Program 3 Report

1. **Unseen observation:** Assign a very small probability (p = 0.000000000001) to it.
2. **Transition bigram smoothing:** Assign a very small count to unseen transition bigrams. I modified plus one smoothing into plus a smaller number (0.000000000001). The formula of smoothing:
   P (tagi|tagi-1) = count(tagi-1, tagi) + 0.000000000001 / count(tagi-1) + 0.000000000001 * V
   The probability for unseen transition:
   P (tagi|tagi-1)unseen = 0.000000000001 / count(tagi-1) + 0.000000000001 * V
3. **Access Viterbi program:** Use "*yuanchai_program3.py"* to assign POS to a file:
   python yuanchai_program3.py <training file name> <test file name>
   (**Note:** It takes about a minute to process the given test file)
4. **Program assessment:**

Training and Development File. I assess this program by dividing the original file *"berp-POS-train.txt"* into a training file and a development file using a separate program *"split.py"*. One out of every five sentences are pulled out from the original file and written into a file as the gold standard of the development file. All the tags are stripped from the gold standard file and the observations are written into a development file with one word a line, sentences separated by a line. The remaining sentences are written into a training file. *"split.py"* yields a training file, a development file, and a gold standard file. The program then uses the training file to set parameters and assigns POS to the development file. The observations and the tags of development file are written into *"output_yuanchai.txt"*.

Test File. The whole original file *"berp-POS-train.txt"* is used as training file. The program assigns POS to the test file *"berp-POS-test-sentences.txt"*. *"berp-POS-test-gold.txt"* is the gold standard file which the output would be compared with.

Evaluation Program. An evaluation program *"evalPOS.py"* is built to see how well the tagging task is done. The evaluation program would produce: 1) An overall accuracy; 2) The precision and recall accuracy of each tag produced by the **system**; 3) A confusion matrix for each tag in the **gold standard**.

Access Evaluation Program:

python evalPOS.py <gold standard file name> <output file name>

Format of Evaluation Results:

