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.

**נתחו את סוגי הטעויות העיקריים** - הסבירו מדוע דווקא הן קורות.

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) | 0.97983871 | 0.857142857 | 1 | 0.909090909 | 0.971962617 | 0.986794718 |
| Accuracy (faild) | 0.02016129 | 0.142857143 | 0 | 0.090909091 | 0.028037383 | 0.013205282 |
| Total count | 248 | 21 | 343 | 99 | 856 | 1666 |
| Column1 | 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) | 0.933803458 | 0.956228956 | 0.960784314 | 0.901931201 | 0.919650655 | 0.976854287 |
| Accuracy (faild) | 0.066196542 | 0.043771044 | 0.039215686 | 0.098068799 | 0.080349345 | 0.023145713 |
| Total count | 4396 | 297 | 51 | 3314 | 2290 | 1901 |
| Column1 | 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) | 0.947368421 | 0.985157699 | 0.962466488 | 0.888888889 | 0.929012346 | 0.967905405 |
| Accuracy (faild) | 0.052631579 | 0.014842301 | 0.037533512 | 0.111111111 | 0.070987654 | 0.032094595 |
| 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) | 0.98438476 | 1 | 0.992913522 | 0.926944564 | 0.979716024 | 0.989473684 |
| Accuracy (faild) | 0.01561524 | 0 | 0.007086478 | 0.073055436 | 0.020283976 | 0.010526316 |
| 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 |