CSCI381/CSCI780

Homework 2

Due 11/13/2017 11:59 pm

1 The Naïve Bayes classifier

1.1 Movie review classification using Naïve Bayes

[**10 points**] Assume that you have trained a Naïve Bayes classifier for the task of sentiment classification (please refer to Chapter 6, pp. 1-9 in the J&M book). The classifier uses only bag-of-word features. Assume the following parameters for each word being part of a positive or negative movie review, and the prior probabilities are 0.4 for the positive class and 0.6 for the negative class.

	pos	**neg**
I	0.09	0.16
always	0.07	0.06
like	0.29	0.06
foreign	0.04	0.15
films	0.08	0.11

Question: What class will Naïve Bayes assign to the sentence "I always like foreign films"?
Show your work.
Answer

$$p(pos) * p(S|pos) = 0.4 * 0.09 * 0.07 * 0.29 * 0.04 * 0.08$$

$$p(neg) * p(S|neg) = 0.6 * 0.16 * 0.06 * 0.06 * 0.15 * 0.11$$

After simplifying the two products above, we conclude that $p(neg) * p(S|neg)$ is greater than $p(pos) * p(S|pos)$. Thus, the model predicts the class negative for this sentence.

1.2 TRAINING THE NAÏVE BAYES CLASSIFIER FOR MOVIE REVIEW CLASSIFICATION

[**40 points**]

1. Implement in Python a Naïve Bayes classifier with bag-of-word features and add-1 smoothing. Note: Smoothing should be used for the context features (bag-of-word features) only. Do not use smoothing for the prior parameters.

2. Use the following small corpus of movie reviews to train your classifier. Save the parameters of your model in a file called movie-review.NB

   a) fun, couple, love, love **comedy**

   b) fast, furious, shoot **action**

   c) couple, fly, fast, fun, fun **comedy**

   d) furious, shoot, shoot, fun **action**

   e) fly, fast, shoot, love **action**

3. Test you classifier on the new document below: *{fast, couple, shoot, fly}*. Compute the most likely class. Report the probabilities for each class.

   Answer

   We augment our training data with two more examples (one for each class), as follows (to ensure there're no features with 0 counts):[1]

   a) fun, couple, love, love **comedy**

   b) fast, furious, shoot **action**

   c) couple, fly, fast, fun, fun **comedy**

   d) furious, shoot, shoot, fun **action**

   e) fly, fast, shoot, love **action**

   f) fun, couple, love, fast, furious, shoot, fly **comedy**

   g) fun, couple, love, fast, furious, shoot, fly **action**

   The prior probabilities are:

$$p(action) = \frac{4}{7}$$

$$p(comedy) = \frac{3}{7}$$

   Bag-of-word feature parameters:

$$p(fun|comedy) = 1$$

[1] If you used a different approach of smoothing (which you explained in your write-up), this would be also accepted.

$$p(couple|comedy) = 1$$
$$p(love|comedy) = \frac{2}{3}$$
$$p(fast|comedy) = \frac{2}{3}$$
$$p(furious|comedy) = \frac{1}{3}$$
$$p(shoot|comedy) = \frac{1}{3}$$
$$p(fly|comedy) = \frac{2}{3}$$

$$p(fun|action) = \frac{1}{2}$$
$$p(couple|action) = \frac{1}{4}$$
$$p(love|action) = \frac{1}{2}$$
$$p(fast|action) = \frac{3}{4}$$
$$p(furious|action) = \frac{3}{4}$$
$$p(shoot|action) = 1$$
$$p(fly|action) = \frac{1}{2}$$

Based on the above parameters, we apply our model to the test example, as follows:

$$p(comedy) * p(S|comedy) = \frac{3}{7} * \frac{2}{3} * 1 * \frac{1}{3} * \frac{2}{3} = 0.063$$
$$p(action) * p(S|action) = \frac{4}{7} * \frac{3}{4} * \frac{1}{4} * 1 * \frac{1}{2} = 0.053$$

Thus our model assigns a higher score to the class *comedy*.

