Paper Code	Examiner	Office
INT309	Fangyu Wu	SD555

2023/24 SEMESTER 1 - In-class Test

BEng Digital Media Technology - Year 4

Multimedia Information Retrieval and Technology

Time Allowed: 2 Hours

Instructions to Candidates

1. Total marks available are 100.
2. Answer all questions.
3. In general, it is particularly important to give reasons for your answer. Only partial marks will be awarded for correct answers with inadequate reasons.
4. Only Casio FS82ES / 83ES calculators are allowed.
5. Answer should be written in the provided answer booklet(s).

Question 1 (20 marks)

Consider the query and document:

Query: best car insurance **Document:** car insurance auto insurance

Collection with N = 1,000,000 documents where the document frequencies of auto, best, car and insurance are respectively 5000, 50000, 10000 and 1000. Compute and report the cosine similarity of the query and document. Logarithmic term weighting for query and raw term frequency for document, the query is converted to a unit vector using tf-idf weighting $w_{t,d} = tf_{t,d} \times \log_{10}(N/df_t)$, and length normalization for document only.

Document: car insurance auto insurance
Query: best car insurance

Term	Query					Document			Prod
	tf-raw	tf-wt	df	idf	tf-idf	tf-raw	tf-wt	Normalization	
auto	0	0	5000	2.3	0	1	1	1/√6	0
best	1	1	50000	1.3	1.3	0	0	0	0
car	1	1	10000	2.0	2.0	1	1	1/√6	2*1/√6
insurance	1	1	1000	3.0	3.0	2	1.3	2/√6	3*2/√6

Score = 0+0+2*1/√6+3*2/√6 ≈ 3.27

Question 2 (30 marks)

Consider an information need for which there are 5 relevant documents in the collection, which are ranked as very relevant, relevant, somewhat relevant (VR, R, SR) vs. nonrelevant (N). For computation of NDCG(Normalized Discounted Cumulative Gain), treat these ordinal values as the numbers 3, 2, 1, and 0, respectively.

1. (4 points) Contrast two systems running a query for the information need on this collection. Their top 10 results were judged for relevance as follows (the leftmost item is the top ranked search result):

   System 1 = VR, R, R, N, SR, VR, N, N, N, N
   System 2 = VR, VR, R, R, N, N, N, SR, N, N

   What is the NDCG of each system? Which has the higher NDCG? Show calculations.

Solution: The NDCG score for system 1 is 0.90
The NDCG score for system 2 is 0.99
The system 2 has the higher NDCG

2. (6 points) Now assuming that we collapse distinction relevance levels to a binary relevance assessment (R vs. N). Which system is better when the evaluation is based on Precision @3?

Solution: Precision@3 for system 1: 100%
Precision@3 for system 2: 100%
Same

3. (6 points) What is the MAP(Mean average precision) of each system? Which has a higher MAP?

Solution: $AP_1 = \frac{1+1+1+\frac{4}{5}+\frac{5}{6}}{5} = \frac{4.633}{5} \approx 0.927$
$AP_1 = \frac{1+1+1+1+\frac{5}{8}}{5} = \frac{4.633}{5} \approx 0.925$
System 1 has a higher AP.

4. (4 points) Intuitively, which system seems better for web search? Why?

Solution: System 1.
System 1 is able to search more accurate results.

5. (4 points) To evaluate the effectiveness of information retrieval (IR) system, what are the three basic elements required?

Solution:

1)A benchmark collection of documents; Must be representative;

2)A benchmark suite of information needs, expressible as queries; Must be representative;

3)An assessment of either Relevant or Nonrelevant judgments for each query-document pair.

6. (4 points) What is relevance feedback?

Solution:
The idea of relevance feedback (RF) is to involve the user in the retrieval process so as to improve the final result set.
1.The user marks some returned documents as relevant or nonrelevant.
2.The system computes a better representation of the information need based on the user feedback.
3.The system displays a revised set of retrieval results

Question 3 (50 marks)

In this question, we will explore various techniques for classifying documents. Consider the following supervised corpus of documents:

	DocID	Contents	Class *c=China*
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no

And the query document "Chinese Chinese Tokyo Japan"

1. (20 points) Compute the raw term frequence (no idf, normalized) representations of the five documents in the table below. Then calculate the two class centroids. Finally, predict the class of the query document using the Rocchio classification algorithm.

