דו"ח תרגיל בית 1 עיבוד שפה טבעית

תאיסה שוט סטיינברג - 341241897
רוני כהן - 206134867

אימון

המאפיינים אשר הגדרנו הם:

- $f_{100} - f_{107}$
- f_{number} - המילה מכילה ספרה או מוכלת בקבוצת מילים שמהווה מספר
- f_{allCap} $f_{CapCapCap}$ f_{CapCap} $f_{Capital}$ - מחלקת משפחות עבור אותיות גדולות, כאשר ההבדל בין כל משפחה הוא מספר המילים המתחילות עם אות גדולה ביחס למילה הנוכחית. והמשפחה האחרונה היא של מילים אשר כל האותיות בהן היא אות גדולה
- f_{plural} - המילה מסתיימת באות s, כלומר "רבים"
- $f_{bio\_pre\_suf}$ - יש למילה סיפא/רישא מתוך רשימה של סיפאות/רישאות נפוצות בתחום
- f_{hyfen} - המילה מכילה – (מקף)
- f_{bio_terms} f_{econ_terms} - מופיעה במילה ביטוי מוכר מהדומיין הרלוונטי עבור כל מודל
- f_{dot} - המילה מסתיימת בנקודה

כאשר במודל 1 השתמשנו ב 9784 פרמטרים מהמשפחות: פפיצ'רים מדומיין הכלכלה, אותיות גדולות, מקף, נקודה, רבים, מספרים.
ובמודל 2 השתמשנו ב1942 פרמטרים מהמשפחות: פיצ'רים מדומיין הביולוגיה, רישות וסיפות ביולוגיות, אותיות גדולות, מקפים. בנוסף, מאחר וסט האימון קטן אך מספר המאפיינים האפשריים יחסית גדול, הספים ברובם נמוכים.
עבור כל אחד מהמודלים הגדרנו ספים שונים עבור המאפיינים שציינו מעלה על פי הרלוונטיות וניסוי אמפירי של האלגוריתם.

הסקה

ההסקה התבצעה ע"י אלגוריתם $Viterbi$ עם אופטימיזציה של $Beam\ Search$, השינוי אשר ביצענו באלגוריתם היה בחישוב ה$score$, כך שבמקום כפל אקספוננטים המרנו לחיבור לוגריתמי של אותם הביטויים. השינוי תרם לכך שזמני החישוב התקצרו משמעותית וכן לא נוצר מצב של $stack\ overflow$ בנוסף מאחר ו$Viterbi$ יש צורך להוסיף $padding$ בתחילת המשפט, הורדנו אותם לאחר מכן בפונקציית הtag_all_test.
אחוז הדיוק על קובץ האימון עבור מודל 1 הוא 90.75% ואחוז הדיוק על קובץ האימון עבור מודל 2 הוא 88.3%.

מבחן

אחוז הדיוק של מודל 1 עבור קובץ המבחן הוא:87.9%.

```
Model 1 - test accuracy
Token-level accuracy on test set: 87.88%
```

עבור המודל השני, עקב כמות דוגמאות קטנה באימון, ביצענו $5-fold\ Cross\ Validation$. בחרנו את המודל עם התוצאות הטובות ביותר והשתמשנו בו להסקה. הדבר שיפר משמעותית את היציבות והדיוק של המודל בהשוואה לשימוש במודל יחיד.

תחרות

לאחר ניתוח הטעויות בעזרת $Confusion\ Matrix$, דייקנו את הספים עבור הפיצ'רים שבהם נמצאה רמת בלבול גבוהה. בנוסף, העשרנו את רשימת הסיפות במאפיין f_{101}, לטיפול ספציפי בבלבולים. אנו צופות לקבל אחוז דיוק דומה לזה שהושג בקובץ המבחן עבור המודל הראשון, אך לא גבוה ממנו משמעותית עקב $overfit$ מסוים שנוצר לאור הנתונים המוגבלים שהיו לנו. עבור המודל השני נצפה לראות אחוז דומה לזה שקיבלנו בזמן האימון בתוך ה$Cross\ validation$, כלומר לכל היותר 72%.