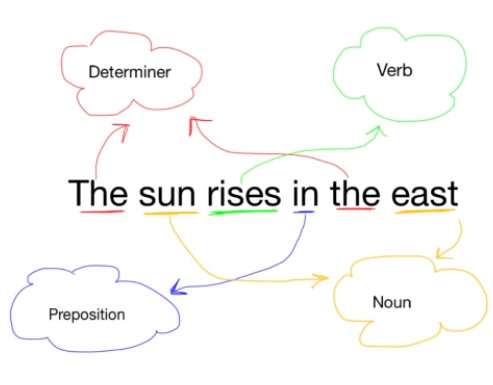
עיבוד שפה טבעית 097215

תרגיל רטוב 1 – Pos Tagging



מיכאל נוביצקי 311773915

עומר אנטורג 201095510

**אימון**

|  |  |  |  |
| --- | --- | --- | --- |
| Count model 2 | Count model 1 on comp 1 | Count model 1 on test1 | Feature |
| *663* | 4263 | 3727 | - |
| *1438* | 6559 | 5997 | – up to length 4 |
| *1950* | 9617 | 8711 | – up to length 4 |
| *578* | 3552 | 3218 | - trigram |
| *210* | 778 | 745 | - bigram |
| *30* | 44 | 44 | - unigram |
| *671* | 4906 | 4165 | - |
| *663* | 5059 | 4307 | - |
| *283* | 1021 | 982 | Skip bigram |
| 1 | 1 | 1 |  |
| 1 | 1 | 1 |  |
| 1 | 1 | 1 |  |
| 1 | 1 | 1 | - series of digits that can contain ‘.’ Or ‘,’ inside |
| 1 | 1 | 1 |  |
| 1 | 1 | 1 |  |
| 6492 | 35805 | 31902 | Total |

**מודל 1**

*פרמטרים נוספים שהגדרנו:*

1. *השתמשנו ב* threshold=3 *כך שכל פיצ'ר שהופיע פחות מ3 פעמים בסט האימון נזרק (חוץ מהפיצ'רים של שאותם לא סיננו). המוטיבציה לשיפור הזה היא ביצוע רגולריזציה של הפיצ'רים שלנו כך שלא נלמד דברים ספציפיים מדי לדאטה של האימון.*
2. *בנינו מילון של מילים וטאגים שהופיעו יותר מ10 פעמים בקורפוס וב 99.5% מהפעמים היה להם את אותו התיוג (השתמשנו במילון לאחר מכן בשלב של ויטרבי). המילון חסך זמני ריצה וגם הוסיף לדיוק של סט המבחן.*
3. *המילה האחרונה במשפט לא נלמדת בשלב האימון והסימנים שסוגרים משפט – ".", "?", "!" נוספו למילון שהגדרנו בסעיף הקודם עם התיוג "." . בשלב של ההסקה, כל הסמלים מהצורה ".", "?", "!" קיבלו אוטומטית את התיוג "." . המוטיבציה לשיפור היא הוספת ידע שלנו על העולם של הבעיה – ראינו שכל הסמלים הללו קיבלו את התיוג "." בסט האימון כשהם הופיעו בסוף משפט.*
4. *השתמשנו ב* lambda=0.7 *עבור רגולריזציית* L2.
5. *שלב ההכנה של הדאטה לאימון מבוצע בצורה מקבילית בעזרת 4 תהליכים.*
6. *ביצענו* grid search *על מרחב פרמטרים עם ערכי* threshold *ו-*lambda *שונים.*

*זמן עיבוד מידע + אימון על* train1 *בלבד – 4996 שניות* ( *כל דיווחי הזמנים הם ביחס להרצות על השרת שקיבלנו בקורס שמכיל 4 ליבות ו16 ג'יגהבייט של* RAM ).

*אחוזי דיוק על סט האימון –* 97.668%.

*זמן עיבוד מידע + אימון על* train1,test1 *–* 5020 *שניות.*

**מודל 2**

השתמשנו באותו סט מאפיינים כמו במודל אחד, אחרי שניסינו להוסיף\להחסיר פיצ'רים נוספים שלא העלו לנו את אחוז הדיוק (פיצ'רים שמסתכלים האם יש סמל במילה, שני טאגים קודמים ומילה נוכחית, וכו').

