תרגיל בית 4 – **H**idden **M**arkov **M**odels

חיים שאללשוילי 200832780

צחי כפיר 200681476

# Accuracy = 0.8159975356812814

חישוב מדד ה-accuracy נעשה באופן הבא:

1. בניית מודל מרקובי בהתבסס על קובץ האימון (training\_pos\_tagging.txt),
2. תיוג קובץ ההערכה (gold\_pos\_tagging.txt) באמצעות המודל המרקובי ואלגוריתם Viterby
3. השוואה בין תיוג האלגוריתם שלנו לבין התיוג הנתון בקובץ ההערכה (gold\_pos\_tagging.txt)
4. חישוב מספר התגים הזהים לחלק לסך כל התגים.

# Confusion Matrix

פלט התוכנית כולל קובץ conf\_matrix.txt, המכיל את confusion matrix בהתאם לקובץ האימון וקובץ ההערכה שהתקבלו כקלט. מצ"ב (נספחים - development confusion matrix) מטריצת הבלבול עבור קובץ האימון וקובץ ההערכה שבעזרתם פותחה התוכנית.

כפי שניתן לראות במטריצה, ברוב הטעויות המסווג בחר לסווג מילים כשמות עצם:

* שמות אשר סווגו כשמות עצם : 230
* שמות תואר אשר סווגו כשמות עצם: 267
* MWE (המוגדר כתואר) אשר סווגו כשמות עצם: 204
* בינוני אשר סווגו כשמות עצם: 159

לדוגמא, סיווג שמות כשמות עצם –שמות הם שמות עצם ולכן הסתברויות המעבר שלהם לשאר סוגי התיוג יהיו דומות. לכן, ובגלל שכמות שמות העצם בטקסט הלימוד היא הגדולה ביותר, הגיוני שהמסווג , הבנוי על מידע סטטיסטי בלבד, יטעה ויסווג שמות רבים כשמות עצם.

שאר טעויות הסיווג נובעות ממילים בעלות מספר סיווגי תיוג בטקסט הלימוד:

לדוגמא : יש לבחור מילה המופיע בקובץ הלימוד גם כשם תואר וגם כשם עצם ולהציגה

מכיוון שכמות שמות העצם בטקסט הלימוד היא הגבוהה ביותר, הגיוני שכאשר המסווג נתקל במילים בעלות כפל סוגי תיוג, המסווג יטעה ויבחר בשם עצם בשל הסתברות המעבר הגבוהה יותר שנוצרת על ידי מודל ה-HMM הנבנה על פי MLE.

1. hmm\_output.txt

כדאי לכם להסתכל על פלט זה על מנת להתרשם מביצועי האלגוריתם

# Baseline

**כיצד הייתם מגדירים baseline של תיוג, שאליו יהיה אפשר להשוות את התוצאות שלכם? דוגמא קיצונית (ולא טובה) לבייסליין היא עבור כל מילה, להגריל תג מתוך קבוצת התגים האפשריים. הציעו דרך טובה יותר לתייג בייסליין (אך אינכם צריכים ליישמה).**

