Natural Language Processing, Fall 2020
Rasika Bhalerao
Midterm Exam
Due October 26 at 4:30pm

Instructions: You are allowed any notes, online resources, or textbooks. Please submit your solutions as a PDF on NYU Classes. Please number each question in your solutions clearly.

Name:

Net ID:

Total: 100 points

1) (6 points) Please give two examples of NLP tasks that email services (such as Gmail) perform for you. Please explain in a sentence or two for each example.

2) (14 points) The following paragraph is from Buzzfeed's article on 16 Mind-Blowing Fruit Facts.

   "Despite the fact that there are more than 1,000 banana varieties on earth, almost every single imported banana on the commercial market belongs to a single variety, called the Cavendish. These bananas became dominant throughout the industry in the 1960s because they were resistant to a fungal disease (called Panama Race One) that wiped out what had previously been the most popular banana, the Gros Michel. But signs point, pretty convincingly, to the Cavendish's own demise within the next decade."

   Suppose you want to find all instances of the word "the" using this regular expression: [tT]he
   (Recall from lecture 1, slide 29, that this matches "the" and "The" regardless of the letters before or after it.)

   a) How many true positives are there in the passage?
   b) How many positives are output by the algorithm?
   c) How many Type 1 errors are there?
   d) How many Type 2 errors are there?
   e) What is the precision of this algorithm?

f) What is the recall of this algorithm?
g) What is the F1 score of this algorithm?

3) (20 points) Here is a training set of 5 documents:

Document	**Category**
paw tail whiskers purr	cat
meow whiskers whiskers	cat
meow meow paw bark	cat
woof woof woof bark	dog
paw paw bark woof	dog

Please use the Naive Bayes algorithm to classify this sentence:

“woof bark whiskers whiskers”

Please do this by hand (no code) and use Laplace (add-1) smoothing. Please show your work.

4) (20 points) Here is a training set of 15 words that have been typed and the five words that they meant to type:

Correct word	**Typed 1**	**Typed 2**	**Typed 3**
hint	hinn	hint	tint
int	inn	int	int
inn	inn	nin	inn
tint	tini	tint	tint
nit	nit	nit	int

a) Consider each correct letter to be a state, and each typed letter to be an emission. Each correct word is a sequence of states, paired with each of the three sequences of emissions. Please create the table for emission probabilities.

b) Similarly, create the table for transition probabilities.

c) Please use the Viterbi algorithm to decode the correct state sequence for "int".

5) (20 points) Here is a grammar:

```
s  =  np vp
vp =  v np pp | v np | v pp | v
np =  n | n pp
pp =  p np
```

Here is a training corpus of 5 sentences:

```
(s (np (n Cats)) (vp (v purr) (pp (p in) (np (n solidarity)))))
(s (np (n Cats)) (vp (v eat) (np (n fish)) (pp (p with) (np (n speed)))))
(s (np (n Humans)) (vp (v hike) (pp (p with) (np (n dogs)))))
(s (np (n Cats) (pp (p in) (np (n America)))) (vp (v eat)))
(s (np (n Dogs)) (vp (v help) (np (n humans) (pp (p in) (np (n need))))))
```

Please ignore capitalization and punctuation.

a) Use the training corpus to train a Probabilistic CFG. Please list the probabilities for each production.

b) Please find both parses of the sentence "Cats help humans with dogs". What is the probability for each parse?

6) (10 points) Please find all coreference chains in the following paragraphs from Buzzfeed's article on 16 Mind-Blowing Fruit Facts. You don't need to include singletons.

"Pilots get paid hundreds of dollars a day to be on stand-by during the summer in case it rains and trees need an emergency blow-drying. It sounds ridiculous, but it's worth it for farmers who raise the delicate, expensive fruit. The job is dangerous; pilots are often injured in orchard crashes."

"In sub-tropical growing regions [...] there are never temperatures cold enough to break down the chlorophyll in the fruit's skin, which means it may still be yellow or green even when it's ripe. But because American consumers can't fathom such a

phenomenon, imported oranges get treated with ethylene gas to get rid of the chlorophyll and turn them orange."

"This also means that Florida oranges tend to be yellower than California oranges, because they're grown further south."

7) (10 points) Please describe a situation in which cultural biases need to be considered before an ethical NLP algorithm or model is deployed. Please write up to a paragraph (between 50 and 100 words). You can either describe a real-world situation in which biases were not sufficiently considered, yielding negative outcomes, or describe a hypothetical scenario where an NLP algorithm could feasibly be used in the real world.

Feedback: This part is not graded.

1. What would you like improved about this course?

2. Are there any topics in which you would be particularly interested for the "Selected Topics" lecture on November 23?