful, best, better, bother, brillian, crap, cry, delightful, enjoy, entertain, even, excellen, favorite, fool, genius, gorgeous, great, highly, horribl, however, humorous, incredible, liar, love, money, n't, never, not, perfect, performance, plot, poor, rare, should, strong, stupid, superb, surpris, tear, tense, terrible, thank, touching, waste, well-made, wonderful, wors

עשינו אופטימיזציה על כל Classifier בנפרד, והגענו לתוצאות הבאות:

עבור SVM: 81%

עבור KNN: 75%

עבור Decision Tree, עם 15 Features בלבד: 75%

ה15 Features של הDecision Tree:

Avoid, awful, bad, brillian, disappoint, excellen, great, horribl, love, n't, strong, stupid, tedious, waste, wonderful

הבחנו בתופעה מעניינת, בה הFeatures של Naïve Bayes ושל KNN די מסכימים, כלומר Features שטובים באחד הם די טובים גם בשני, אך יש תופעה הפוכה עבור KNN וDesicion Tree: תת קבוצה של Features שטובים בKNN או בDecision Tree גורמים לתוצאות מאוד, מאוד גרועות בשאר המסווגים (55% הצלחה בNaive Bayes עבור Features של Decision Tree)

** השתמשנו ב**Shuffling** עבור הKFold Cross Validation, וזה היה קריטי בכדי לא לעשות overfitting בחיפוש תת הקבוצה הטובה ביותר של הFeatures.

## Monte Carlo Tree Search Feature Selection (FUSE)

ניסינו לפתור את בעיית הFeature Selection גם בעזרת עצי חיפוש מונטה קרלו, לפי המאמר של אלגוריתם הFUSE:

Gaudel, Romaric, and Michele Sebag. "Feature selection as a one-player game." *International Conference on Machine Learning*. 2010.

מרחב החיפוש שלנו הוא Lattice של תתי קבוצות, כאשר השורש הוא הקבוצה הריקה, ויש קשת מכל תת קבוצה לתת קבוצה בהיוסף איבר נוסף.

כלומר, יש לנו מרחב חיפוש עם Branching Factor של 161-i ברמה הi.

כמובן שזה בלתי אפשרי לעשות חיפוש רגיל במרחב מצבים שכזה, ואף לא ניתן לפתוח 3 רמות בעץ.

אלגוריתם עצי החיפוש מונטה קרלו מהווה אלגוריתם טוב לפתרון הבעיה, מכיוון שהוא מפתח עץ לא סימטרי, ומשקיע יותר עבודה בתתי עצים מבטיחים, ומהווה פיתרון לבעיית הN-Bandit Problem ע"י איזון טוב בין Exploration לExpoitation.

בנוסף לא נרצה לפתוח את כל ה~161 בנים במסלול לעלה, לכן אנו צריכים להשתמש בכל מידע אפשרי שאנו אוספים במהלך החיפוש וביקור בצמתים אחרים בעץ:

אלטרנטיבת הRAVE לעצי חיפוש מונטה קרלו:

G-Rave: ניתן חלק מה'ציון' להאם לבחור בקשת בעץ ע"י הסתכלות על כל תתי הקבוצות שכבר ביקרנו בהם המכילים את הFeature שאנו מחליטים לגביו כעת.

L-Rave: נסתכל בכל אחוזי הדיוק של הFeature בכל תתי הקבוצות בתת העץ הנוכחי.

אז בעזרת שיפורים אלה, יש לנו יותר מידע על תתי עצים שלא ביקרנו בהם עוד, ואנו נעבור לאט-לאט מהציון הגלובלי, ללוקאלי ואז לאמיתי ככל שאנו מאששים את הראיות שלנו ביותר ביקורים בתת העץ.

בסופו של דבר, קיבלנו תוצאות תחרותיות לאלה שהשגנו באלגוריתם הגנטי, אך ישנה בעייה בבחירת ההיפר פרמטרים הנכונים לאלגוריתם כך שהוא יהיה יציב ויתכנס היטב.

**שאלה מספר 2**

התקבלו 22878 מילים שונות בBag of Words לאחר CountVectorizer עם אי התחשבות בWords-Stop.

התוצאות של המסווגים:

- SVM: 0.79 (+- 0.07)
- Naive Bayes: 0.90 (+- 0.05)
- DecisionTree: 0.69 (+- 0.08)
- KNN: 0.86 (+- 0.06)

מאוד הפתיע אותנו עד כדי כמה טוב עובד Naïve Bayes בשלב זה – 90% הצלחה!

בנוסף יש שיפור מאוד ניכר בKNN, כנראה כי יש צפיפות גבוהה בהרבה בייצוג של כל המילים בBag of Words, ולכן k השכנים הקרובים ביותר יהיו קרובים בהרבה מאשר במילון מצומצם יותר, ולכן גם הדיוק של בחירת הרוב מk איברים אלה תהיה גבוהה יותר.

הייתרון של Bag of Words הוא שיש לו גישה לכל המילים ככלל בכל הטקסטים, שהמסווגים יכולים להשתמש בהם, ואם אנו מגבילים ידנית את אוסף המילים, ייתכן שפיספסנו מילים משמעותיות שיכולות לעזור לסיווג.

SVM וDecision Tree קיבלו אחוזי הצלחה נמוכים יותר ממקודם, אך זה בטווח השגיאה בשינוי שכזה בFeatures.

**שאלה מספר 3**

50 המילים שSelectKBest נותן לנו הן:

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

חלק גדול של המילים בהחלט הגיוני שיפריד בין Reviews חיוביים לשליליים, אך ישנן מילים רבות שאין להן שום קשר לחיוביות הreview, והם artifact סטטיסטי ומהווים overfitting מוחלט למספר reviews בdataset, והן מילים כגון hitchcock, portman, mother, war

**שאלה מספר 4**

לאחר שימוש ב50 המילים שהתקבלו בשאלה 3 עם SelectKBest, תוצאות המסווגים הן כלהלן:

- SVM: 0.80 (+- 0.07)
- Naive Bayes: 0.81 (+- 0.05)
- DecisionTree: 0.69 (+- 0.05)
- KNN: 0.76 (+- 0.07)

אנו רואים כאן נפילה משמעותית מאוד בביצועים של המסווגים של Naïve Bayes וKNN, שבהם היה שיפור משמועיתי בBag of Words על גבי בחירת Features ידנית.

ככל הנראה נבחרו מילים לא טובות מספיק ע"י SelectKBest – אנו בחרנו Features יותר טובים בשאלה 1, וניתן היה לראות אילו Features הם בעייתיים, כפי שרשמנו בשאלה 3.

בנוסף, ייתכן שחלק מהמסווגים פשוט עובדים יותר עם יותר Features, וכמה שיותר, יותר טוב, ושההגבלה ל50 מילים לא נותנת מספיק מידע כדי לסווג את כל הטקסטים היטב.

**בדיקת הסיווגים על קורפוסים נוספים:**

בדקנו את הסיווגים שלנו גם על הקורפוס המלא של Standford של הIMDB Reviews, וקיבלנו שקבוצת המילים שלנו קיבלה תוצאות דומות גם שם, כלומר הייתה לנו יכולת הכללה טובה, בניגוד לקבוצת המילים שהתקבלה בסעיף ג' בקורפוס הקטן, עליה הייתה ירידה בכמה אחוזים בקורפוס המלא.