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

Class Test 2, Autumn 2012

Attempt all questions.	*Time: 1 hour*
Use of calculator is allowed.	*Full Marks: 20*
State any assumptions made clearly.	

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)**

Ans.

$$\hat{P}(q|M_{d_1}) = \frac{1}{7} * \frac{1}{7} * \frac{1}{7} = 0.0029$$

$$\hat{P}(q|M_{d_2}) = \frac{2}{9} * \frac{1}{9} * \frac{1}{9} = 0.0027$$

$$\hat{P}(q|M_{d_3}) = \frac{3}{6} * \frac{1}{6} * \frac{1}{6} = 0.0139$$

$$\hat{P}(q|M_{d_4}) = \frac{1}{8} * \frac{2}{8} * \frac{1}{8} = 0.0039$$

Ranking: $d_3 > d_4 > d_1 > d_2$

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 = *sports*?
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*	?

Ans. Multinomial parameters:

$$\hat{P}(sports) = \frac{3}{5} = 0.6$$

$$\hat{P}(\overline{sports}) = \frac{2}{5} = 0.4$$

Conditional probabilities (No. of terms in vocabulary = 8):

$$\hat{P}(football|sports) = \frac{5+1}{8\ +\ 8} = 0.3750$$

$$\hat{P}(cricket|sports) = \frac{1+1}{8\ +\ 8} = 0.1250$$

$$\hat{P}(hockey|sports) = \frac{1+1}{8\ +\ 8} = 0.1250$$

$$\hat{P}(termite|sports) = \frac{0+1}{8\ +\ 8} = 0.0625$$

$$\hat{P}(football|\overline{sports}) = \frac{1+1}{6\ +\ 8} = 0.1429$$

$$\hat{P}(cricket|\overline{sports}) = \frac{1+1}{6\ +\ 8} = 0.1429$$

$$\hat{P}(hockey|\overline{sports}) = \frac{0+1}{6\ +\ 8} = 0.0714$$

$$\hat{P}(termite|\overline{sports}) = \frac{1+1}{6\ +\ 8} = 0.1429$$

$$Then, \hat{P}(sports|d_6) = 0.6\ *\ 0.3750 *\ 0.1250 * 0.1250 * 0.0625 = 0.00022$$

$\hat{P}(\overline{sports}|d_6)$ = 0.4 * 1429 * 0.1429 * 0.0714 * 0.1429 = 0.00008

Thus, the classifier assigns the test **document 6 to *sports*.**

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)**

Ans. Let c = sports and c-ba

	Term frequencies								
Vector	*football*	*cricket*	*termite*	*grasshopper*	*hockey*	*goal*	*obama*	*romney*	Class
$\overrightarrow{d_1}$	2	1	0	0	0	0	0	0	Yes
$\overrightarrow{d_2}$	0	1	1	1	0	0	0	0	No
$\overrightarrow{d_3}$	2	0	0	0	1	0	0	0	Yes
$\overrightarrow{d_4}$	1	0	0	0	0	1	0	0	Yes
$\overrightarrow{d_5}$	1	0	0	0	0	0	1	1	No
$\overrightarrow{d_6}$	1	1	1	0	1	0	0	0	?
$\overrightarrow{\mu_{sport}}$	-	-	-	-	-	-	-	-	-
$\overrightarrow{\mu_{\overline{sports}}}$	-	-	-	-	-	-	-	-	-

	Term weights								
Vector	*football*	*Cricket*	*termite*	*grasshopper*	*hockey*	*goal*	*obama*	*romney*	Class
$\overrightarrow{d_1}$	0.1030	0.3010	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	Yes
$\overrightarrow{d_2}$	0.0000	0.3010	0.4771	0.7782	0.0000	0.0000	0.0000	0.0000	No
$\overrightarrow{d_3}$	0.1030	0.0000	0.0000	0.0000	0.4771	0.0000	0.0000	0.0000	Yes
$\overrightarrow{d_4}$	0.0792	0.0000	0.0000	0.0000	0.0000	0.7782	0.0000	0.0000	Yes
$\overrightarrow{d_5}$	0.0792	0.0000	0.0000	0.0000	0.0000	0.0000	0.7782	0.7782	No
$\overrightarrow{d_6}$	0.0792	0.3010	0.4771	0.0000	0.4771	0.0000	0.0000	0.0000	?
$\overrightarrow{\mu_{sports}}$	0.0951	0.1003	0.0000	0.0000	0.1590	0.2594	0.0000	0.0000	-
$\overrightarrow{\mu_{\overline{sports}}}$	0.0396	0.1505	0.2386	0.3891	0.0000	0.0000	0.3891	0.3891	-

$$\left|\overrightarrow{\mu_{sports}} - \overrightarrow{d_6}\right| = 0.6608$$

$$\left|\overrightarrow{\mu_{\overline{sports}}} - \overrightarrow{d_6}\right| = 0.8735$$

Thus, Rocchio **assigns *d*$_6$ to *sports*** as $\overrightarrow{d_6}$ is closer to the mean vector of *sports* class than the $\overline{sports}$ class.

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)

Ans. For term *obama*,

	Document	R	NR
Term present	x_t = 1	0.67	0.33
Term absent	x_t = 0	0.33	0.67

For term *health*,

	Document	R	NR
Term present	x_t = 1	0.67	0.67
Term absent	x_t = 0	0.33	0.33

For term *plan*,

	Document	R	NR
Term present	x_t = 1	0.33	0.33
Term absent	x_t = 0	0.67	0.67

$$RSV_1(for\ health) = \log_{10} \frac{0.67*0.33}{0.67*0.33} = 0.00$$

$$RSV_2(for\ obama\ and\ plan) = \log_{10} \frac{0.67*0.67}{0.33*0.33} + \log_{10} \frac{0.33*0.67}{0.33*0.67} = 0.60$$

$$RSV_3(for\ obama) = \log_{10} \frac{0.67*0.67}{0.33*0.33} = 0.60$$

$$RSV_4(for\ health) = \log_{10} \frac{0.67*0.33}{0.67*0.33} = 0.00$$

$$RSV_5(for\ health) = \log_{10} \frac{0.67*0.33}{0.67*0.33} = 0.00$$

$$RSV_6(for\ obama, health\ and\ plan) = \log_{10} \frac{0.67*0.67}{0.33*0.33} + \log_{10} \frac{0.67*0.33}{0.67*0.33} + \log_{10} \frac{0.33*0.67}{0.33*0.67} = 0.60$$

Ranking: $d_2 = d_3 = d_6 > d_1 = d_4 = d_5$