

Information Retrieval and Text Mining: Assignment 4

Problem 1. (10 points)

Familiarize yourself with wordnet at <http://wordnet.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.

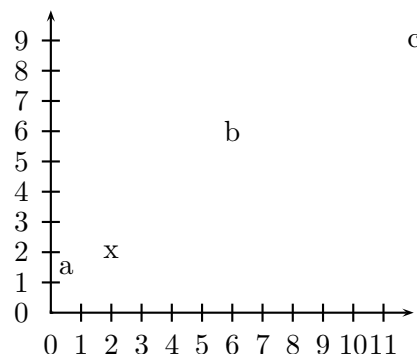
Problem 2. (20 points)

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 $c = \text{Japan?}$
training set	1	Kyoto Osaka Taiwan	yes
	2	Japan Kyoto	yes
	3	Taipei Taiwan	no
	4	Macao Taiwan Shanghai	no
test set	5	Taiwan Taiwan Kyoto	?

Problem 3. (10 points)

In the figure below, which of the three vectors \vec{a} , \vec{b} , and \vec{c} is (i) most similar to \vec{x} according to dot product similarity ($\sum_i x_i \cdot y_i$), (ii) most similar to \vec{x} according to cosine similarity ($\sum_i x_i \cdot y_i / (|x||y|)$), (iii) closest to \vec{x} according to Euclidean distance ($|x - y|$)? The vectors are $\vec{a} = (0.5 \ 1.5)^T$, $\vec{x} = (2 \ 2)^T$, $\vec{b} = (6 \ 6)^T$, and $\vec{c} = (12 \ 9)^T$.

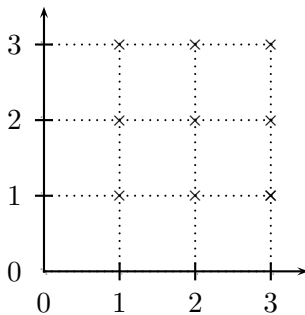


Problem 4. (10 points)

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?

Problem 5. (10 points)

(i) Perform a 3-means clustering for the points below. Draw a different diagram for each iteration to show the assignments and the centroids. If a tie occurs during an assignment step, you can freely choose any of the possible assignments. (ii) There are several clusterings that 3-means can converge to in this case. Give an example of such a clustering that is different from the one in (i).



Due date: Tuesday, June 23, 2009, 15:45 (turn assignments in before class in V38.03 or by email)