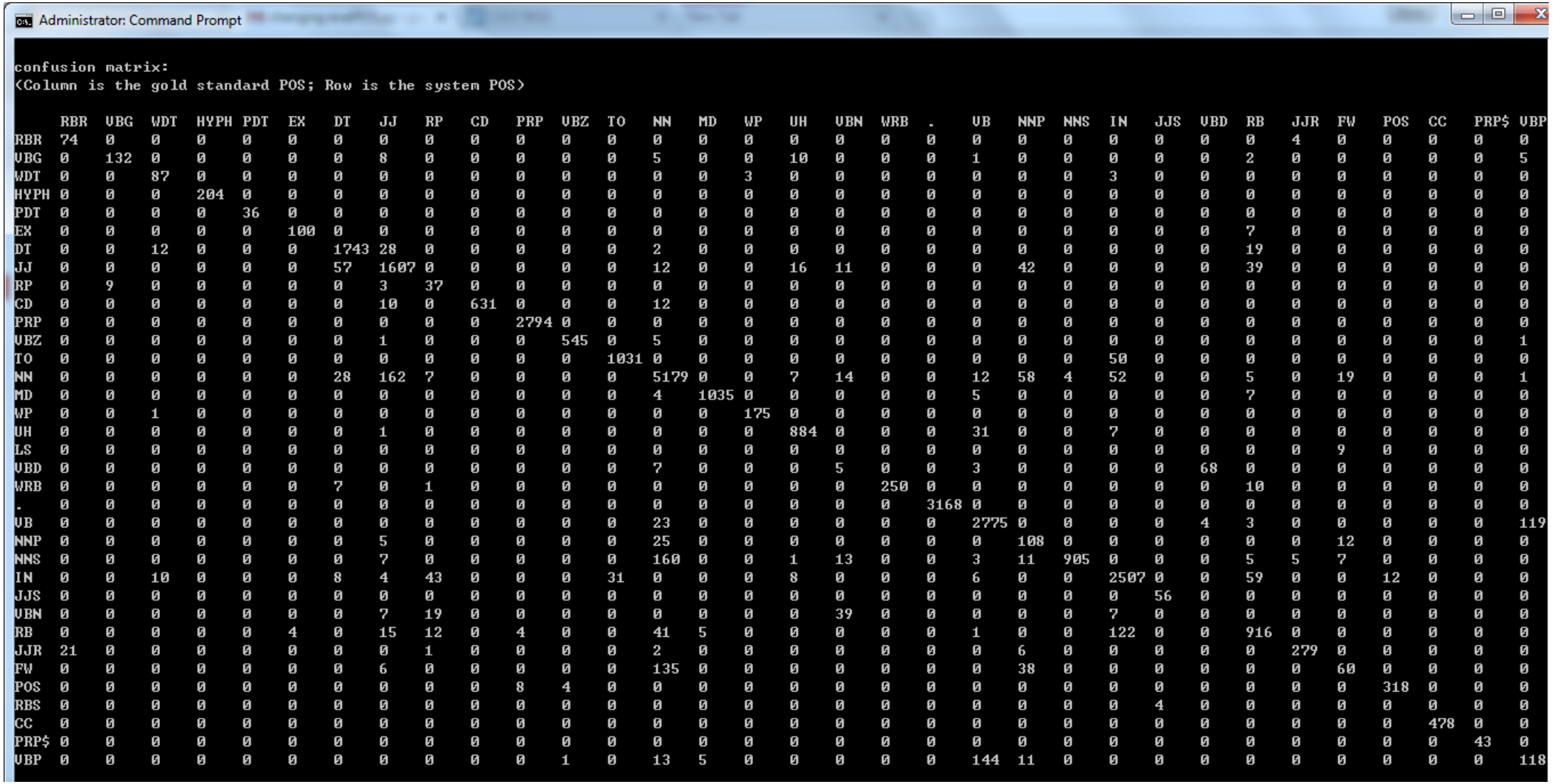
1) **Overall accuracy:**
   Accuracy = Count of correct POS / Count of all the POS
2) **Per POS class error rate:**
   The precision and recall accuracy for each POS class assigned by the system is calculated.
   Use "NNS" as an example:
   Precision = Count of correct "NNS" captured by the system / Count of all "NNS" captured by the system
   Recall = Count of correct "NNS" captured by the system / Count of all "NNS" in gold standard
3) **Confusion matrix:**
   A confusion matrix is built for each mistagged POS in the gold standard. The column is the gold standard POS; the row is the POS produced by system.
   (Note: For PC, if the command prompt window is not wide enough to show the entire matrix, please right click on the buffer window, choose "property → layout → window width" and set width as 180.)

5. **How well the Assessment Works:** It provides a thorough reflection of the accuracy of the entire program and each POS. The confusion matrix could be used to tune the parameters of the program.

