תרגיל בית 1

מגישים: יונתן לוין ותומר כץ.

ת"ז: 213125164 325028967

# אימון

## מודל 1

ניתוח ריצה: (זמן בשניות)



המאפיינים בהם השתמשנו:

f100: כפי שהוגדר בהרצאה. מכיל 15415 מאפיינים.

f101: כפי שהוגדר בהרצאה. מכיל 13265 מאפיינים.

f102: כפי שהוגדר בהרצאה. מכיל 22393 מאפיינים.

f103: כפי שהוגדר בהרצאה. מכיל 8150 מאפיינים.

f104: כפי שהוגדר בהרצאה. מכיל 1060 מאפיינים.

f105: כפי שהוגדר בהרצאה. מכיל 44 מאפיינים.

f106: כפי שהוגדר בהרצאה. מכיל 32132 מאפיינים.

f107: כפי שהוגדר בהרצאה. מכיל 30793 מאפיינים.

capital: מקבל 1 אם המילה הנוכחית מכילה אות גדולה והתג הוא התג הנוכחי. מכיל 76 מאפיינים.

number: מקבל 1 אם המילה הנוכחית מכילה מספר והתג הוא התג הנוכחי. מכיל 50 מאפיינים.

contains hyphen: מקבל 1 אם המילה הנוכחית מכילה '-' והתג הוא התג הנוכחי. מכיל 57 מאפיינים.

pp\_word: מקבל 1 אם 2 מילים אחורה היא מילה מסוימת והתג הוא התג הנוכחי. מכיל 38050 מאפיינים.

f100\_lower: מקבל 1 אם המילה הנוכחית באותיות קטנות היא מילה מסוימת והתג הוא התג הנוכחי. מכיל 14719 מאפיינים.

f101\_lower: כפי שהוגדר בהרצאה, אבל ממירים את המילה הנוכחית לאותיות קטנות. מכיל 12608 מאפיינים.

f102\_lower: כפי שהוגדר בהרצאה, אבל ממירים את המילה הנוכחית לאותיות קטנות. מכיל 20156 מאפיינים.

f106\_lower: כפי שהוגדר בהרצאה, אבל ממירים את המילה הקודמת לאותיות קטנות. מכיל 30860 מאפיינים.

f107\_lower: כפי שהוגדר בהרצאה, אבל ממירים את המילה הבאה לאותיות קטנות. מכיל 30257 מאפיינים.

nn\_word: מקבל 1 אם המילה היא המילה 2 קדימה והתג הוא התג הנוכחי. מכיל 36386 מאפיינים.

c\_word\_n\_word: מקבל 1 אם המילה הראשונה היא המילה הנוכחית, המילה השנייה היא המילה הבאה והתג הוא התג הנוכחי. מכיל 67418 מאפיינים.

length: מקבל 1 אם אורך המילה הוא המספר הנתון. מכיל 305 מאפיינים.

2words: כפי שהוגדר בהרצאה. מכיל 68274 מאפיינים.

3words: כפי שהוגדר בהרצאה. מכיל 102107 מאפיינים.

## מודל 2

ניתוח ריצה: (זמן בשניות)

A black background with white numbers

Description automatically generated

השתמשנו בתת קבוצה מהמאפיינים של המודל הקודם והם:

f100: מכיל 1805 מאפיינים.

f101: מכיל 2556 מאפיינים.

f102: מכיל 3978 מאפיינים.

f103: מכיל 1168 מאפיינים.

f104: מכיל 316 מאפיינים.

f105: מכיל 32 מאפיינים.

f106: מכיל 2801 מאפיינים.

f107: מכיל 2694 מאפיינים.

capital: מכיל 47 מאפיינים.

number: מכיל 36 מאפיינים.

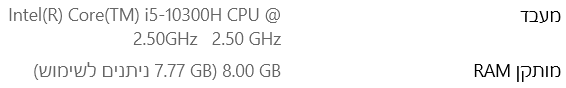
contains hyphen: מכיל 38 מאפיינים.

pp\_word: מכיל 3065 מאפיינים.

length: מכיל 199 מאפיינים.

הורדנו מאפיינים מהמודל השני שהופיעו במודל הראשון כדי למנוע overfit.

## מפרט חומרה



# הסקה

מימשנו Viterbi-beam-search עם beam=3 כפי שנראה בתרגול.

החריגות מייעלות את זמן ריצת האלגוריתם אך לא משפיעות על התוצאה הסופית.

החריגות הן:

## שינוי במבנה הנתונים

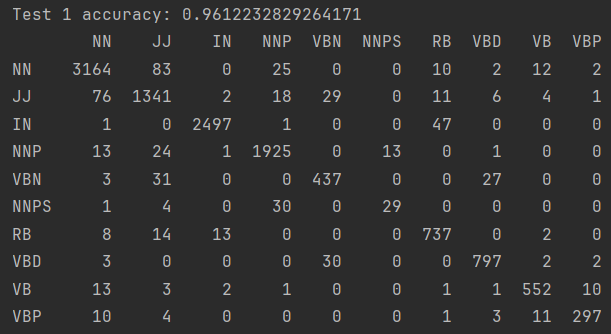
במקום לשמור שני טנזורים כדי לשמור את ההסתברויות והמסלולים, השתמשנו רק בשתי מטריצות דלילות בצורת מילון. המטריצה הראשונה מחזיקה את ה beams של השלב הקודם, עם המסלול אליהם וההסתברות להגיע אליהם. המטריצה השנייה מחזיקה את השיפור הטוב ביותר שניתן לבצע מתיוג קודם לתיוג נוכחי, עם ההסתברות והמסלול אליו.

## שינוי בחיפוש מסלול טוב ביותר

במקום שלכל תא בטנזור נחפש את המקסימום על כל התיוגים 2 אחורה, אנחנו מתחזקים במטריצה השנייה את המסלול הטוב ביותר שנמצא עד אותו שלב.

# מבחן

## מודל 1



כדי להתמודד עם הבלבול בין VB ל NN, ניצור מאפיין שמסתכל על המילה הנוכחית, המילה הבאה והתיוג הנוכחי. שילוב זה אמור לעזור להבחין בין אם המילה הנוכחית היא פועל לבין שם עצם.

## מודל 2

כדי להתמודד עם בעיית חוסר המבחן. ביצענו cross-validation עם 5-folds על קובץ האימון. תוצאות:



כדי להתמודד עם הגודל הקטן של קובץ האימון, הורדנו בחרנו תת קבוצה ממש מהמאפיינים של מודל 1, כדי למנוע overfit.

# תחרות

לא עשינו שינויים למודלים עבור תיוג קבצי התחרות. אנחנו מצפים לקבל דיוק 0.961 עבור המודל הראשון ו 0.933 עבור המודל השני. זה מתבסס על הבדיקה על קובץ המבחן ועל הרצת ה cross-validation בהתאמה. עשויים להיות הבדלים כי המודלים למדו טוב על מילים נדירות ולכן שונות המילים הנדירות בתחרות יכולה להשפיע הרבה על הדיוק.

# חלוקת עבודה

כל העבודה נעשתה ביחד במפגשים פרונטליים ובזום.