Natural Language Processing (CO3086)
Lab 4 - NLP 242

HO CHI MINH UNIVERSITY OF TECHNOLOGY

Vietnam National University Ho Chi Minh

Problem 1

Assume the following likelihoods for each word being part of a positive or negative movie review, and equal prior probabilities for each class.

Word	pos	neg
I	0.09	0.16
always	0.07	0.06
like	0.29	0.06
foreign	0.04	0.15
films	0.08	0.11

What class

Answer.

Problem 2

Given the following short movie reviews, each labeled with a genre, either comedy or action:

1. fun, couple, love, love **comedy**
2. fast, furious, shoot **action**
3. couple, fly, fast, fun, fun **comedy**
4. furious, shoot, shoot, fun **action**
5. fly, fast, shoot, love **action**

and a new document D: `fast, couple, shoot, fly`
Compute the most likely class for D. Assume a naive Bayes classifier and use add-1 smoothing for the likelihoods.

Answer.

Problem 3

Train two models, multinomial naive Bayes and binarized naive Bayes, both with add-1 smoothing, on the following document counts for key sentiment words, with positive or negative class assigned as noted.

doc	good	poor	great	class
$d1$	3	0	3	pos
$d2$	0	1	2	pos
$d3$	1	3	0	neg
$d4$	1	5	2	neg
$d5$	0	2	0	neg

Use both naive Bayes models to assign a class (pos or neg) to this sentence:

A good, good plot and great characters, but poor acting.

With naive Bayes text classification, we simply ignore (throw out) any word that never occurred in the training document. (We don't throw out words that appear in some classes but not others; that's what add-one smoothing is for.) Do the two models agree or disagree?

Answer.

Problem 4

Consider that our document collection S has the following documents: $D_1, ..., D_5$:

document	words
D_1	Data Base System Concepts
D_2	Introduction to Algorithms
D_3	Computational Geometry: Algorithms and Applications
D_4	Data Structures and Algorithm Analysis on Massive Data Sets
D_5	Computer Organization

Our dictionary $DICT$ consists of 8 words: $\{w_1 = \text{data}, w_2 = \text{system}, w_3 = \text{algorithm}, w_4 = \text{computer}, w_5 = \text{geometry}, w_6 = \text{structure}, w_7 = \text{analysis}, w_8 = \text{organization}\}$. We consider that, by stemming, "computer" and "computational" are regarded as the same word, and so are "algorithms" and "algorithm".

1. Let $tf(w, D)$ denote the term frequency of term w in a document D. Give the value of $tf(w_i, D_j)$ for all $1 \leq i \leq 8$ and $1 \leq j \leq 5$.

2. Let $idf(w)$ denote the inverse document frequency of term w as defined in our lecture notes. Give the value of $idf(w_i)$ for all $1 \leq i \leq 8$.

3. Convert each document in S into an 8-dimensional point according to the tf-idf model as defined in our lecture notes.

4. Assume that we have received a query with terms "Geometry Algorithm Concepts". Convert the query to an 8-dimensional point.

5. Rank the documents in descending order of their relevance to the query in Problem 4 according to the cosine metric.

Answer.

1) The value of $tf(w_i, D_j)$ *for all* $1 \leq i \leq 8$ *and* $1 \leq j \leq 5$

	D_1	D_2	D_3	D_4	D_5
w_1	1	0	0	2	0
w_2	1	0	0	0	0
w_3	0	1	0	0	0
w_4	0	0	1	0	1
w_5	0	0	1	0	0
w_6	0	0	0	1	0
w_7	0	0	0	1	0
w_8	0	0	0	0	1

2) The value of $idf(w_i)$ *for all* $1 \leq i \leq 8$

w_1	1.32
w_2	2.32
w_3	0.74
w_4	1.32
w_5	2.32
w_6	2.32
w_7	2.32
w_8	2.32

For example, $idf(w_1) = \log_2(|S|/2) = \log_2(5/2) = 1.32$. *In particular, the 2 in the denominator is because* w_1 *appears in two documents* D_1 *and* D_4.

3) Consider D_i $(1 \leq i \leq 5)$. *Let* p_i *be the point converted from* D_i. *The* j*-th coordinate* $p_i[j]$ *of* p_i *equals* $\log_2(1 + tf(w_j, D_i)) \cdot idf(w_j)$. *For example, when* $i = j = 1$, $p_1[1] = \log_2(1+1) \cdot 1.32 = 1.32$. *In this way, we can obtain* $p_1, ..., p_5$ *as:*

	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8
p_1	1.32	2.32	0	0	0	0	0	0
p_2	0	0.74	0	0	0	0	0	0
p_3	0	0.74	1.32	2.32	0	0	0	0
p_4	2.09	0	0.74	0	0	2.32	2.32	0
p_5	0	0	0	1.32	0	0	0	2.32

4) Answer: $(0, 0, 0.74, 0, 2.32, 0, 0, 0)$.

5) Let q *be the point converted from* Q. *The cosine metric calculates the score of* p_i *and* q *as:*

$$score(p_i, q) = \frac{p_i \cdot q}{|p_i| \cdot |q|}$$

Consider, for example, p_2. *We have* $p_2 \cdot q = 0.74 \cdot 0.74 = 0.55$ *(all the other terms in the dot product are 0). This, together with* $|p_2| = 0.74$ *and* $|q| = 2.44$, *gives* $score(p_2, q) = \frac{0.55}{0.74 \times 2.44} = 0.30$. *The following table gives the scores of all documents:*

	D_1	D_2	D_3	D_4	D_5
	0	0.30	0.74	0.06	0

The relevance ranking is D_3, D_2, D_4, D_1 *and* D_5.