Homework 4

1. Hidden Markov Model

a. training data

```
# tokens    46469
# tag types 21
# word types    10586
```

b. probabilities

tag bigram probabilities

```
p(B-person | O) = 0.009605
p(O | O) = 0.967824
p(I-person | B-person) = 0.442697
p(I-person | I-person) = 0.086124
p(O | I-person) = 0.913876
```

tag-word probabilities

```
p(God | B-person) = 0.018911
p(God | O) = 0.000136
p(Justin | B-person) = 0.039692
p(Justin | O) = 0.000024
p(Lindsay | B-person) = 0.025145
p(Lindsay | O) = 0.000002
```

c. first five sentences of the development set

```
STOP O O
WHAT O O
YOU'RE O O
DOING O O
AND O O
GO O O
GET O O
#ExpelledMovieToNumberOne O O
ON O O
ITUNES B-other O
BECAUSE O O
IT'S O O
ONLY O O
2ND O O
!! O O
@camerondallas O O
Shs O O
```

```
RT O O
@Phil_Heim O O
: O O
Safe O O
to O O
say O O
Super B-other O
Bowl I-other B-facility
Sunday I-other O
is O O
my O O
favourite O O
holiday O O
of O O
the O O
year O O

RT O O
@shashiranjanttv O O
: O O
@shashiranjanttv O O
second O O
is O O
Bawana B-geo-loc O
constituency O O
on O O
Feb O O
4 O O
- O O
Final O O

I've O O
watched O O
all O O
my O O
dreams O O
' O O
episodes O O
so O O
I O O
decided O O
to O O
choose O O
another O O
series O O
! O O
It O O
may O O
```

```
be O O
inspired O O
from O O
the O O
movie O O
I'll O O
watch O O
it O O
now O O
! O O
#GoodNight O O

Final O O
grades O O
go O O
up O O
tomorrow O O
:( O O
```

d. FB1 score 14.35%

```
processed 16261 tokens with 661 phrases; found: 245 phrases; correct: 65.
accuracy:  92.98%; precision:  26.53%; recall:   9.83%; FB1:  14.35
          company: precision:  63.64%; recall:  17.95%; FB1:  28.00  11
         facility: precision:   0.00%; recall:   0.00%; FB1:   0.00  19
          geo-loc: precision:  61.54%; recall:  27.59%; FB1:  38.10  52
            movie: precision:   0.00%; recall:   0.00%; FB1:   0.00  11
      musicartist: precision:   0.00%; recall:   0.00%; FB1:   0.00  9
            other: precision:  27.27%; recall:   6.82%; FB1:  10.91  33
           person: precision:  13.33%; recall:   7.02%; FB1:   9.20  90
          product: precision:  27.27%; recall:   8.11%; FB1:  12.50  11
       sportsteam: precision:  40.00%; recall:   2.86%; FB1:   5.33  5
           tvshow: precision:   0.00%; recall:   0.00%; FB1:   0.00  4
```

2.

```
1: train accuracy = 0.901440, dev accuracy = 0.901913
2: train accuracy = 0.906927, dev accuracy = 0.930632
3: train accuracy = 0.911360, dev accuracy = 0.931185
4: train accuracy = 0.916439, dev accuracy = 0.845581
5: train accuracy = 0.916460, dev accuracy = 0.930816
6: train accuracy = 0.917472, dev accuracy = 0.844044
7: train accuracy = 0.920312, dev accuracy = 0.894594
8: train accuracy = 0.920764, dev accuracy = 0.870057
9: train accuracy = 0.921776, dev accuracy = 0.829469
10: train accuracy = 0.923132, dev accuracy = 0.931493
11: train accuracy = 0.922185, dev accuracy = 0.932230
12: train accuracy = 0.925305, dev accuracy = 0.925958
13: train accuracy = 0.923132, dev accuracy = 0.890843
```

```
14: train accuracy = 0.925757, dev accuracy = 0.866613
15: train accuracy = 0.924208, dev accuracy = 0.928049
16: train accuracy = 0.925951, dev accuracy = 0.860833
17: train accuracy = 0.925154, dev accuracy = 0.875961
18: train accuracy = 0.927371, dev accuracy = 0.928418
FB1:  15.09
```

```
processed 16261 tokens with 661 phrases; found: 214 phrases; correct: 66.
accuracy:  92.84%; precision:  30.84%; recall:   9.98%; FB1:  15.09
        company: precision:  50.00%; recall:  15.38%; FB1:  23.53  12
       facility: precision:   5.88%; recall:   2.63%; FB1:   3.64  17
        geo-loc: precision:  56.00%; recall:  24.14%; FB1:  33.73  50
          movie: precision:   0.00%; recall:   0.00%; FB1:   0.00  11
   musicartist: precision:  33.33%; recall:   2.44%; FB1:   4.55  3
          other: precision:   7.84%; recall:   3.03%; FB1:   4.37  51
         person: precision:  42.31%; recall:  12.87%; FB1:  19.73  52
        product: precision:  22.22%; recall:   5.41%; FB1:   8.70  9
     sportsteam: precision:  40.00%; recall:   2.86%; FB1:   5.33  5
         tvshow: precision:   0.00%; recall:   0.00%; FB1:   0.00  4
```

