NLP HW2 report

Elay Sason 322995358 Ron Dagani 318170917

תהליך ה embedding לדאטא היה זהה בכל המודלים לכן נפרט עליו במשותף ולאחר מכן נפרט על המשך כל מודל בנפרד.

Embedding

ראשית עוברים על המסמך ומחלקים את הדאטא למשפטים תוך הפרדת התגים מהמילים (באשר נצפתה שורה ריקה הופסק הרישום למשפט הקודם והתחיל הרישום למשפט החדש).

לאחר מכן, כל מילה מקבלת ייצוג לפי pre trained glove אליה שורשר ייצוג של המילה שקדמה לה במשפט ושהגיעה אחריה במשפט, מיוצגות גם הן, לפי glove.

במידה והמילה הייתה הראשונה, שורשר וקטור של 1 באורך 200 (אורך היצוג של glove) כמילה הקודמת, על מנת לסמן את תחילת המשפט.

במידה והמילה הייתה המילה האחרונה במשפט, שורשר וקטור של 0 באורך 200 כמילה הבאה של המילה הנוכחית על מנת לסמן את סוף המשפט.

יש לציין כי לפני חיפוש המילה בglove היא עבדה לאותיות קטנות ועברה עיבוד ראשוני שכלל הורדת סימנים וכו' על מנת לאפשר למילה להימצא ללא הפרעות.

במקרה בו המילה לא נמצאה בglove היא קיבלה ייצוג של 'unk' – גם אם מדובר היה במילה הראשית וגם אם היה מדובר במילה כקונטקסט למילה אחרת.

כעת נפרט על המודלים:

המודל הראשון – SVM:

לאחר ניסיון לממש KNN עם 1- 20 שכנים, random forest ומודלים נוספים נמצא כי מודל SMV עם קרנל rbf (Soft) נתן את תוצאת הf1 המיטבית.

אופן מימוש המודל: טעינת הדאטא ועשיית embedding כתואר לעיל, הגדרת מודל SVM עם קרנל rbf, אימון המודל על קובץ האימון ובחינתו על קובץ המבחן.

תוצאת מודל זה :

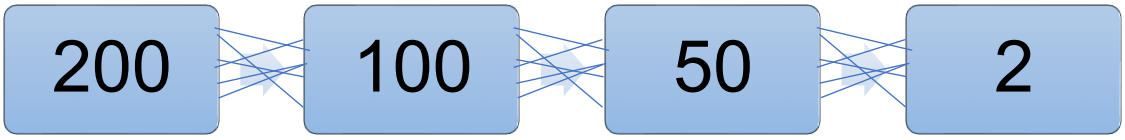
```
svm F1 score:0.529
```

המודל השני והשלישי – רשת fully connected:

המודל השלישי הינו גם המודל השני מכיוון ששניהם מבוססים רשת fc.

נציג את המודל:

המודל הינו רשת fully connected בעלת 3 שכבות כאשר ממדי השכבות הינם:



בין שכבות הרשת הופעלה פונקציית אקטיבציה Relu ובוצע Dropout בשיעור של 0.6.

הדאטא שעבר ברשת חולק לבאצ'ים בגודל 100.

חישוב הloss בוצע באמצעות פונקצית loss המותאמת לבחינת הפסד על F1 שבה הערך לכל batch הוא 1 פחות ממוצע הf1, זאת על מנת לדייק ולאפטם את הscore הסופי. ערך השגיאה מושקל בהתאם לכמות התיוגים 1 שהיו בבאצ', זאת על מנת לעשות מאין Upsampling ותיקון לחוסר האימון בדאטא בכך ששגיאה על 1 תיחשב יותר משגיאה על 0 ובכך ינתן יותר משקל לשגיאה זו למרות המחסור בדוגמאות.

לבסוף בוצע Softmax על מנת לקבל את הסיווג הסופי.

תוצאות מודל משולב זה (2+3) הינו: 0.645

גרסאות קודמות של מודל זה, שתוצאתם הייתה נמוכה יותר הכילו נסיונות שינוי של אחוז הdropout, גודל הבאצ', גודל הhidden dim, גודל המכפיל של הדגימות של 1 (שהוזכר לעיל), סוג פונ' האקטיבציה, סוג פונ' loss, דיוק זיהוי המילים של glove ע"י preprocessing מורכב יותר, שינוי ממדי השכבות ומספרם, שינוי למודלים אחרים בdeep learning ועוד.