דוח תרגיל בית 1 עיבוד שפה טבעית:

מגישים : אמית פרידמן 209146299 ושלומי ברזניק 208838409

הערות :

- הוספנו פרמטר שכל קובץ מקבל- אם מקבל 1 מריץ עבור החלק 1 ואם 2 עבור חלק 2.
- הפיצ'רים שהוספנו + הפיצ'רים האותיות הגדולות והמספרים מוגדרים בקובץ features_functions
- כל החלקים פרט להסקה בחלק 1 הורצו על המחשב האישי שלנו כאשר יש לו מעבד עם 2 ליבות אחת של 2.3 GHz והשנייה 2.4GHz ויש למחשב 8 ג'יגה זיכרון RAM.
- את ההסקה של מודל 1 הרצנו על המכונה.

חלק 1:

אימון:

- הגדרנו את ההיסטוריה באופן הבא: (word(i-2),word(i-1),word, tag(i-2),tag(i-1),tag(i))
- הורדנו מאפיינים שהופיעו רק פעם אחת כדי למנוע overfitting וכדי לשפר את זמני האיטרציות של הgradient descent לצערנו בכל זאת הtraining לקח בערך 21 שעות.
- הוספנו עבור שני החלקים את הפיצ'רים הבאים:
  - F_DT(history)={ 1 if word in DT list and tag ='DT' ,0 otherwise}
  - פיצ'רים דומים עבור מילים שייכות לרשימות : WP,WPS,MD,PRP,PRP$,IN,CC,WRB
  - כל פיצ'ר כזה הוסיף מאפיין אחד.

מבחן:

Time Test 1=7203.015183210373

Accuracy test 1=

```
0.91355272756748
```

| Predicted | JJ | NNP | NN | NNS | RB | VBN | VB | IN | VBD | VBP |
|---|---|---|---|---|---|---|---|---|---|---|
| Actual | | | | | | | | | | |
| JJ | 1355 | 34 | 35 | 4 | 13 | 22 | 3 | 0 | 8 | 1 |
| NNP | 87 | 1725 | 84 | 17 | 2 | 0 | 5 | 0 | 1 | 4 |
| NN | 273 | 90 | 2844 | 58 | 5 | 1 | 17 | 2 | 1 | 7 |
| NNS | 15 | 13 | 7 | 1404 | 0 | 0 | 1 | 0 | 0 | 0 |
| RB | 34 | 9 | 13 | 4 | 691 | 0 | 0 | 14 | 0 | 2 |
| VBN | 64 | 3 | 6 | 1 | 0 | 390 | 2 | 0 | 27 | 0 |
| VB | 15 | 7 | 14 | 0 | 1 | 0 | 525 | 1 | 3 | 12 |
| IN | 6 | 11 | 7 | 5 | 63 | 0 | 2 | 2411 | 0 | 0 |
| VBD | 17 | 5 | 2 | 0 | 0 | 58 | 1 | 2 | 726 | 10 |
| VBP | 5 | 3 | 4 | 1 | 1 | 0 | 15 | 0 | 2 | 295 |

שיפורים לויטרבי-

- עברנו לכל u,v רק על הw שעבורם Pi(k,w,u) הוא הגדול ביותר.
- השתמשנו ב multi processing עם 3 תהליכים מקבילים.

דרך לשפר את המודל:

- להתייחס למילה אחרי כדי לשפר את היכולת של המודל להבדיל בין שמות עצם לתארים (NN,JJ). פיצ'ר שכזה הוא הפיצ'ר:

$$f_{107}(h,t) = \begin{cases} 1 & \text{if next word } w_{i+1} = \textit{the} \text{ and } t = \text{Vt} \\ 0 & \text{otherwise} \end{cases}$$

חלק 2:

אימון:

האימון התבצע בצורה זהה לחלק 1 כאשר threshold =1 .

זמן אימון = 535.8619837760925 secs

- לא הוספנו פיצ'רים נוספים חוץ מאלה שהוספנו בחלק הראשון.

מבחן:

את החלק השני בדקנו בעזרת דאטה המבחן מהחלק הראשון עם אותו אלגוריתם הסקה בדיוק.
קיבלנו דיוק של 64% וזהו דיוק טוב יחסית לעובדה שחלק מהטאגים כלל לא קיימים בדאטה האימון.

תחרויות:

- בתחרות הראשונה אנחנו צופים לדיוק הדומה לדיוק בחלק המבחן מכיוון שדאטה האימון בחלק זה גדול מספיק בשביל לכלול לפחות את רוב המילים מהדאטה של התחרות.
- בתחרות השנייה אנחנו צופים לדיוק נמוך משום שדאטה האימון קטן ולכן יש סיכוי רב שהוא לא יכיר חלק גדול מהמילים ואף מהטאגים של המילים בקובץ התחרות.

חלוקת העבודה:

- שלומי עשה את ה pre training .
- את ה training עשינו ביחד.
- אמית עשתה את ההסקה ואת קבצי התחרות.
- כל הקבצים שימשו את שני החלקים.
- את הדוח עשינו ביחד.