2 Part-of-Speech Tagging

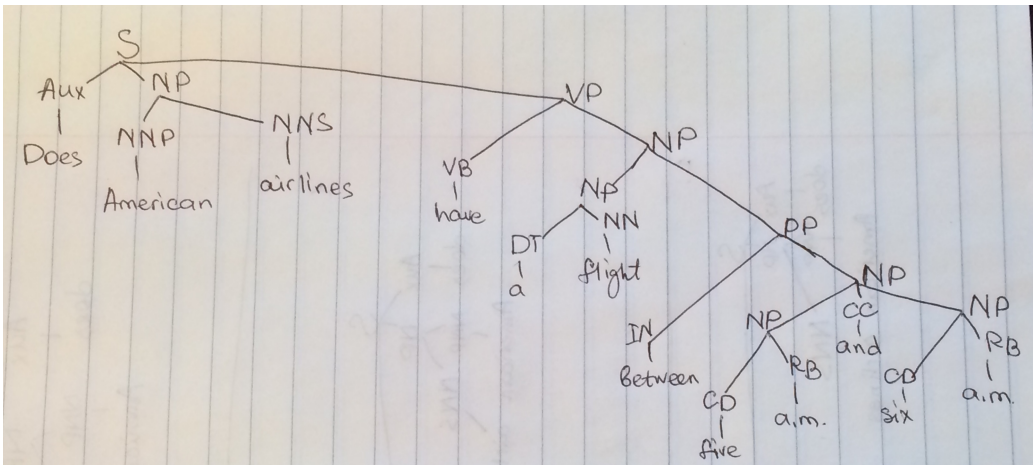
[**20 points**] Do exercise 10.1 at the end of Chapter 10 of the book:
`https://web.stanford.edu/~jurafsky/slp3/10.pdf`
Answer (Corrected tags are in bold)

1. I/PRP need/VBP a/DT flight/NN from/IN **Atlanta/NNP**
2. Does/VBZ this/DT flight/NN serve/VB **dinner/NN**
3. I/PRP **have/VBP** a/DT friend/NN living/VBG in/IN Denver/NNP
4. **Can/MD** you/PRP list/VB the/DT nonstop/JJ afternoon/NN flights/NNS

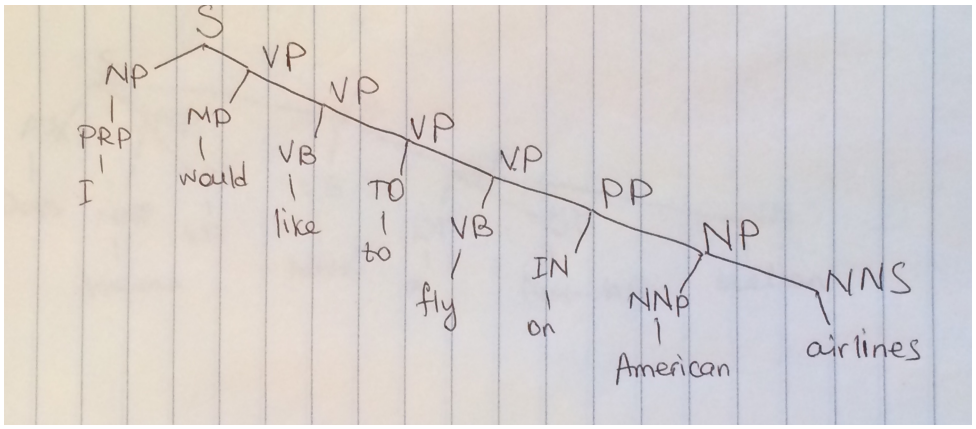
3 PARSING

[**30 points**] Do exercise 11.2 at the end of Chapter 11 of the book:
`https://web.stanford.edu/~jurafsky/slp3/11.pdf`