# נספחים:

## Development Confusion Matrix

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Accuracy (success) | 0.84375 | 0 | 1 | 0.55 | 0.823529412 | 0.92 |
| Accuracy (faild) | 0.15625 | 1 | 0 | 0.45 | 0.176470588 | 0.08 |
| Gold Total count | 32 | 3 | 60 | 20 | 136 | 275 |
| Training Total count | 248 | 21 | 343 | 99 | 856 | 1666 |
|  | interrogative | numberexpression | clitic | wprefix | copula | conjunction |
| interrogative |  |  |  |  |  |  |
| numberexpression |  |  |  |  |  |  |
| clitic |  |  |  |  |  |  |
| wprefix |  |  |  |  |  |  |
| copula |  |  |  |  |  |  |
| conjunction |  |  |  |  | 1 |  |
| adjective |  |  |  | 1 |  |  |
| existential |  |  |  |  | 7 |  |
| title |  |  |  |  |  |  |
| mwe | 1 |  |  | 3 |  |  |
| participle |  |  |  |  |  |  |
| adverb |  |  |  |  |  | 9 |
| interjection |  |  |  |  |  |  |
| negation |  |  |  |  |  |  |
| modal |  |  |  |  |  |  |
| unknown |  |  |  |  |  |  |
| propername |  |  |  | 1 |  |  |
| quantifier |  |  |  |  |  |  |
| numeral |  |  |  |  |  |  |
| punctuation |  |  |  |  |  |  |
| noun | 4 | 3 |  | 4 | 1 | 2 |
| verb |  |  |  |  |  |  |
| pronoun |  |  |  |  | 15 | 1 |
| preposition |  |  |  |  |  | 10 |
|  |  |  |  |  |  |  |
| Total | 5 | 3 | 0 | 9 | 24 | 22 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Accuracy (success) | 0.580691643 | 0.759259259 | 0.75 | 0.39703154 | 0.457227139 | 0.848275862 |
| Accuracy (faild) | 0.419308357 | 0.240740741 | 0.25 | 0.60296846 | 0.542772861 | 0.151724138 |
| Gold Total count | 694 | 54 | 8 | 539 | 339 | 290 |
| Training Total count | 4396 | 297 | 51 | 3314 | 2290 | 1901 |
|  | adjective | existential | title | mwe | participle | adverb |
| interrogative |  |  |  | 3 |  | 1 |
| numberexpression |  |  |  |  |  |  |
| clitic |  |  |  |  |  | 1 |
| wprefix |  |  |  | 1 |  |  |
| copula |  | 12 |  |  |  |  |
| conjunction |  |  |  | 12 |  | 2 |
| adjective |  |  |  | 14 | 5 | 4 |
| existential |  |  |  | 2 |  |  |
| title |  |  |  |  |  |  |
| mwe | 5 |  |  |  | 2 | 8 |
| participle | 5 |  |  | 7 |  |  |
| adverb | 1 |  |  | 17 |  |  |
| interjection |  |  |  |  |  | 1 |
| negation |  |  |  | 7 |  |  |
| modal | 4 | 1 |  | 1 |  | 1 |
| unknown |  |  |  |  |  |  |
| propername | 4 |  |  | 5 | 1 | 1 |
| quantifier |  |  |  | 6 | 1 | 1 |
| numeral |  |  |  | 6 |  |  |
| punctuation |  |  |  |  |  |  |
| noun | 267 |  | 2 | 204 | 159 | 22 |
| verb | 5 |  |  | 6 | 15 | 1 |
| pronoun |  |  |  | 3 |  |  |
| preposition |  |  |  | 31 | 1 | 1 |
|  |  |  |  |  |  |  |
| Total | 291 | 13 | 2 | 325 | 184 | 44 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Accuracy (success) | 0 | 0.91011236 | 0.777777778 | 0 | 0.574074074 | 0.788888889 |
| Accuracy (faild) | 1 | 0.08988764 | 0.222222222 | 1 | 0.425925926 | 0.211111111 |
| Gold Total count | 1 | 89 | 63 | 7 | 594 | 90 |
| Training Total count | 19 | 539 | 373 | 63 | 3564 | 592 |
| Column1 | interjection | negation | modal | unknown | propername | quantifier |
| interrogative |  |  |  |  |  | 1 |
| numberexpression |  |  |  |  |  |  |
| clitic |  |  |  |  | 1 |  |
| wprefix |  |  |  |  | 1 |  |
| copula |  |  |  |  |  |  |
| conjunction |  |  |  |  | 2 | 2 |
| adjective |  |  | 2 |  | 2 |  |
| existential |  | 6 | 2 |  |  |  |
| title |  |  |  |  |  |  |
| mwe |  |  |  |  | 10 | 2 |
| participle |  |  | 1 |  |  |  |
| adverb | 1 |  | 1 |  | 1 | 4 |
| interjection |  |  |  |  |  |  |
| negation |  |  | 1 |  |  |  |
| modal |  |  |  |  |  |  |
| unknown |  |  |  |  |  |  |
| propername |  |  |  | 2 |  |  |
| quantifier |  |  |  |  |  |  |
| numeral |  |  |  |  | 4 |  |
| punctuation |  |  |  |  |  |  |
| noun |  |  | 5 | 5 | 230 | 6 |
| verb |  |  | 1 |  | 2 |  |
| pronoun |  |  |  |  |  |  |
| preposition |  | 2 | 1 |  |  | 4 |
|  |  |  |  |  |  |  |
| Total | 1 | 8 | 14 | 7 | 253 | 19 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Accuracy (success) | 0.886363636 | 1 | 0.954845815 | 0.572864322 | 0.87654321 | 0.938042131 |
| Accuracy (faild) | 0.113636364 | 0 | 0.045154185 | 0.427135678 | 0.12345679 | 0.061957869 |
| Gold Total count | 220 | 1601 | 2724 | 796 | 243 | 807 |
| Training Total count | 1601 | 9531 | 17357 | 4654 | 1479 | 4750 |
| Column1 | numeral | punctuation | noun | verb | pronoun | preposition |
| interrogative |  |  | 1 |  | 2 |  |
| numberexpression |  |  |  |  |  |  |
| clitic |  |  | 1 |  |  |  |
| wprefix |  |  |  |  |  | 1 |
| copula |  |  |  |  | 8 |  |
| conjunction |  |  | 3 |  | 1 | 10 |
| adjective |  |  | 14 |  |  | 3 |
| existential |  |  |  |  |  |  |
| title |  |  | 8 |  |  |  |
| mwe | 3 |  | 43 | 3 | 6 | 8 |
| participle |  |  | 15 | 6 |  |  |
| adverb |  |  | 9 |  | 2 |  |
| interjection |  |  |  |  |  |  |
| negation |  |  |  |  |  |  |
| modal |  |  |  |  |  |  |
| unknown |  |  |  |  |  |  |
| propername | 1 |  | 13 | 3 |  | 3 |
| quantifier |  |  | 2 | 1 | 6 |  |
| numeral |  |  |  |  |  |  |
| punctuation |  |  |  |  |  |  |
| noun | 21 |  |  | 321 | 5 | 24 |
| verb |  |  | 12 |  |  | 1 |
| pronoun |  |  |  |  |  |  |
| preposition |  |  | 2 | 6 |  |  |
|  |  |  |  |  |  |  |
| Total | 25 | 0 | 123 | 340 | 30 | 50 |