NLP – HW2

עומר כהן 308428127

נמרוד אדמוני 203860721

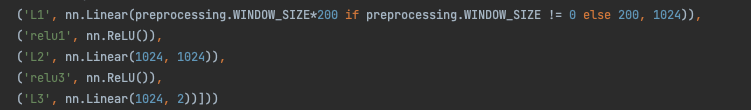
אימון

1. מודל ראשון:

במודל זה השתמשנו בglove-twiter-200- כדי לקבל ווקטורים מהמילים, כל מילה שאין לה קידוד, קידדנו אותנו לווקטור של אפסים. לאחר מכן השתמשנו במסווג SVM כדי לבצע את משימת הסיווג. במהלך התרגיל בדקנו כמה גדלי חלונות, ומצאנו שככל שמגדילים את גודל החלון, הביצועים יורדים. זה מכיוון שמסווגים ליניארים כמו SVM לא מגיבים טוב להגדלה של מימד המאפיינים.

1. מודל שני:

במודל זה השתמשנו באותו מקודד, ושוב קידדנו כווקטור אפס, מילים שהמקודד אינו מכיר. במודל זה השתמשנו בגודל חלון של 1. ארכיטקטורת הרשת שבה השתמשנו כפי שקודדנו בpytorch:



1. במודל זה השתמשנו ברשת עמוקה FC. השתמשנו בגודל חלון של 5, כלומר שתי מילים מימין ומשמאל למילה הנוכחית. במודל זה עשינו עיבוד מקדמים לקובץ האימון, וגם כשנעשה inference נבצע את אותו עיבוד מקדים:
2. מילים שמתחילות ב # ישונו לו hashtag
3. מילים שמתחילות ב $ ישונו לamount
4. מילים שמכילות רק סימוני ניקוד נמחקו
5. מילים שמתחילות ב @ ישונו ל username
6. מילים שמתחילות ב http ישונו ל website
7. מספרים בפורמט של שנה שונו ל year
8. מילים שבפורמט של שעה שונו ל time
9. מספרים שונו ל number
10. מילים של ימות השבוע שונו ל day
11. מילים של חודשי השנה שונו ל month
12. סימני ניקוד בתוך מילה נמחקו
13. אימוגים שונו למילה emoji

ארכיטקטורת הרשת היא:

Graphical user interface, text

Description automatically generated

בנוסף ביצענו חיפוש מעמיק עבור היפר-פרמטרים שנתנו ביצועים טובים על קובץ הdev.

תוצאות על קובץ הdev:

1. מודל ראשון : f1=0.57
2. מודל שני :f1=0.58
3. מודל שלישי: f1=0.68