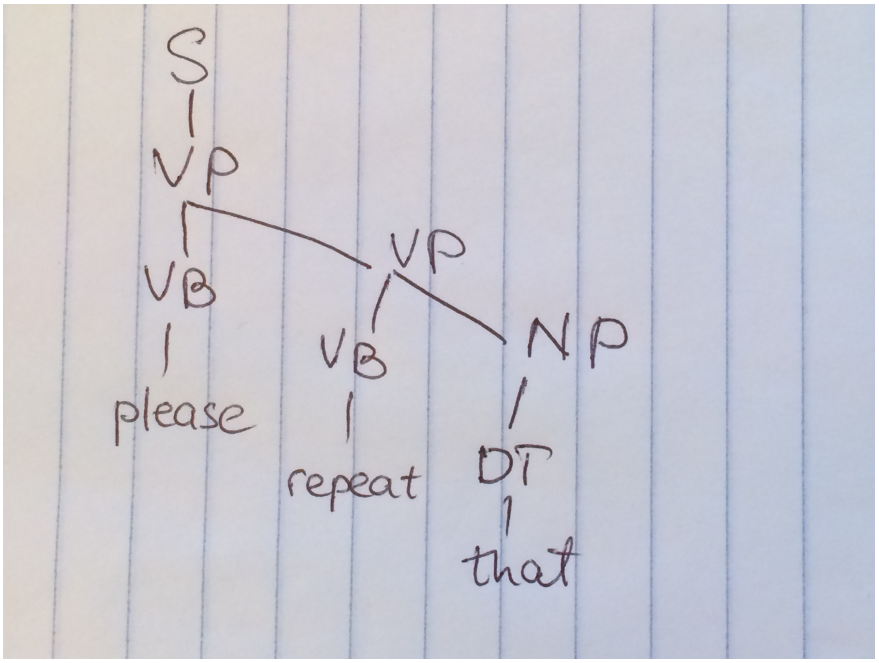
1. Does American airlines have a flight between five a.m. and six a.m.?

