תרגיל בית 1

שיטות בעיבוד שפה טבעית

עומר כהן 308428127
נמרוד אדמוני 203860721

אימון:

סוגי המאפיינים בהם השתמשנו הינם:

מודל	מודל 1 (גדול)	מודל 2 (קטן)
מאפיינים	$f_{100} - f_{107}$ $f_{108} - capital\ letter\ feature$ $f_{109} - only\ number\ feature$	$f_{100} - f_{107}$

כאשר המאפיינים f_{108}, f_{109} ממומשים כך:

f_{108} הוא מאפיין שמזהה שמילה מסוימת היא מספר:

$$f_{108} = \begin{cases} 1 & if\ current\ word\ w_i\ is\ a\ Number\ and\ tag\ is\ t = CD \\ 0 & otherwise \end{cases}$$

f_{109} הוא מאפיין שמזהה שהמילה הנוכחית מתחילה באות גדולה:

$$f_{109} = \begin{cases} 1 & if\ current\ word\ w_i\ begins\ with\ T\ and\ t = NN \\ 0 & otherwise \end{cases}$$

$$f_{110} = \begin{cases} 1 & \langle pp_{word}, p_{word}, c_{word}, c_{tag} \rangle =< w_1, w_2, w_3, tag > \\ 0 & otherwise \end{cases}$$

$$f_{111} = \begin{cases} 1 & \langle suffix, c_{word}, c_{tag} \rangle =< s, w_3, tag > \\ 0 & otherwise \end{cases}$$

$$f_{112} = \begin{cases} 1 & if\ word\ w_i\ length\ is > 4\ and\ \ tag\ is\ t \\ 0 & otherwise \end{cases}$$

$$f_{113} = \begin{cases} 1 & if\ current\ word\ w_i\ contains\ puncuation\ and\ tag\ is\ t \\ 0 & otherwise \end{cases}$$

מספר המאפיינים בכל מודל הם:

מודל 1 (גדול)	מודל 2 (קטן)
f100 15415 f101 13265 f102 22393 f103 8150 f104 1060 f105 44 f106 32132 f107 30793 f108 46 f109 76	f100 75 f101 333 f102 349 f103 118 f104 107 f105 21 f106 70 f107 66 f108 22 f109 29 f110 4 f111 57 f112 34 f113 24

השיפורים שהוספנו למודלים הם:
מודל 1 – אותיות גדולות וזיהוי מספרים
מודל 2 – בנוסף לזיהוי אותיות גדולות ומספרים גם את משפחות המאפיינים 110-113.
זמן האימון עבור מודל 1 הוא 5 דקות עם אחוז דיוק של על קובץ המבחן . עבור מודל 2 זמן האימון הוא שניות בודדות ואחוז הדיוק המשוערך הוא 83.8% , בפועל המודל מבצע אופטימיזציה על קובץ האימון ל 88%
שני המודלים אומנו על Macbook pro 13'' 2021 עם ערכת שבבים M1.

<u>הסקה:</u>

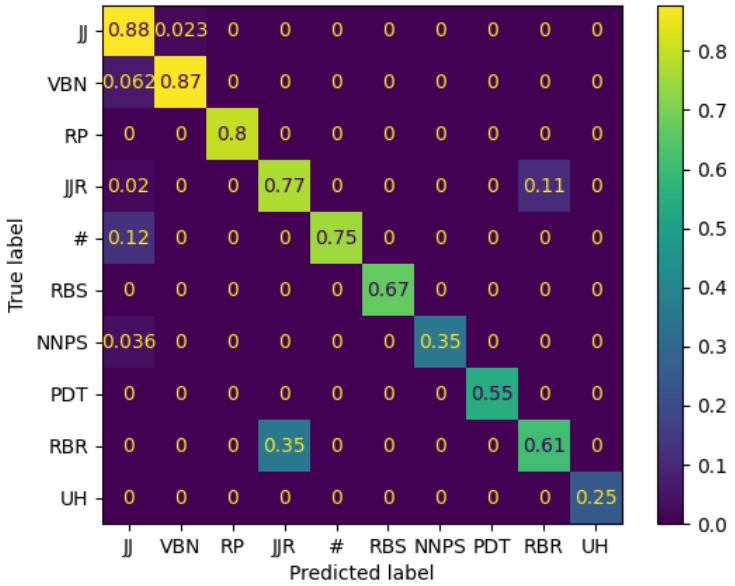
ההסקה התבצעה על אלגוריתם $Viterbi$ עם $beam = 2$ עבור מודל 1 ו $beam =$ עבור מודל 2 ללא שינויים.

<u>מבחן:</u>

עבור מודל 1 אחוז הדיוק עבור קובץ המבחן $test1.wtag$ הוא 95.8%:

```
you have 123374 features!
123374
f100 15415
f101 13265
f102 22393
f103 8150
f104 1060
f105 44
f106 32132
f107 30793
f108 46
f109 76
the following tags the model confused the most
JJ confused 12.224448897795593% of the time.
VBN confused 12.650602409638555% of the time.
JJR confused 23.0% of the time.
RBS confused 33.333333333333336% of the time.
# confused 25.0% of the time.
RP confused 20.408163265306122% of the time.
RBR confused 39.21568627450981% of the time.
PDT confused 45.45454545454546% of the time.
NNPS confused 65.06024096385542% of the time.
UH confused 75.0% of the time.
Accuracy is: 0.9589000591366056
```

קיבלנו עבור המבחן את ה $Confusion - Matrix$ הבאה:



ניתן לראות כי המודל מתבלבל בהסתברות גבוהה בין התיוגים JJR ו- RBR, נוסיף מאפיין שמחזיק את המילה, סיומת בת 2/3 אותיות והתיוג וכך נוכל להבדיל בין תיוגי JJR שנגמרים ב'*er-*' לבין תיוגי RBR שנגמרים ב'*ly-*' .

עבור מודל 2 נרצה להגדיל את ה-lambda כדי להגדיל רגולריזציה ולתת עדיפות ל - varians נמוך על פני bias נמוך. בנוסף, נרצה גם להגדיל את ערך הסף עבור המאפיינים כדי להשאיר את המאפיינים הדומיננטיים ולהימנע מoverfit על ה-dataset. כדי להתמודד עם המחסור ב-data להערכת ביצועי המודל, פתרון אפשרי הוא להקצות חלק מה-data שברשותנו לvalidation. לדוגמא, לחלק את ה-dataset ל80% אימון ו20% validation. נרצה לאמן מספר פעמים את המודל עם הדאטה המחולק כך שכל פעם התוכן בכל חלק יהיה שונה, ולבצע ממוצע של הערכת הביצועים כדי לקבל שיערוך של ביצועי המודל. לבסוף המודל האמיתי יתאמן על כל ה-dataset.

<u>תחרות:</u>
עבור שני המודלים אנחנו מצפים לקבל אחוזי דיוק כמו על קבצי הtest מכוון שאת שניהם לא ראינו באימון. המודלים שהשתמשנו להם זהים למודלים שדווחו באימון. אם זאת במהלך האימון של המודל נקודת ההשוואה שלנו היא קובץ הtest והשינויים שעשינו תאמו לתוצאות שקיבלנו עליו, לכן הוא למעשה evaluation.

<u>חלוקת עבודה:</u>
כל העבודה על תרגיל הבית נעשתה במשותף.