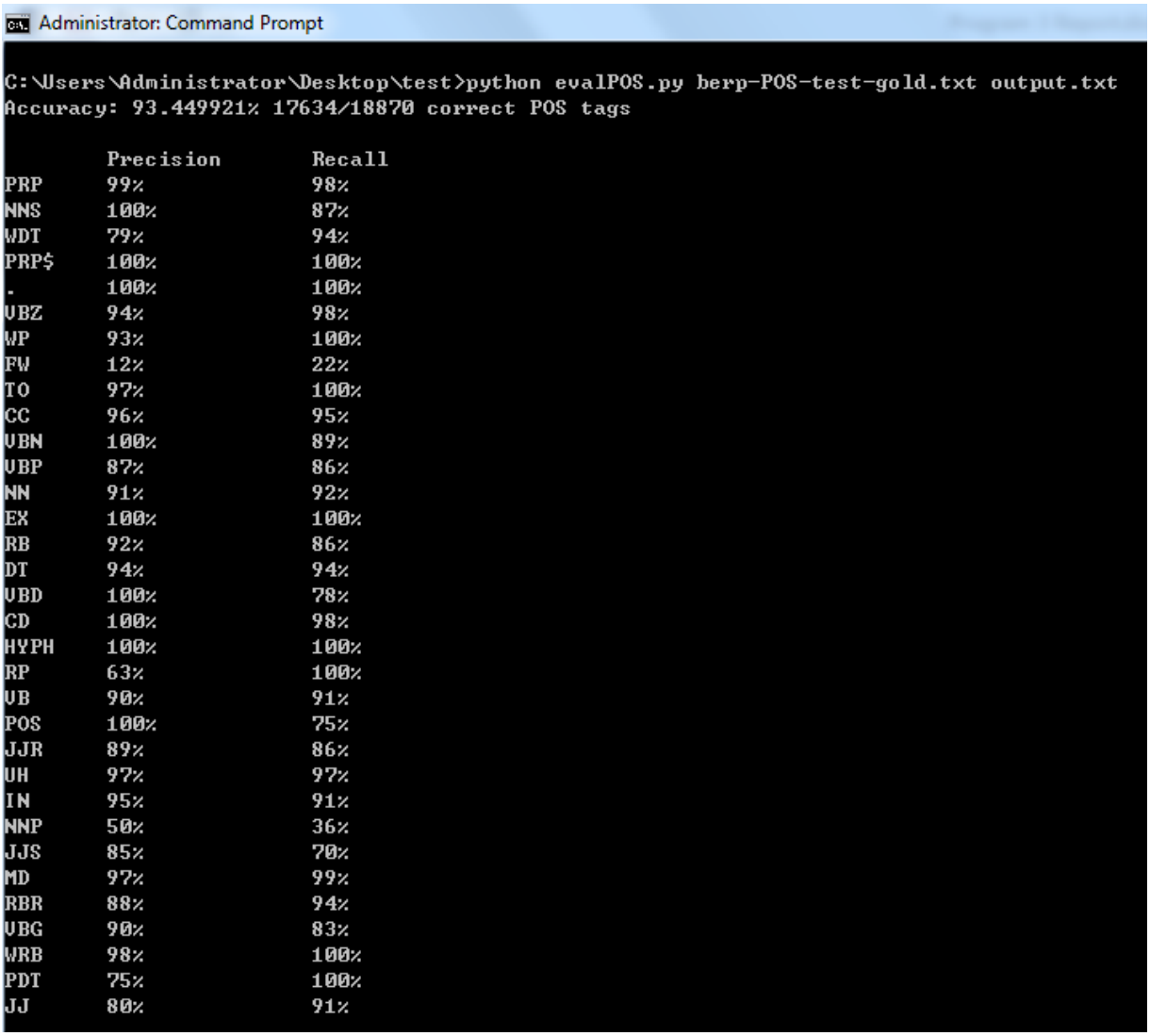
The result of training/development file evaluation is as follow:

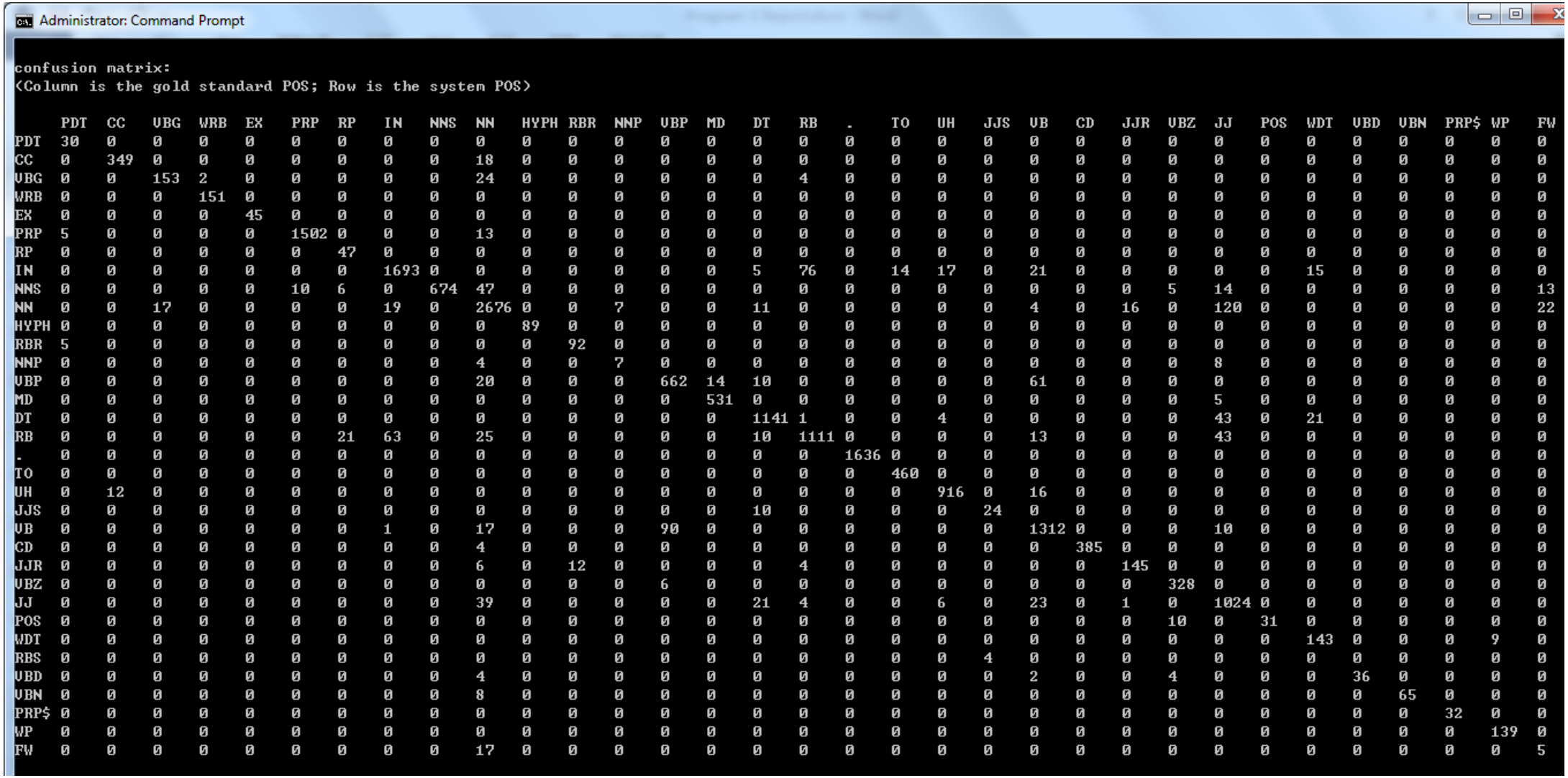
```
C:\Users\Administrator\Desktop\test>python evalPOS.py standard.txt output.txt
Accuracy: 93.450530% 29450/31514 correct POS tags

        Precision       Recall
POS     96%             96%
VBG     93%             80%
VBN     47%             54%
JJ      86%             90%
JJR     96%             90%
PRP$    100%            100%
FW      56%             25%
MD      99%             98%
RP      30%             75%
UH      95%             95%
NN      92%             93%
JJS     93%             100%
IN      91%             93%
VBZ     99%             98%
NNP     39%             72%
VBP     90%             87%
NNS     99%             81%
WP      98%             99%
HYPH    100%            100%
VBD     94%             81%
TO      97%             95%
PRP     99%             100%
WDT     79%             93%
PDT     100%            100%
DT      94%             96%
WRB     100%            93%
CD      100%            96%
RBR     77%             94%
VB      93%             94%
EX      96%             93%
RB      85%             81%
.       100%            100%
CC      100%            100%
```