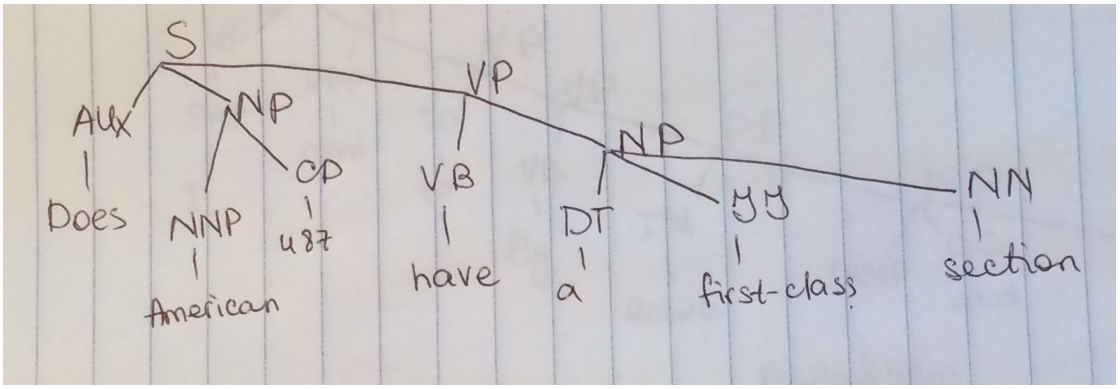


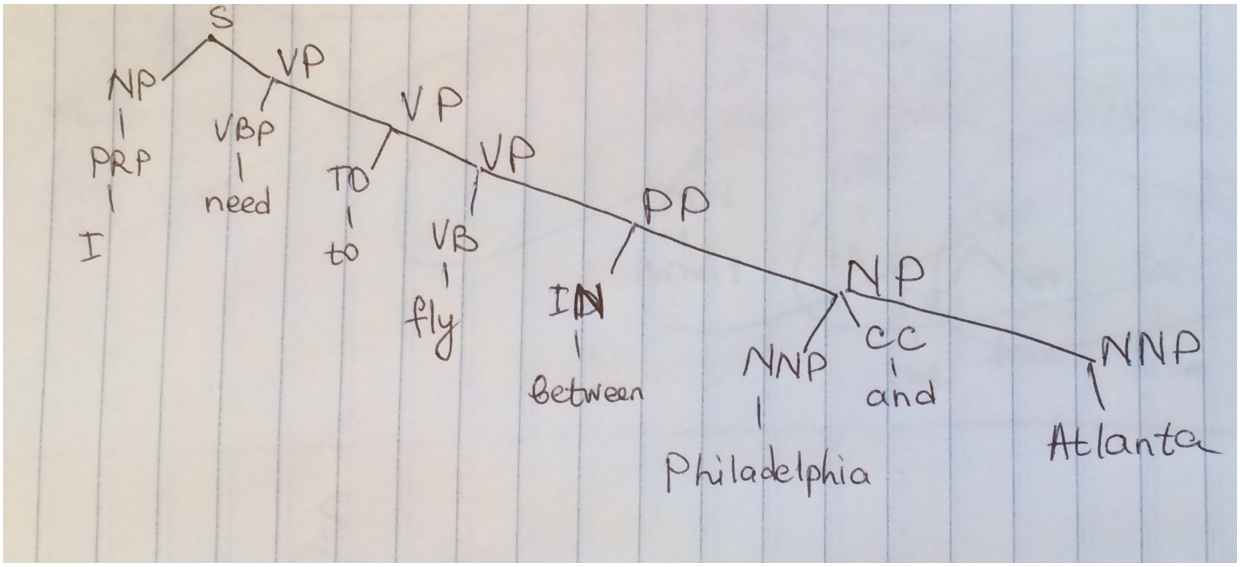
2. I would like to fly on American airlines.