*פרמטרים נוספים שהגדרנו הם כמו במודל הראשון, מלבד* lambda=0.1, threshold=1.

*זמן עיבוד מידע + אימון – 125 שניות*, *הרצנו על השרת שקיבלנו בקורס עם נתונים שפירטנו במודל 1*.

*אחוזי דיוק על סט ה*validation *של המודל הכי טוב* בCV *-92.86%.*

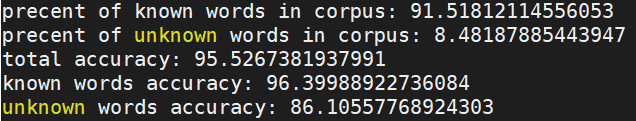
**הסקה**

שינויים שעשינו מהגרסא הרגילה של ויטרבי:

1. על מנת להוריד את סיבוכיות הזמן של ההסקה, מימשנו beam search ובפועל לא ראינו תוספת משמעותית לביצועים כאשר השתמשנו ברוחב אלומה גדול מ2.
2. השתמשנו במילון של מילים וטאגים שעליו הסברנו בסעיף האימון, כל מילה שהופיע במילון זה קיבלה אוטומטית את התיוג שבמילון, המוטבציה לשיפור היא שיפור הדיוק וזמני הריצה של האלגוריתם על מילים שהופיעו יותר מ99.5% מהפעמים עם אותו תיוג.
3. שלב ההסקה מתבצע בצורה מקבילית בעזרת 4 תהליכים.

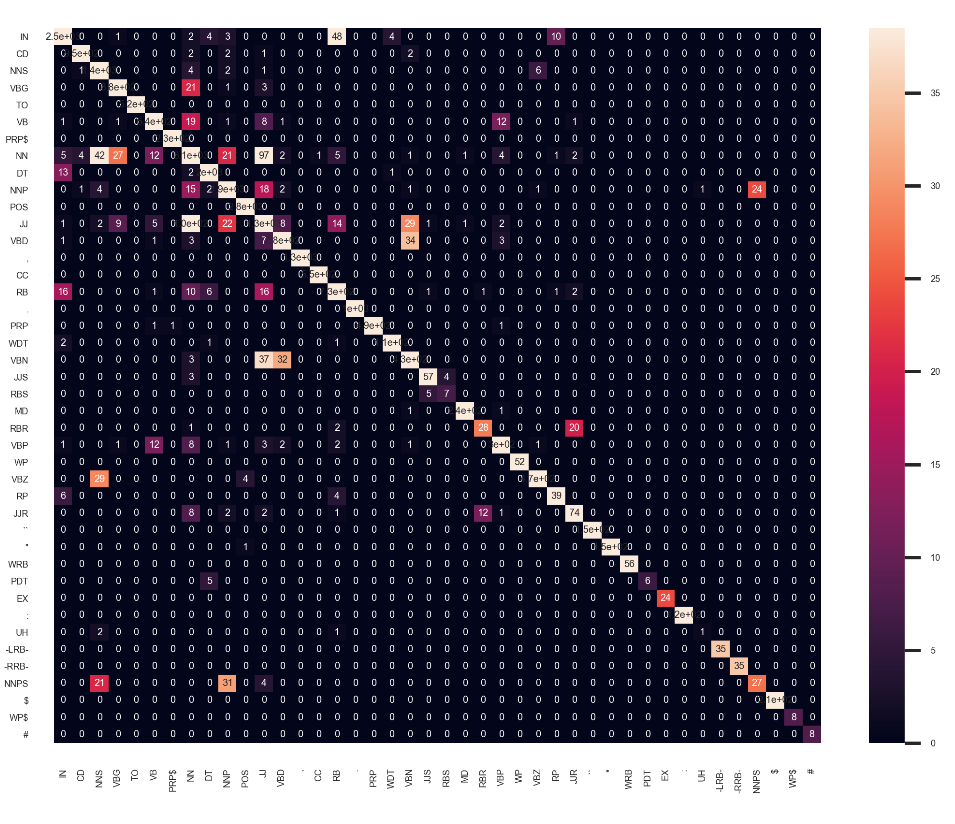
זמן הסקה עבור מודל 1 על test1 – 412 שניות.

