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 |  |  |

כאשר המאפיינים ממומשים כך:

הוא מאפיין שמזהה שמילה מסוימת היא מספר:

הוא מאפיין שמזהה שהמילה הנוכחית מתחילה באות גדולה:

מספר המאפיינים בכל מודל הם:

|  |  |
| --- | --- |
| מודל 1 (גדול) | מודל 2 (קטן) |
|  | 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

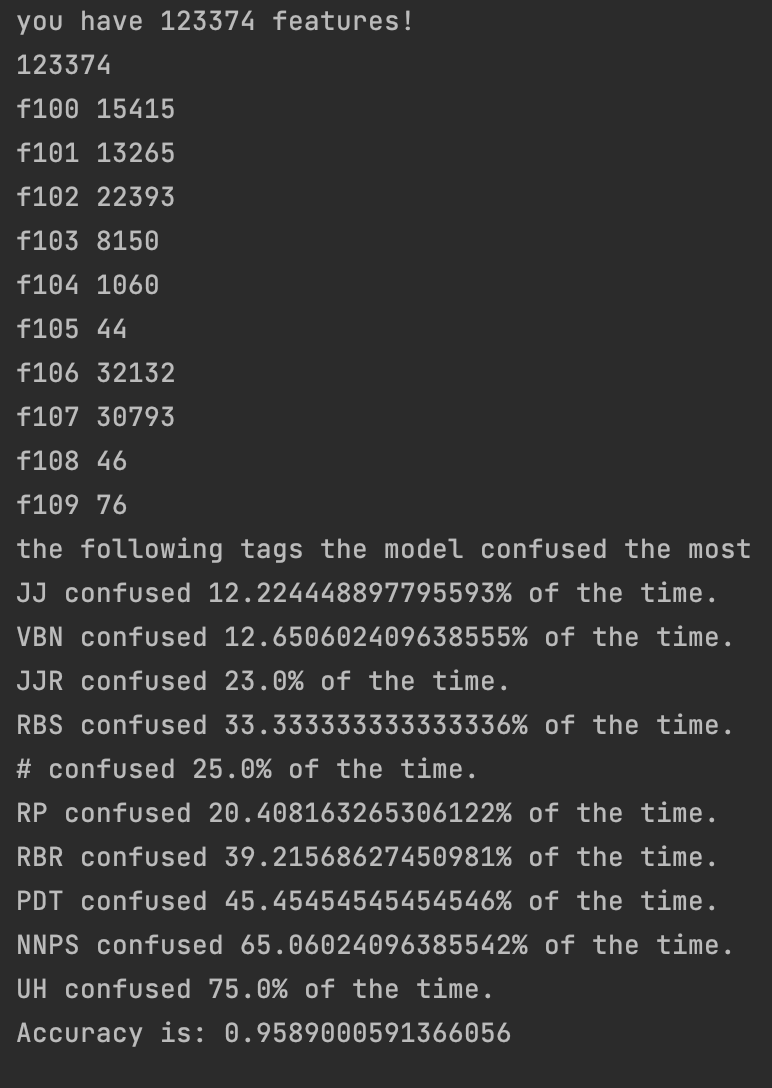
מודל 2 – בנוסף לזיהוי אותיות גדולות ומספרים גם את משפחות המאפיינים 110-113.

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.

הסקה:  
 ההסקה התבצעה על אלגוריתם עם עבור מודל 1 ו עבור מודל 2 ללא שינויים.

מבחן:  
עבור מודל 1 אחוז הדיוק עבור קובץ המבחן הוא 95.8%:



קיבלנו עבור המבחן את ה 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 מ*overfit *על ה-*dataset*. כדי להתמודד עם המחסור ב-*data *להערכת ביצועי המודל, פתרון אפשרי הוא להקצות חלק מה-*data *שברשותנו ל*validation*. לדוגמא, לחלק את ה-*dataset *ל80% אימון ו*20%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.

תחרות:  
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.

חלוקת עבודה:

כל העבודה על תרגיל הבית נעשתה במשותף.