WRITEUP

Or Zinger 200687572

Matan Vetzler

1) עבור מילים נדירות באימון עשינו לקסיקון של קטגוריות-
TWODIGITS: עבור שני ספרות (26)
FORDIGITS: עבור ארבע ספרות (2016)
DIGITS: ספרות (129012438)
containsDigitAndComma: מכיל ספרות פסיק ונקודה (26,000.06)
containsDigitAndPeriod: מכיל ספרות ונקודה (26.16)
containsDigitAndAlpha: מכיל ספרות ואותיות (26A10b26)
containsDigitAndSlash: מכיל ספרות וסלאש (26/06/2016)
capPeriod: כינויי שם (Mr., Mis.)
initCap: מילה מתחילה באות גדולה (Aristo)
allCaps: כל המילה באותיות גדולות (AIRBNB)
portmanteau: הלחמות (give-up)
וחישבנו את מספר ההופעות והצירופים לPOS אחרים שיש לקטגוריות אלו.
בשלב המבחן, מילה שלא נראתה באימון, תמופה באמצעות אותו לקסיקון למילה מהלקסיקון ולה כבר קיימת סטטיסטיקה.

2) האסטרטגיה הייתה לשמור עבור כל מילה שנראתה באימון את רשימת התגיות **שהיא הופיעה איתה**, לשמור את זה בקובץ extra file ובאלגוריתם Viterbi במקום לרוץ על כל התגיות הקיימות שיש רצים רק על התגיות שמשוייכות למילה.

3) מודל HMM VITERBI על *ass1-tagger-dev*:

Accuracy: 94.92%

מודל HMM GREEDY על *ass1-tagger-dev*:

Accuracy: 92.71%

מודל HMM VITERBI על *NER dev*:

Accuracy: 94.63%

Prec: 73.21% **Rec:** 72.41% **F-M:** 72.81%

מודל HMM GREEDY על *NER dev*:

Accuracy: 93.56%

Prec: 63.22% **Rec:** 70.34% **F-M:** 66.62%

maxent-greedy accuracy - 94.37

memm-viterbi accuracy - 96.94

memm-viterbi-ner-per-token accuracy - 94.22

memm-viterbi-ner-span accuracy - 93.47

memm-viterbi-ner-span precision - 57.94

memm-viterbi-ner-span recall - 62.15

memm-viterbi-ner-span F score - 59.97

4) גד

5) יש הבדל בין הDS הקיימים (בין הNER לבין הtagger), ההבדל הוא כמובן בתגיות השונות. בNER יש תגיות המדברות על יישויות (Entities) ובTAGGER יש תגיות המדברות על POS מהמחלקות הפתוחות והסגורות. כמו כן, הבדל נוסף הוא שבNER אנחנו רואים בבירור כי התגית O משוייכת להרבה מילים ומופיעה כ80% מהDS, כלומר זו תגית שלא מייחדת מילה כמו התגיות POS בTAGGER.

6) היות והתיוג O מופיע הכי הרבה פעמים כי הוא תיוג די נפוץ, אולי היינו "מענישים" אותו וממשקלים אותו באיזשהוא משקל כמו שעושים באינטרפולציה (כלומר, בנוסף למשקולים שנותנים באינטרפולציה, להסתברות שיופיע O היינו מכפילים במשקל מסויים) וכך היינו מפחיתים סטטיסטית את השימוש בO ומעלים את הסיכויים של תיוגים אחרים גם.

7)

adding more specific tags can provide more versatile tags set which can provide more information leading to better performances. moreover, looking at bigger window of data will provide more information which can help have better understanding of the model of the data. another option is to add more morphology features to the words to analyze and test which features provide good and useful information.

8)