**מבחן**



אחוזי הדיוק של מודל1 על test1 – **95.526%**

התמודדות עם מיעוט דוגמאות – כפי שציינו בסעיף התחרות בהמשך, ביצענו CV עם K=5, בחרנו את הפרמטרים שעבורם קיבלנו את הדיוק הכי גבוה, אימנו את המודל בעזרת הפרמטרים הללו וביצענו חיזוי על comp2. הוספת\החסרת פיצ'רים שונים לא עזרה בבדיקות שעשינו.

**מטריצת בלבול**

ניתן לראות שמספר הטעויות השלישי הגדול ביותר הוא כשהמודל חוזה VBN כשבפועל הטאג הוא VBD, כלומר הבלבול הוא בין פועל בזמן past participle לבין פועל שהוא past tense. קיים פיצ'ר (מילה קודמת וטאג נוכחי) שמסתכל על המילה הקודמת ונדלק אם היא בקבוצה הבאה של מילים – {have, was, has} (המילים הללו מופיעות הרבה פעמים לפני מילים שהתיוג שלהם הוא VBN), אבל אנחנו רואים שהוא לא מספיק, לכן אנחנו מציעים להעניק לו משקל גבוה יותר מ-1 כך שהחשיבות שלו בוקטור משקולות תגדל ואנחנו צופים שיהיה שיפור בטעויות מסוג זה.

זמן הסקה עבור מודל 1 על comp1 – 377.06 שניות.

זמן הסקה עבור מודל 2 על comp2 – 117 שניות.

**תחרות**

תחזית עבור מודל 1 – במודל זה כמות הדוגמאות אימון היא גבוהה (5000 משפטים מתוייגים) וגם השתמשנו בסט מבחן (test1) על מנת לבדוק שאנחנו לא ביצענו overfitting ולכן אם הפילוג של comp1 דומה לפילוג של train1, test1 אנחנו צופים לקבל אחוז דיוק גבוה בדומה למה שקיבלנו בסט המבחן (סביב ה95.5%). בנוסף, עבור ביצוע התחזית על סט המבחן comp1, ביצענו אימון גם על train1 וגם על test1 אחרי שראינו שהתחזית על test1 היא טובה ועל מנת להגדיל את השונות של הדאטה של האימון ולתפוס מקרים נוספים שלא ניצפו בtrain1.

תחזית עבור מודל 2 – על מנת להתמודד עם המחסור בכמות הדאטה המתוייג, ניסינו בהתחלה לבצע חיזוי בעזרת המודל הראשון על סט האימון השני על מנת לבדוק אם אחוז הדיוק גבוה ואז נוכל להניח שהפילוג שלהם דומה ולהשתמש בדאטה ממודל 1 בשביל להגדיל את סט האימון. התוצאות שקיבלנו לא היו מספיק טובות (בערך 82%) ונמוכות משמעותית מהתוצאה שקיבלנו כשעשינו תחזית של מודל1 על test1, לכן החלטנו לבדוק שהמודל יודע להכליל לסט מבחן בעזרת ביצוע Cross Validation על סט האימון train2 עם K=5 וביצוע grid search על מרחב הפרמטרים של לאמדה וטרשולד. לאחר מכן בחרנו את המודל עם הפרמטרים הכי טובים שמצאנו ובעזרתו ביצענו חיזוי על comp2. אנחנו צופים שהביצועים של מודל 2 יהיו נמוכים משמעותית מהביצועים של מודל 1 כי כמות הדאטה באימון קטנה הרבה יותר, אבל הם עדין יהיו סבירים כי השתמשנו בCV על מנת לבחור את הפרמטרים. לסיכום בחלק זה אנחנו צופים כ80% דיוק.

**חלוקת עבודה**

מיכאל – עיבוד מקדים, אימון וחלק 1

עומר – ויטרבי, חלק 2 וניתוח תוצאות