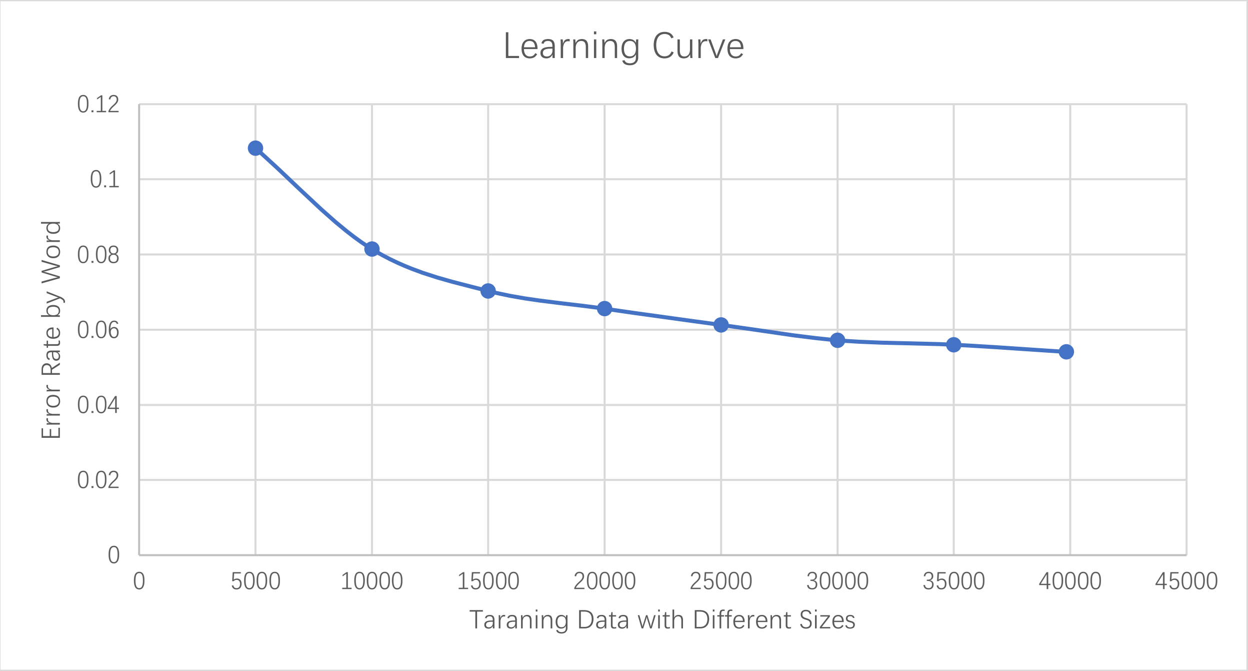
Richard Xie
915505564

Task 1:

Output with different training data sizes:

Size	Error rate by word
5000	0.108258
10000	0.081437
15000	0.070294
20000	0.065583
25000	0.061271
30000	0.057183
35000	0.055986
39832	0.054092



Thoughts:

It will take a longer and longer time for the system to apply Viterbi algorithm to the POS-tagged data when it gets larger and larger.

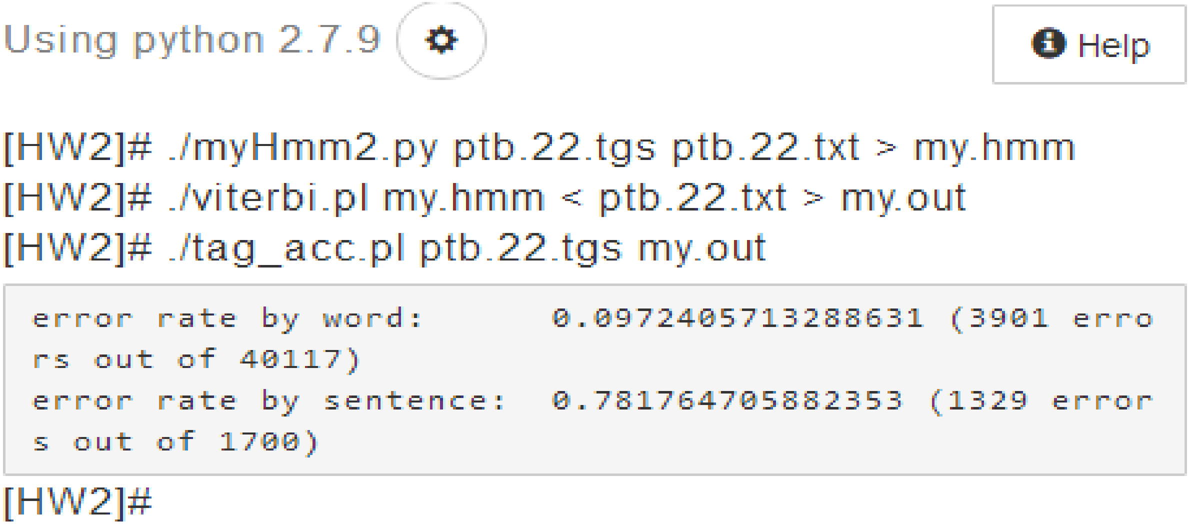
Reference: Post @147 on Piazza.