Solution:

DocID	Beijing	Chinese	Japan	Macao	Shanghai	Tokyo
d1	1	2	0	0	0	0
d2	0	2	0	0	1	0
d3	0	1	0	1	0	0
d4	0	1	1	0	0	1
d5	0	2	1	0	0	1

Based on the above raw term frequency, we can compute the normalized query vector
d1: $(\frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}}, 0, 0, 0, 0)$
d2: $(0, \frac{2}{\sqrt{5}}, 0, 0, \frac{1}{\sqrt{5}}, 0)$
d3: $(0, \frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}}, 0, 0)$
d4: $(0, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, 0\ , 0, \frac{1}{\sqrt{3}})$
q: $(0, \frac{2}{\sqrt{6}}, \frac{1}{\sqrt{6}}\ , 0, 0, \frac{1}{\sqrt{6}})$

We measure the distance between the query vector and each centroid, and choose the class with the smallest distance.
The centroid for class c: (0.15, 0.83, 0, 0.24, 0.15, 0)
The distance between centroid for class c and query document:
$\sqrt{(0.15)^2 + (0.83 - 0.82)^2 + (0.41)^2 + (0.24)^2 + (0.15)^2 + (0.41)^2} \approx 0.66$

The centroid for class $\overline{c}$: (0, 0.58, 0.58, 0, 0, 0.58)
The distance between centroid for class $\overline{c}$ and query document:
$\sqrt{(0.58 - 0.82)^2 + (0.58 - 0.41)^2 + (0.58 - 0.41)^2} \approx 0.34$

The query belongs to the class of not china.

2. (10 points) What are the three nearest neighbors to the query? What class does this document belong to if we consider 3NN classification? Use raw term frequency, no idf, and cosine similarity.

Solution:

d1≈0.73
d2≈0.73
d3≈0.58
d4≈0.94
Three nearest neighbors: Doc4, Doc1, Doc2.
The query belongs to class c.

3. (15 points) Predict the class of the query using Bernoulli model or Multinomial event model in Naive Bayes Classification (select one model to predict the class).

Example 13.1: For the example in Table 13.1, the multinomial parameters we need to classify the test document are the priors $\hat{P}(c) = 3/4$ and $\hat{P}(\overline{c}) = 1/4$ and the following conditional probabilities:

$$\begin{aligned}
\hat{P}(\text{Chinese}|c) &= (5+1)/(8+6) = 6/14 = 3/7 \\
\hat{P}(\text{Tokyo}|c) = \hat{P}(\text{Japan}|c) &= (0+1)/(8+6) = 1/14 \\
\hat{P}(\text{Chinese}|\overline{c}) &= (1+1)/(3+6) = 2/9 \\
\hat{P}(\text{Tokyo}|\overline{c}) = \hat{P}(\text{Japan}|\overline{c}) &= (1+1)/(3+6) = 2/9
\end{aligned}$$

The denominators are $(8+6)$ and $(3+6)$ because the lengths of $text_c$ and $text_{\overline{c}}$ are 8 and 3, respectively, and because the constant B in Equation (13.7) is 6 as the vocabulary consists of six terms.

$\hat{P}(c \mid q) \propto 3/4 \cdot (3/7)^2 \cdot 1/14 \cdot 1/14 \approx 0.0007$

$\hat{P}(\overline{c} \mid q) \propto 1/4 \cdot (2/9)^2 \cdot 2/9 \cdot 2/9 \approx 0.0006$

Thus, the query belongs to the class of china.

4. (5 points) What are the application scenarios for text classification?

Solution: Spam filter and Google alerts

Appendix A: Equation List

The NDCG:

$$NDCG_{rank\ n} = \frac{ActualDCG_{rank\ n}}{IdealDCG_{rank\ n}} \tag{1}$$

Multinomial:

$$P(c|d) \propto P(c)\Pi_{1 \leqslant k \leqslant N_d} P(t_k|c) \tag{2}$$

$$P(t_k|c) = \frac{T_{ct}+1}{\sum_{t' \in V(T_{ct'}+1)}} \tag{3}$$

Bernoulli:

$$P(c|d) \propto P(c)\Pi_{t_k \in Q} P(t_k|c)\Pi_{t_k \notin Q}[1 - P(t_k|c)] \tag{4}$$

$$P(t_k|c) = \frac{df_{ct}+1}{N_c + Numberofclasses} \tag{5}$$

——— *End of paper* ———