Administrator: Command Prompt
confusion matrix:
(Column is the gold standard POS; Row is the system POS)
RBR 74 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 4 0 0 0 0 0
VBG 0 132 0 0 0 0 0 8 0 0 0 0 0 5 0 0 10 0 0 0 1 0 0 0 0 0 2 0 0 0 0 0 5
WDT 0 0 87 0 0 0 0 0 0 0 0 0 0 0 0 3 0 0 0 0 0 0 0 3 0 0 0 0 0 0 0 0 0
HYPH 0 0 0 204 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
PDT 0 0 0 0 36 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
EX 0 0 0 0 0 100 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 7 0 0 0 0 0 0
DT 0 0 12 0 0 0 1743 28 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 19 0 0 0 0 0 0
JJ 0 0 0 0 0 0 57 1607 0 0 0 0 0 12 0 0 16 11 0 0 0 42 0 0 0 0 39 0 0 0 0 0 0
RP 0 9 0 0 0 0 0 3 37 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
CD 0 0 0 0 0 0 0 10 0 631 0 0 0 12 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
PRP 0 0 0 0 0 0 0 0 0 0 2794 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
VBZ 0 0 0 0 0 0 0 1 0 0 0 545 0 5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
TO 0 0 0 0 0 0 0 0 0 0 0 0 1031 0 0 0 0 0 0 0 0 0 0 0 50 0 0 0 0 0 0 0 0
NN 0 0 0 0 0 0 28 162 7 0 0 0 0 5179 0 0 7 14 0 0 12 58 4 52 0 0 5 0 19 0 0 0 1
MD 0 0 0 0 0 0 0 0 0 0 0 0 0 4 1035 0 0 0 0 0 5 0 0 0 0 0 7 0 0 0 0 0 0
WP 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 175 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
UH 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 884 0 0 0 31 0 0 7 0 0 0 0 0 0 0 0 0
LS 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 9 0 0 0 0
VBD 0 0 0 0 0 0 0 0 0 0 0 0 0 7 0 0 0 5 0 0 3 0 0 0 0 68 0 0 0 0 0 0 0
WRB 0 0 0 0 0 0 7 0 1 0 0 0 0 0 0 0 0 0 250 0 0 0 0 0 0 0 10 0 0 0 0 0 0
. 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 3168 0 0 0 0 0 0 0 0 0 0 0 0 0
VB 0 0 0 0 0 0 0 0 0 0 0 0 0 23 0 0 0 0 0 0 2775 0 0 0 0 4 3 0 0 0 0 0 119
NNP 0 0 0 0 0 0 0 5 0 0 0 0 0 25 0 0 0 0 0 0 0 108 0 0 0 0 0 0 12 0 0 0 0
NNS 0 0 0 0 0 0 0 7 0 0 0 0 0 160 0 0 1 13 0 0 3 11 905 0 0 0 5 5 7 0 0 0 0
IN 0 0 10 0 0 0 8 4 43 0 0 0 31 0 0 0 8 0 0 0 6 0 0 2507 0 0 59 0 0 12 0 0 0
JJS 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 56 0 0 0 0 0 0 0 0
VBN 0 0 0 0 0 0 0 7 19 0 0 0 0 0 0 0 0 39 0 0 0 0 0 7 0 0 0 0 0 0 0 0 0
RB 0 0 0 0 0 4 0 15 12 0 4 0 0 41 5 0 0 0 0 0 1 0 0 122 0 0 916 0 0 0 0 0 0
JJR 21 0 0 0 0 0 0 0 1 0 0 0 0 2 0 0 0 0 0 0 0 6 0 0 0 0 0 279 0 0 0 0 0
FW 0 0 0 0 0 0 0 6 0 0 0 0 0 135 0 0 0 0 0 0 0 38 0 0 0 0 0 0 60 0 0 0 0
POS 0 0 0 0 0 0 0 0 0 0 8 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 318 0 0 0
RBS 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 4 0 0 0 0 0 0 0
CC 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 478 0 0
PRP$ 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 43 0
VBP 0 0 0 0 0 0 0 0 0 0 0 1 0 13 5 0 0 0 0 0 144 11 0 0 0 0 0 0 0 0 0 0 118