Task 2:

Applied add-k smoothing method to the original train_hmm.py file, renamed to myHMM.py

Reference: https://www.youtube.com/watch?v=GwP8gKa-ij8

Performance on ptb.22.* :



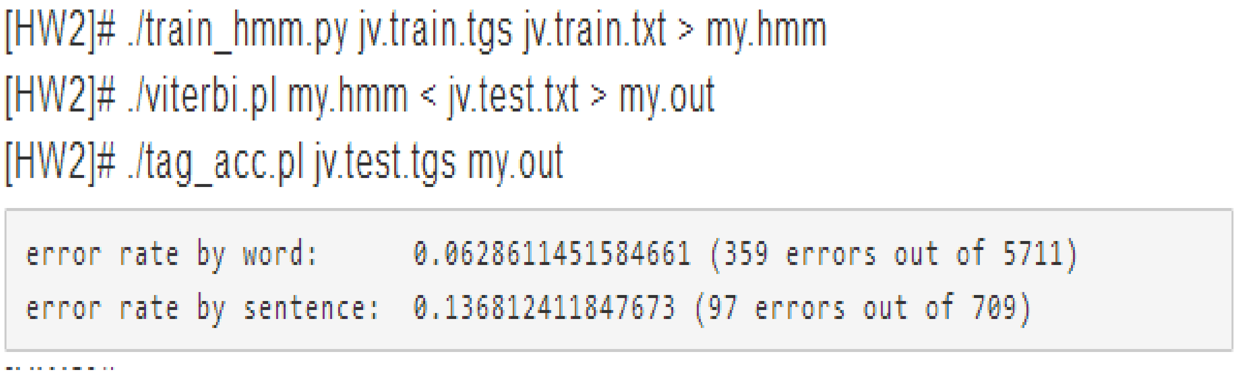
```
[HW2]# ./myHmm2.py ptb.22.tgs ptb.22.txt > my.hmm
[HW2]# ./viterbi.pl my.hmm < ptb.22.txt > my.out
[HW2]# ./tag_acc.pl ptb.22.tgs my.out
error rate by word:      0.0972405713288631 (3901 erro
rs out of 40117)
error rate by sentence:  0.781764705882353 (1329 error
s out of 1700)
[HW2]#
```

Can’t show the difference since ptb.23.tag is not provided.
Can be tested using myHmm2.py and the default viterbi.pl

Task 3:

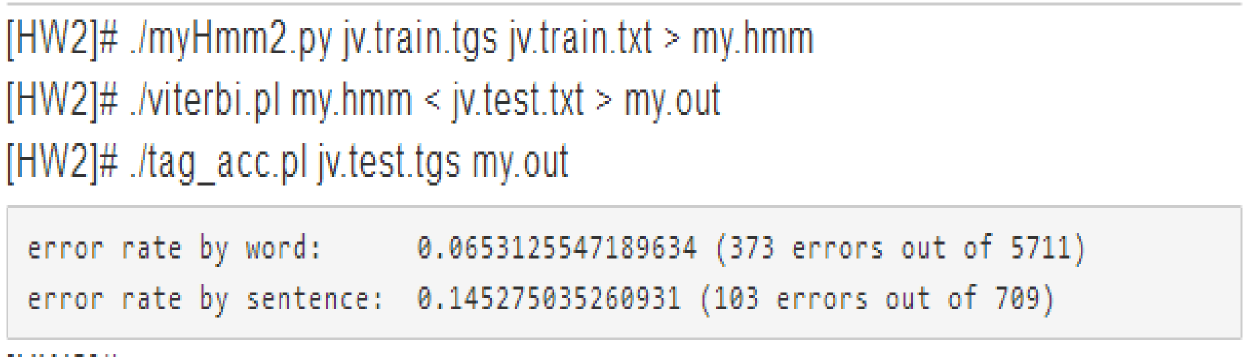
Japanese:

