CSCI 544 – Applied Natural Language Processing

Homework 3

Due: February 15, 2016, at 23:59 Pacific Time (11:59 PM)

Total: 6 pages, 50 points. This homework counts for 5% of the course grade.
Assignments turned in after the deadline but before February 18 are subject to a 30% grade penalty.

General instructions

1. Do not write your name on any sheet.
2. 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.
3. The assignment must be submitted through the personalized link provided. Do not share the link with others. It is linked to your email.
4. The arithmetic is straightforward, though in some cases you may want to use a calculator. If in doubt, show your work.
5. Write concisely: enough space is provided to write your answers. Long and rambling answers will be penalized. You cannot add additional sheets.
6. The completed assignments will be accepted only through the online system. In person/email submissions will not be considered.

Problem 1. In this problem you will use probabilities to segment Arabic words into prefixes, stems and suffixes. Since we are able to give little data about stems, we concentrate only on prefixes and suffixes. The following segmented words are used as training data (the transliteration follows Habash, page 25; ∅ denotes an null prefix or suffix).

Arabic	Analysis	Meaning	Arabic	Analysis	Meaning
لولده	l + wld + h	to his child	وعدك	∅ + wʕd + k	your promise
وكتبه	w + ktb + h	and his books	فكتبي	f + ktb + y	and my books
فعمله	f + ʕml + h	and his work	لعملك	l + ʕml + k	to your work
وشغل	w + šɣl + ∅	and work	باذنه	b + Aðn + h	with his permission
صحتك	∅ + SHt + k	your health	فابني	f + Abn + y	and my son

a. (4 points) Give estimates for the probability of each prefix (don't forget the null prefix):

b. (4 points) Give estimates for the probability of each suffix (don't forget the null suffix):

c. (4 points) For segmenting words, we make the simplifying assumption that any sequence of characters is possible and equally likely as a stem; however, we do impose a constraint that a stem is at minimum three characters. Given this constraint, find the most likely segmentation for each word (use the transliteration, not the Arabic characters):

Arabic	Transliteration	Segmentation	Likelihood of prefix-suffix combination
فعلي	fʕly		
وضحك	wDHk		

d. (8 points) Does the segmenter always give the most common prefix? The most common suffix? Why?

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 a text. 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
New York (3)	New Shoes (2)
New Delhi (4)	Red (7)
Bank of America (2)	Bank (8)
Mr Red (1)	River Bank (3)
America (5)	New (13)
	Red Shoes (2)

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. 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
New				
Bank				
Mr				
Red				
America				
River				
York				
Delhi				
of				
Shoes				

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
New				
Bank				
Mr				
Red				
America				
River				
York				
Delhi				
of				
Shoes				

d. (5 points) Use your classifier to predict for each of the following phrases whether or not they are a named entity.

	P(Named Entity)	P(Not Named Entity)	Chosen label
Mr America			
Mr Shoes			
New Bank			
Bank of Delhi			
Delhi			
Red York			
New America			
New York			
New River			
Red River			

e. (6 points) Why do we construct the feature as “any word” rather than “word other than first”? (Hint: how would we classify *Delhi* 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 Mr Shoes, New America, and New York if the first word only contributed one feature?