עיבוד שפה טבעית - תרגיל בית 1 - רטוב
דו"ח מסכם

רוברט ספקטור 305320400, עופרי אולשביצקי 311287791

מודל 1

אימון -בהסקה על סט האימון השגנו דיוק של 99.5%. חששנו מ overfitting אך על סט המבחן קיבלנו תוצאות טובות. המודל מומש עם סוג המאפיינים $f_{100} - f_{105}$, כאשר עבור המאפיינים f_{101} ו- f_{102} הגדלנו את אורך ה-suffix ו-prefix עד ל 10 אותיות. כמו כן מימשנו את f_{106} , f_{107} כפי שנלמדו בתרגול ובנוסף מימשנו את סוגי המאפיינים הבאים:

- $f_{201} = 1$ if there exists an uppercase letter in current word and $t = V_t$
- $f_{202} = 1$ if there exists a hyphen in current word and $t = V_t$
- $f_{203} = 1$ if there exists a digit in current word and $t = V_t$
- $f_{204} = 1$ if $< current\ word,\ t_{-1},\ t > \ = \ < 'the',\ NN,\ VB >$
- $f_{205} = 1$ if $< current\ word,\ previous\ word,\ t > \ = \ < 'world', 'hello',\ VB >$
- $f_{206} = 1$ if $< current\ word,\ next\ word,\ t > \ = \ < 'hello',\ 'world',\ VB >$
- $f_{207} = 1$ if all letters in current word are uppercase and $t = V_t$
- $f_{208} = 1$ if current word has the characteristic of XXxx_dd and $t = V_t$ where X is an uppercase letter, x is a lowercase letter, _ is _ and d is a digit
- $f_{209} = 1$ if previous word has the characteristic of XXxx_dd and $t = V_t$
- $f_{210} = 1$ if next word has the characteristic of XXxx_dd and $t = V_t$

השראה לחלק מהמאפיינים נלקחו מהמאמרים:

Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network
Part-of-Speech Tagging from 97% to 100%:Is It Time for Some Linguistics?Christopher D. Manning

מצאנו שאפילו מאפיינים שהופיעו פעם אחת בכל הקורפוס תרמו לשיפור ה accuracy על סט המבחן, ההערכה שלנו היא שהמאפיינים האלה תרמו לזיהוי תיוגים של מילים נדירות, לכן כל מאפיין נלקח לאימון אם הוא הופיע לפחות פעם אחת בקורפוס. את הבדיקה ביצענו על הגבלה של 2,5,10 ו-50. מספר המאפיינים מכל סוג:

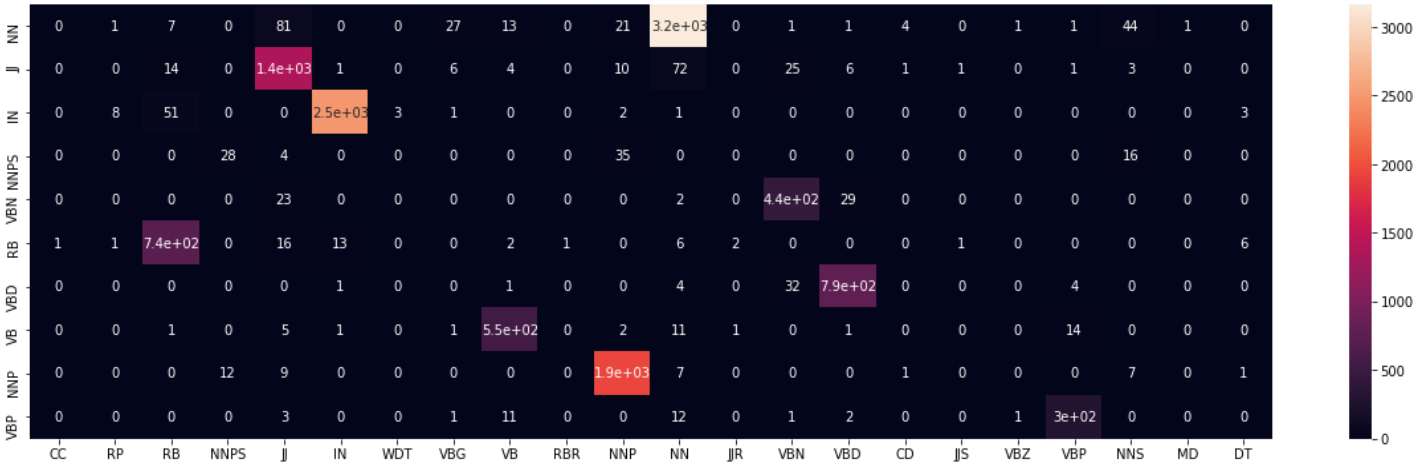
f_{100} - 14719, f_{101} - 58321, f_{102} - 45848, f_{103} - 8150, f_{104} - 1060,
f_{105} - 44, f_{106} - 32132, f_{107} - 30793, f_{201} - 35, f_{202} - 15, f_{203} - 6,
f_{204} - 32976, f_{205} - 68274, f_{206} - 67418, f_{207} - 22, f_{208} - 1025,
f_{209} - 2448, f_{210} - 2465

