Information Retrieval: Assignment 4

Problem 1. (20 points)

Based on the data below, estimate a multinomial Naive Bayes classifier (the type of NB classifier we introduced in class) and apply the classifier to the test document.

	docID	words in document	in c = Japan?
training set	1	Kyoto Osaka Taiwan	yes
	2	Japan Kyoto	yes
	3	Taipei Taiwan	no
	4	Macao Taiwan Shanghai	no
	5	London	no
test set	6	Taiwan Taiwan Kyoto	?

You only need to provide the subset of the parameters that you need to classify the test set (e.g., it's not necessary to estimate the lexical probabilities for "London").

Problem 2. (10 points)

Rank the documents in collection $\{d_1, d_2\}$ for query q using the language model approach to IR introduced in class. Use the mixture coefficient $\lambda = 0.4$.

- d_1: Scottish sheep getting smaller due to climate change study says
- d_2: The analysis has shown a dramatic shift in the natural ranges for US Bird species in response to climate change
- Query q: climate change

Problem 3. (10 points)

Familiarize yourself with wordnet at http://wordnetweb.princeton.edu/perl/webwn. For example, search for *suit* and notice the different sets of synonyms (synsets) for the different senses of the word: suit of clothes, lawsuit etc.

Give (a) an information need (b) a corresponding query and (c) a wordnet synset for one of the query terms such that expanding the query with one of the synset words *improves* search results on Google (compared to running the query (b)). Describe what is better about the modified results.

Give (a) an information need (b) a corresponding query and (c) a wordnet synset for one of the query terms such that expanding the query with one of the synset words makes search results *worse* on Google (compared to running the query (b)). Describe what is worse about the modified results.

It's probably easiest to expand only one of the query terms with only one of the synset terms in each case. You can choose freely which query terms and which synset terms you want to use. No need to include a printout – just describe original queries, modified queries, original results and better/worse results.

Problem 4. (10 points)

Consider the statistics for the first 100,000 documents in Reuters-RCV1 below. (i) Which two terms will be selected in frequency-based feature selection and why? (clarification in case this isn't clear from the slides: frequency selection is based on number of documents that are in the class and contain the term) (ii) Compute the MI values and order the terms according to MI. Which two terms will be selected in MI-based feature selection?

term	N_{00}	N_{01}	N_{10}	N_{11}
brazil	98,012	102	1835	51
council	96,322	133	3525	20
producers	98,729	119	1118	34
roasted	99,824	143	23	10

Problem 5. (10 points)

Design an algorithm that performs an efficient 1NN search in 1 dimension where efficiency is with respect to the number of documents N. The goal is to make 1NN classification faster, not 1NN training. In other words, the algorithm should be faster than the generic $O(N)$ algorithm for 1NN classification that simply scans all N documents. (i) Describe the algorithm you propose. What is the time complexity of your algorithm as a function of N (ii) for training a 1NN classifier and (iii) applying the trained 1NN classifier to a test document?

<u>Due date:</u> Wednesday, June 26, 2013, 12:15

Please turn in your assignment in class if possible. Email submissions are only accepted if you have a good reason why you cannot attend the review meeting. You will receive one extra point of credit if you use a staple or paper clip.