CSCI 544 – Applied Natural Language Processing, Spring 2017

Homework 1

Due: January 20, 2016, at 23:59 Pacific Time (11:59 PM)

Total: 7 pages, 50 points. This homework counts for 5% of the course grade.
Assignments turned in after the deadline but before February 3 are subject to a 30% grade penalty.
Students who registered on or after January 12 are not subject to the grade penalty.

General instructions

1. Do not write your name on any sheet.
2. This is an individual assignment. You may not work in teams or collaborate with other students. You must be the sole author of 100% of the material you turn in.
3. You may not look for solutions on the web, or use answers you find online or anywhere else.
4. Each student receives a personalized download copy of the assignment. You should download, print, write your answers and upload the finished copy only through the link provided.
5. The assignment must be submitted through the personalized link provided. Do not share the link with others. It is linked to your email.
6. The arithmetic is straightforward, though in some cases you may want to use a calculator. If in doubt, show your work.
7. Type your answers, or use *very neat and clear* handwriting. Answers that are difficult to read will be graded as errors, and these grades will not be changed.
8. Write concisely: enough space is provided to write your answers. Long and rambling answers will be penalized. You cannot add additional sheets.
9. The completed assignments will be accepted only through the online system. In person/email submissions will not be considered.

Problem 1. You are building a classifier for the sentiment of Russian adjectives. The following 100 adjectives have been sampled from the class of positive adjectives, to use as training data. The adjectives have been analyzed into a stem and suffix.

Adjective	Stem + suffix	Count
красивый	красив + ый	10
красивая	красив + ая	18
красивую	красив + ую	12
приятный	приятн + ый	10
приятная	приятн + ая	32
приятную	приятн + ую	18

a. (5 points) Based on the training data, give estimates for the probabilities of the individual stems and suffixes below.

P(красив- | positive) = P(приятн- | positive) =

P(-ый | positive) = P(-ая | positive) = P(-ую | positive) =

b. (6 points) Suppose that the stem and suffix are conditionally independent, given the class (that is, a naive Bayes model). If the probability estimates you just calculated exactly describe the class of positive adjectives, how many instances of each word would you expect to find in a sample of 100 words drawn from the class of positive adjectives?

красивый

красивая

красивую

приятный

приятная

приятную

c. (4 points) Is it possible to construct any sort of model that better fits the observed sample? If so, how? If not, why not?

d. (5 points) Roughly speaking (without calculating numbers), does our observed sample provide strong evidence against using a naive Bayes model for describing the class of positive adjectives? Why or why not?

Problem 2. Named entity recognition (NER) is the problem of identifying the names of persons, organizations, locations etc. In this problem you will construct a naive Bayes classifier to identify named entities in Czech. The table below is a snapshot of the data set, where phrases are labeled as to whether or not they represent a named entity. Each phrase is followed by The number of times it appears in the data.

Named entities	Not named entities
Nové Město (3)	Nové Auto (1)
Nové Dillí (5)	Kostel (9)
Kostel Panny Marie (2)	Červený (7)
Pan Červený (1)	Staré Auto (3)
Marie (4)	Nové (12)
	Červený Muž (3)

a. (2 points) Identify the priors for each class:

Named entity: _____ Not named entity: _____

b. (5 points) You will be constructing two types of features: *first word*, and *any word*. The *first word* feature of a phrase is the first word of the phrase; the *any word* feature of a phrase will have multiple occurrences – one for each word, including the first (so a three-word phrase, for example, will have three *any word* features).

Start by tabulating the number of instances of each feature, for each class.

	First word		Any word	
	Named Entity	Not Named Entity	Named Entity	Not Named Entity
Červený				
Kostel				
Marie				
Nové				
Pan				
Staré				
Auto				
Dillí				
Město				
Muž				
Panny				

c. (5 points) Apply Laplace (add-one) smoothing, and calculate the probabilities of each feature, conditional upon class.

	First word		Any word	
	Named Entity	Not Named Entity	Named Entity	Not Named Entity
Červený				
Kostel				
Marie				
Nové				
Pan				
Staré				
Auto				
Dillí				
Město				
Muž				
Panny				

d. (5 points) Use your classifier to predict for each of the following phrases whether or not they are a named entity: for each phrase, calculate the probability that it belongs to each class, and then select the most probable class. (Some of the phrases below are not proper Czech; don't worry about it for this exercise.)

	P(Named Entity)	P(Not Named Entity)	Chosen label
Červený Kostel			
Červený Město			
Dillí			
Kostel Panny Dillí			
Pan Auto			
Panny Marie			
Nové Kostel			
Nové Marie			
Nové Město			
Staré Dillí			

e. (6 points) Why do we construct the feature as "any word" rather than "word other than first"? (Hint: how would we classify *Dillí* with such features?)

f. (7 points) The first word of each phrase contributes two features for classification (*first word* and *any word*), so in effect it is counted twice. Is this justified? What would happen to *Pan Auto* ("Mr. Auto"), *Nové Marie* ("New Mary"), and *Staré Dillí* ("Old Delhi") if the first word only contributed one feature?