סה"כ 365,751 מאפיינים נלקחו לאימון. פרמטר הרגולריזציה של אמד לוג הנראות שווה ל1, מצאנו שסטיה מעלה או מטה הורידה את הדיוק על המבחן. את מזעור אמד לוג הנראות ביצענו עם אלגוריתם LBFGS מחבילת scipy. זמן האימון של המודל לקח 8 שעות על מעבד i7-4500U.

הסקה - לא ביצענו שינויים על אלגוריתם ההסקה פרט למימוש beam search. ההסקה בוצעה עם beam width=2 ראינו ש beam רחב יותר תרם מעט לדיוק והעלה את זמן הריצה בהרבה. את הבדיקה על רוחב ה beam ביצענו על קורפוס מבחן מצומצם, לא היה הבדל בתוצאת ההסקה בין beam=3 לbeam=10. תיוג כל המשפטים בקורפוס לקח 29 דקות.

מבחן - על סט המבחן השגנו דיוק של 96.2%. `accuracy = 0.962659457632`

Confusion Matrix



למען נוחות הקריאה הורדנו עמודות של אפסים.

ניתן לראות שהמודל טעה הרבה פעמים בין תיוג JJ לNN כלומר בין שמות עצם לשמות תואר סה"כ 81 פעמים. באנגלית יש שמות תואר מסוימים שיכולים לשמש גם כשמות עצם. למשל המשפט:

"How do you treat your **blind** in your province?"

בדוגמא הספציפית הזאת אם נשמיט את החלק של המשפט אחרי המילה blind, יהיה קשה להבין האם המילה blind היא שם עצם או תואר, כלומר, אחת הדרכים לשפר את הדיוק במקרה כזה היא להתייחס למילים ותיוגים **אחרי** המילה, במודל שלנו השתמשנו במילים שאחרי המילה כמאפיין אך לא השתמשנו בתיוגים - שכן זה דורש אלגוריתם הסקה מסובך יותר.

תחרות - לא ביצענו שינויים למודל בשביל התחרות גם כאן השתמשנו ב beam=3 באלגוריתם ההסקה. למעשה ניסינו להגיע למודל שממקסם את ה accuracy על סט המבחן על מנת להשתמש בו גם לתחרות. מצאנו שכל המאפיינים שמימשנו עזרו לנו על סט המבחן ולכן לא החסרנו אף מאפיין בשביל התחרות.

עשוי להיות הבדל בין הדיוק על קובץ התחרות וקובץ ה test, אם למשל, המשפטים בקובץ האימון וקובץ הבדיקה הגיעו מאותו מקור אך המשפטים בקובץ האימון הן ממקור אחר או אם למשל המילים בקובץ התחרות מגיעים ממאגר מילים הרבה יותר עשיר.
כשעברנו על קובץ התחרות מצאנו הרבה משפטים דומים בקונטקסט ובתחביר לקובץ ה test.
כמו כן מצאנו שיש 1668 מילים בקובץ התחרות שלא נמצאים בקובץ האימון, לעומת 1375 מילים שנמצאים בקובץ המבחן ולא נמצאים בקובץ האימון. כלומר, בקובץ התחרות יש יותר מילים "נדירות". מכיוון שמודל ה MEMM מתקשה עם מילים נדירות, נצפה שהדיוק על קובץ התחרות יהיה פחות טוב - ההערכה שלנו היא 94% דיוק.

מודל 2

תחרות - עבור מודל 2 גם לא ביצענו שינויים על מודל האימון. על מנת לבדוק את עצמנו, תחילה תייגנו את קובץ תחרות 2 בעזרת מודל 1 ועליו עשינו את כל הבדיקות (לאחר שהורדנו תיוגים שלא קיימים במודל 2). מכיוון שקובץ האימון הרבה יותר קטן נצפה לירידה משמעותית באחוז הדיוק - להערכתנו 85% דיוק.

חלוקת העבודה

רוברט - מימוש כל החלק של הלמידה כולל מחלקות הפיצ'רים + דו"ח סופי.
עופרי - מימוש כל החלק של ההסקה - אלגוריתם ויטרבי.
מחקר ומבחן - רוברט ועופרי.

ממשקי הרצה של הקוד

דוגמא להרצה:

```
from MEMM import MEMM

train_file_path = 'train1.wtag'
predict_file_path = 'test1.wtag'

memm = MEMM()
memm.fit(train_file_path)
memm.predict(predict_file_path , beam_width=2)
```