Baseline Model:



```
[HW2]# ./train_hmm.py jv.train.tgs jv.train.txt > my.hmm
[HW2]# ./viterbi.pl my.hmm < jv.test.txt > my.out
[HW2]# ./tag_acc.pl jv.test.tgs my.out
error rate by word:      0.0628611451584661 (359 errors out of 5711)
error rate by sentence:  0.136812411847673 (97 errors out of 709)
```

My Model:



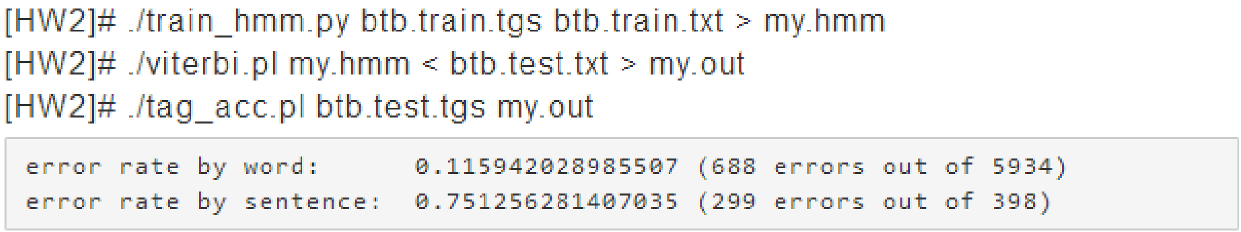
```
[HW2]# ./myHmm2.py jv.train.tgs jv.train.txt > my.hmm
[HW2]# ./viterbi.pl my.hmm < jv.test.txt > my.out
[HW2]# ./tag_acc.pl jv.test.tgs my.out
```

```
error rate by word:      0.0653125547189634 (373 errors out of 5711)
error rate by sentence:  0.145275035260931 (103 errors out of 709)
```

For Japanese, the error rate is typically lower because the language is written in Hiragana, which can be expressed using alphabet characters. I think if the text was written in Japanese characters, the error rate would become much higher.

Bulgarian:

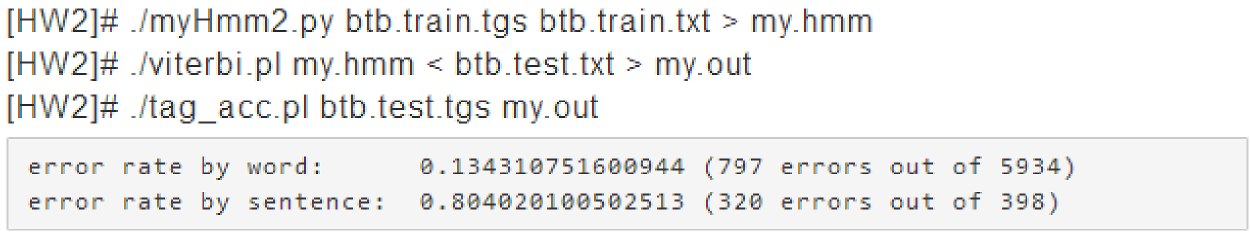
Base Model:



```
[HW2]# ./train_hmm.py btb.train.tgs btb.train.txt > my.hmm
[HW2]# ./viterbi.pl my.hmm < btb.test.txt > my.out
[HW2]# ./tag_acc.pl btb.test.tgs my.out
```

```
error rate by word:      0.115942028985507 (688 errors out of 5934)
error rate by sentence:  0.751256281407035 (299 errors out of 398)
```

My Model



```
[HW2]# ./myHmm2.py btb.train.tgs btb.train.txt > my.hmm
[HW2]# ./viterbi.pl my.hmm < btb.test.txt > my.out
[HW2]# ./tag_acc.pl btb.test.tgs my.out
```

```
error rate by word:      0.134310751600944 (797 errors out of 5934)
error rate by sentence:  0.804020100502513 (320 errors out of 398)
```

As for Bulgarian, there are more special characters in the language, which means that the tagger for English may not be a be able to tag most of the characters. Therefore, generating high error rates.