

Information Retrieval (CS60092)
Computer Science and Engineering, Indian Institute of Technology Kharagpur

Class Test 2, Autumn 2012

Attempt all questions.
Use of calculator is allowed.
State any assumptions made clearly.

Time: 1 hour
Full Marks: 20

Q. 1> Let the document collection contain only the following four documents:

d_1 : cats are small and so are dogs
 d_2 : cats and dogs may live as long as cats
 d_3 : dogs attack cats, cats and cats
 d_4 : dogs and cats may be friends of dogs

Rank the documents in response to the query *cats and dogs*, using the unigram MLE model from the document *only* (not the collection). Show all steps of the computation. (5)

Q. 2> Following the naïve Bayes classifier, what would be the more probable class for document 6 (see Table below)? Use Laplace smoothing in your computations. Clearly state all the multinomial parameters and the conditional probabilities involved. (5)

	docID	words in document	in class = <i>sports</i> ?
training set	1	football cricket football	Yes
	2	cricket termite grasshopper	No
	3	football football hockey	Yes
	4	football goal	Yes
	5	obama romney football	No
test set	6	football cricket hockey termite	?

Q. 3> Consider the set of six documents in the previous question as your entire collection. Use the TF-IDF weight formula $w_{t,d} = (1 + \log_{10} tf_{t,d}) \log_{10}(N/df_t)$, where N is the no. of documents in the collection. Compute the unnormalized weight vectors for each of the six documents. Will the Rocchio classification assign document 6 to sports? Why? (5)

Q. 4> Consider the query *obama health plan*. The document collection consists of six documents only, which are marked as relevant (R) or non-relevant (NR):

d_1 : president rejects rumors about his own bad health (NR)
 d_2 : the plan is to visit obama (NR)
 d_3 : obama raises concerns with us medical reforms (R)
 d_4 : president states a health vision (R)
 d_5 : romney states a health issue (NR)
 d_6 : obama states a health plan (R)

Assume a binary independence model (BIM) of retrieval. Rank the documents in descending order of their retrieval status value (RSV). Use contingency tables to show intermediate steps. Do not use any smoothing. The RSV for a BIM model is given by

$$RSV_d = \sum_{t: x_t = q_t = 1} \log_{10} \frac{p_t(1 - u_t)}{u_t(1 - p_t)}$$

where, for each term t , the probabilities of occurrence p_t and u_t can be represented in the form of the following contingency table:

	Document	R	NR
Term present	$x_t = 1$	p_t	u_t
Term absent	$x_t = 0$	$1 - p_t$	$1 - u_t$

(5)