b. I stopped the program automatically when the accuracy score of the training data reached a plateau, measured as being when the accuracy was smaller than the minimum of the last five epochs. I shuffled the training data each time, iterating through the data in a new random order so as to generate a better model for predicting unknown words. Since I used shuffling, the program did always reach over 15% after the same number of iterations, if it did so after a reasonable amount of time. As such, the automatic termination was more to stop the program if reaching a good accuracy score no longer seemed to be possible. During training, so as to ensure a FB1 score over 15%, I printed out the FB score and stopped the program manually when the FB1 score went over 15%. As the program was quite variable, jumping between a FB1 score of 11-15% and 5-8% between iterations regardless of the accuracy, an automatic detection based solely on the accuracy scores was not dependable. This took 18 epochs to reach in the example above, but given that the program runs fairly quickly per epcoch, it took only a little over 5 minutes to reach this accuracy score, as measured by timing the program with the time command.

3.

I tried updating emit_p with both the observed word and the unknown word character rather than just the unknown word character should the word be unknown while training. Words were determined to be unknown if it was the first time that they were encountered in the training data while the algorithm ran through all lines in the data. This was shuffled around so that new words would be measured as unknown words, as the alphabet (list of encountered words) was reset on every iteration of the training, and the list of training sentences was shuffled to be iterated through in a random iteration each time. This method improved the accuracy marginally from solely updating the unknown word counter, though the accuracy was still variable. The best FB1 accuracy achieved through this method was a little over 15%, and the accuracy for training consistently increased from 90% to around 93%, while the development accuracy fluctuated greatly, probably due to the unknown words changing, typically ending at around 90%. The FB1 score on the testing data as acheived on

the 21st epoch of training was 15.24.

A basic trigram implementation was slow with an accuracy hovering around 35% for training and 40% for dev, with an FB1 of about 0.6. This was probably due to the lack of smoothing onto a bigram model, as the trigram was much more limited. For this model, I used the prev word as well as the word before that as a tuple pair that represented the key to the transition probabilities. As this model did not improve accuracy at all, and only served to cause a severe drop in accuracy, I decided to forego attempting to smooth on a bigram model, as it did not seem to be the correct course of action.

I made a model that changed the model based on the testing data, which was not a valid training sample, but it was surprisingly not as effective as my next modification, which was converting words to all lower case.

Converting to lower case improved the FB1 score by a noticeable amount, with the training accuracy slowly rising from 90%-93% and the dev accuracy staying around 90% with some variation throwing the accuract down to 60-70%. The highest FB1 score acheived on the testing data was 17.38%.

```
processed 61908 tokens with 3473 phrases; found: 1175 phrases; correct: 404.
accuracy:  90.65%; precision:  34.38%; recall:  11.63%; FB1:  17.38
        company: precision:  53.28%; recall:  11.76%; FB1:  19.26  137
       facility: precision:   7.27%; recall:   3.16%; FB1:   4.41  110
        geo-loc: precision:  58.37%; recall:  26.87%; FB1:  36.80  406
          movie: precision:   0.00%; recall:   0.00%; FB1:   0.00  2
    musicartist: precision:   0.00%; recall:   0.00%; FB1:   0.00  5
          other: precision:  14.74%; recall:   4.79%; FB1:   7.24  190
         person: precision:  20.83%; recall:   9.34%; FB1:  12.89  216
        product: precision:  15.38%; recall:   4.88%; FB1:   7.41  78
     sportsteam: precision:   4.76%; recall:   0.68%; FB1:   1.19  21
         tvshow: precision:   0.00%; recall:   0.00%; FB1:   0.00  10
```