The result for test file evaluation is as follow:




Administrator: Command Prompt
confusion matrix:
(Column is the gold standard POS; Row is the system POS)
PDT CC VBG WRB EX PRP RP IN NNS NN HYPH RBR NNP VBP MD DT RB . TO UH JJS VB CD JJR VBZ JJ POS WDT VBD VBN PRP$ WP FW
PDT 30 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
CC 0 349 0 0 0 0 0 0 0 18 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
VBG 0 0 153 2 0 0 0 0 0 24 0 0 0 0 0 0 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0
WRB 0 0 0 151 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
EX 0 0 0 0 45 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
PRP 5 0 0 0 0 1502 0 0 0 13 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
RP 0 0 0 0 0 0 47 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
IN 0 0 0 0 0 0 0 1693 0 0 0 0 0 0 0 5 76 0 14 17 0 21 0 0 0 0 15 0 0 0 0 0
NNS 0 0 0 0 0 10 6 0 674 47 0 0 0 0 0 0 0 0 0 0 0 0 0 0 5 14 0 0 0 0 0 0 13
NN 0 0 17 0 0 0 0 19 0 2676 0 0 7 0 0 11 0 0 0 0 0 4 0 16 0 120 0 0 0 0 0 0 22
HYPH 0 0 0 0 0 0 0 0 0 0 89 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
RBR 5 0 0 0 0 0 0 0 0 0 0 92 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
NNP 0 0 0 0 0 0 0 0 0 4 0 0 7 0 0 0 0 0 0 0 0 0 0 0 0 8 0 0 0 0 0 0 0
VBP 0 0 0 0 0 0 0 0 0 20 0 0 0 662 14 10 0 0 0 0 0 61 0 0 0 0 0 0 0 0 0 0
MD 0 0 0 0 0 0 0 0 0 0 0 0 0 0 531 0 0 0 0 0 0 0 0 0 0 5 0 0 0 0 0 0 0
DT 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1141 1 0 0 4 0 0 0 0 0 43 0 21 0 0 0 0 0
RB 0 0 0 0 0 0 21 63 0 25 0 0 0 0 0 10 1111 0 0 0 0 13 0 0 0 43 0 0 0 0 0 0 0
. 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1636 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
TO 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 460 0 0 0 0 0 0 0 0 0 0 0 0 0 0
UH 0 12 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 916 0 16 0 0 0 0 0 0 0 0 0 0 0
JJS 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 10 0 0 0 0 24 0 0 0 0 0 0 0 0 0 0 0 0
VB 0 0 0 0 0 0 0 1 0 17 0 0 0 90 0 0 0 0 0 0 0 0 1312 0 0 0 10 0 0 0 0 0 0 0
CD 0 0 0 0 0 0 0 0 0 4 0 0 0 0 0 0 0 0 0 0 0 0 0 385 0 0 0 0 0 0 0 0 0 0
JJR 0 0 0 0 0 0 0 0 0 6 0 12 0 0 0 0 4 0 0 0 0 0 0 0 145 0 0 0 0 0 0 0 0 0
VBZ 0 0 0 0 0 0 0 0 0 0 0 0 0 6 0 0 0 0 0 0 0 0 0 0 0 328 0 0 0 0 0 0 0 0
JJ 0 0 0 0 0 0 0 0 0 39 0 0 0 0 0 0 21 4 0 0 6 0 23 0 1 0 1024 0 0 0 0 0 0 0
POS 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 10 0 31 0 0 0 0 0 0
WDT 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 143 0 0 0 9 0
RBS 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 4 0 0 0 0 0 0 0 0
VBD 0 0 0 0 0 0 0 0 0 4 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 4 0 0 0 36 0 0 0 0
VBN 0 0 0 0 0 0 0 0 0 8 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 65 0 0 0
PRP$ 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 32 0 0
WP 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 139 0
FW 0 0 0 0 0 0 0 0 0 17 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 5