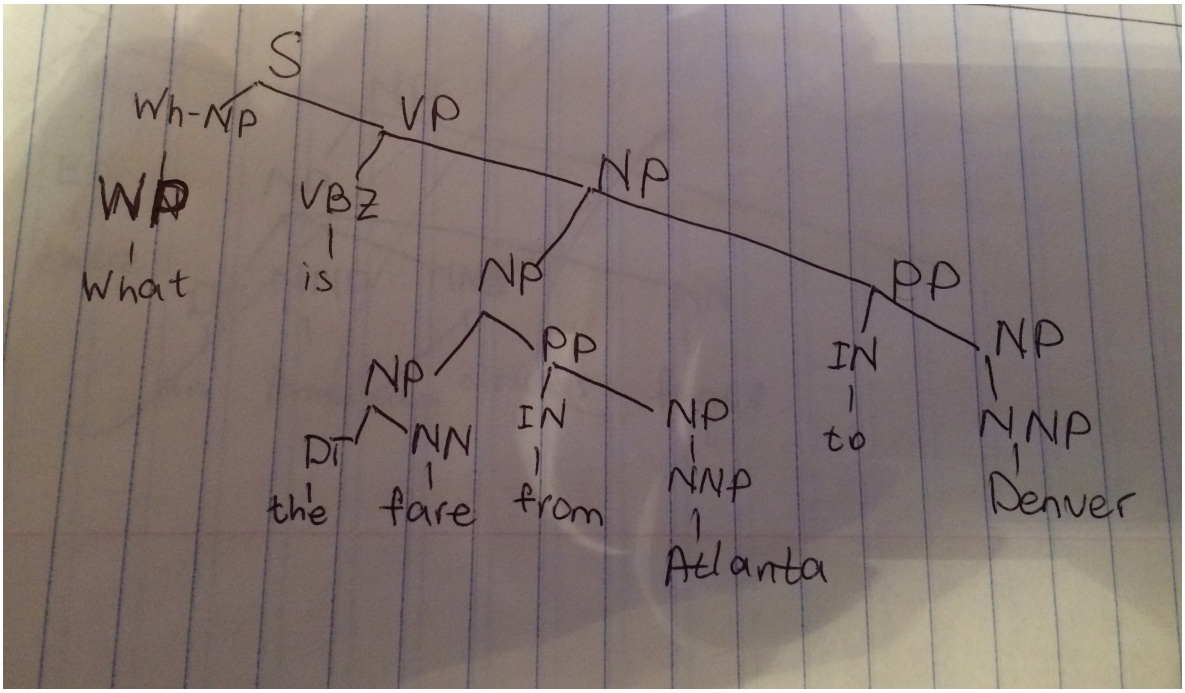


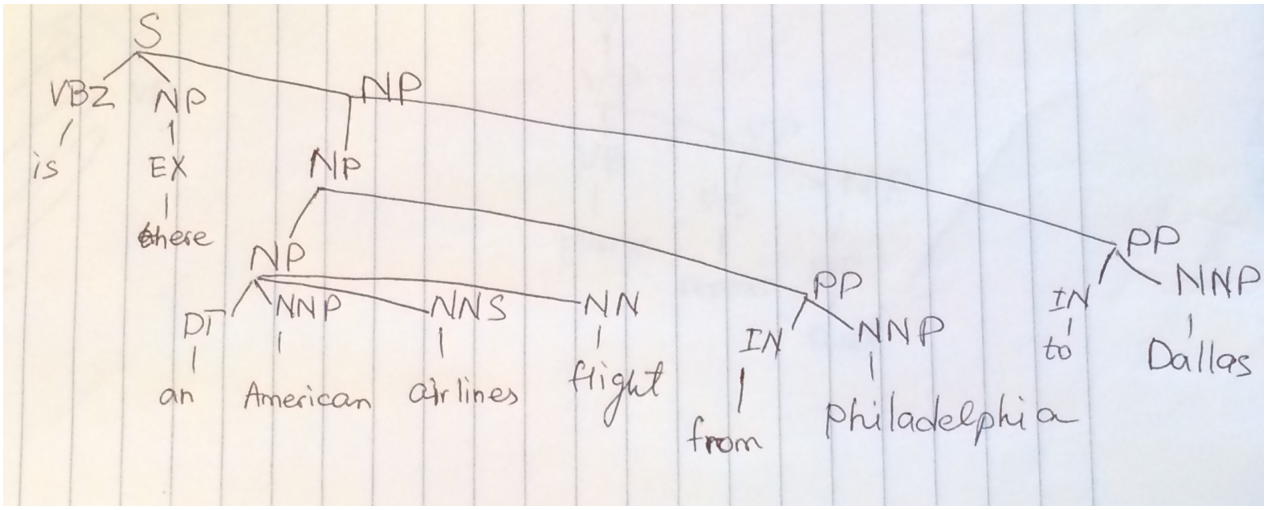
3. Please repeat that.
4. Does American 487 have a first-class section?
5. I need to fly between Philadelphia and Atlanta.
6. What is the fare from Atlanta to Denver?
7. Is there an American airlines flight from Philadelphia to Dallas?


S
VP
VB
please
VP
VB
repeat
NP
DT
that


S
Aux
Does
NP
NNP
American
CD
487
VP
VB
have
NP
DT
a
JJ
first-class
NN
section


S
NP
PRP
I
VP
VBP
need
VP
TO
to
VP
VB
fly
PP
IN
between
NP
NNP
Philadelphia
CC
and
NNP
Atlanta


S
Wh-NP
WP
What
VP
VBZ
is
NP
NP
DT
the
NN
fare
PP
IN
from
NP
NNP
Atlanta
PP
IN
to
NP
NNP
Denver



3.1 SUBMISSION

Please place the following on the server venus.cs.qc.edu and email me the path to the directory:

1. The Python code along with a README file that has instructions on how to run it in order to obtain the answers to questions in Section 1.2
2. The writeup that should be named **Homework2** that includes the answers to the questions.

Your grade will be based on the *correctness* of your answers, the *clarity* and completeness of your responses, and the *quality* of the code that you submitted.
Please refer to the course webpage on late submission policy.