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

Class Test 1

Time: 1 hour
Full Marks: 20

Attempt all questions.
Use of calculator is allowed.

Q. 1> a. Find the Jaccard coefficients of *bord* with *aboard*, *border*, *lord* and *morbid*. **(2)**

Soln. We consider bigrams here.
Bigrams in *bord* = {*bo*, *or*, *rd*}
Bigrams in *aboard* = {*ab*, *bo*, *oa*, *ar*, *rd*}
Jaccard coefficient = |A ∩ B| / |A U B| = 2/6 = **0.33 Ans.**
Bigrams in *border* = {*bo*, *or*, *rd*, *de*, *er*}
Jaccard coefficient = |A ∩ B| / |A U B| = 3/5 = **0.40 Ans.**
Bigrams in *lord* = {*lo*, *or*, *rd*}
Jaccard coefficient = |A ∩ B| / |A U B| = 2/4 = **0.50 Ans.**
Bigrams in *morbid* = {*mo*, *or*, *rb*, *bi*, *id*}
Jaccard coefficient = |A ∩ B| / |A U B| = 1/7 = **0.14 Ans.**

b. Assuming that the components of document vectors are computed using the tf-idf weighting scheme, find the vectors corresponding to d_1 and d_2 (coming from the same document collection, with 2000 documents). Also find the cosine similarity between these two vectors. **(3)**

term	tf (d_1)	tf (d_2)	df$_t$
car	10	30	520
auto	15	12	618
insurance	5	8	430
best	25	10	790

Soln. $\text{idf}_{car} = \log_{10}(N/\text{df}_t) = \log_{10}(2000/520) = \log_{10}3.85 = 0.59$
$\text{idf}_{auto} = \log_{10}(N/\text{df}_t) = \log_{10}(2000/618) = \log_{10}3.24 = 0.51$
$\text{idf}_{insurance} = \log_{10}(N/\text{df}_t) = \log_{10}(2000/430) = \log_{10}4.65 = 0.67$
$\text{idf}_{best} = \log_{10}(N/\text{df}_t) = \log_{10}(2000/790) = \log_{10}2.53 = 0.40$
V(d*₁*) = (10 x 0.59, 15 x 0.51, 5 x 0.67, 25 x 0.40) = (5.90, 7.65, 3.35, 10.00) **Ans.**
V(d*₂*) = (30 x 0.59, 12 x 0.51, 8 x 0.67, 10 x 0.40) = (17.70, 6.12, 5.36, 4.00) **Ans.**
Cosine similarity(d_1, d_2) = $(V(d_1).\ V(d_2))/|V(d_1)||V(d_2)|$
= ((5.90 x 17.70) + (7.65 x 6.12) + (3.35 x 5.36) + (10.00 x 4.00))/(($5.90^2 + 7.65^2 + 3.35^2 + 10.00^2)^{1/2}$ x ($17.70^2 + 6.12^2 + 5.36^2 + 4.00^2)^{1/2}$)
= (104.43 + 46.82 + 17.96 + 40.00)/((34.81 + 58.52 + 11.22 + 100.00) x (313.29 + 37.45 + 28.73 + 16.00))
= 209.21/(204.55 x 395.47) = 209.21/80893.39 = **2.59×10^{-3} Ans.**

Q. 2> a. A collection has 500,000 documents, 250 tokens per documents, four characters per token and 200,000,000 postings. A posting is defined as a doc-id in the postings list, excluding any other information.

i. Find the length of a doc-id.

ii. Find the size of the collection in MBs.

iii. Find the size of the uncompressed posting file. **(0.5 x 3 = 1.5)**

Soln. i. Length of doc-id = $\log_2 500000 = \log_{10} 500{,}000 / \log_{10} 2 = 18.93 \approx$ **19 bits. Ans.**

ii. Size of the collection = 500,000 x 250 x 4 bytes = **476.84 MB Ans.**

iii. Size of the uncompressed posting file = 200,000,000 X 19 bits = 3.80×10^9 bits = **453.00 MB Ans.**

b. Let us assume that gap encoding using variable byte codes is being used. Let the postings list for some term consist of the doc-ids 824, 829, 1234. How should this postings list be represented using the above encoding scheme? **(3.5)**

Soln.

docIDs	824		829	1234	
gaps			5	405	
VB code	00000110	10111000	10000101	00000011	10010101

Q. 3> Consider a document collection that contains the following documents:

*d*₁: *tick goes the clock goes tick tick tick*

*d*₂: *tick tock big time*

*d*₃: *clock tower*

*d*₄: *big tower of clock*

Let a query be *"clock tick"*. Compute the tf-idf scores of each document with respect to this query and provide the resultant document ranking. **(5)**

Soln. $\text{idf}_{clock} = \log_{10}(N/\text{df}_t) = \log_{10}(4/3) = 0.12$

$\text{idf}_{tick} = \log_{10}(N/\text{df}_t) = \log_{10}(4/2) = 0.30$

For d_1, $\text{tf}_{clock} = 1$, $\text{idf}_{clock} = 0.12 \longrightarrow \text{tf-idf}_{clock} = 1 \times 0.12 = 0.12$

$\text{tf}_{tick} = 4$, $\text{idf}_{tick} = 0.30 \longrightarrow \text{tf-idf}_{tick} = 4 \times 0.30 = 1.20$

Score of d_1 = 0.12 + 1.20 = **1.32 Ans.**

For d_2, $\text{tf}_{clock} = 0$, $\text{idf}_{clock} = 0.12 \longrightarrow \text{tf-idf}_{clock} = 0 \times 0.12 = 0.00$

$\text{tf}_{tick} = 1$, $\text{idf}_{tick} = 0.30 \longrightarrow \text{tf-idf}_{tick} = 1 \times 0.30 = 0.30$

Score of d_2 = 0.00 + 0.30 = **0.30 Ans.**

For d_3, $\text{tf}_{clock} = 1$, $\text{idf}_{clock} = 0.12 \longrightarrow \text{tf-idf}_{clock} = 1 \times 0.12 = 0.12$

$\text{tf}_{tick} = 0$, $\text{idf}_{tick} = 0.30 \longrightarrow \text{tf-idf}_{tick} = 0 \times 0.30 = 0.00$

Score of d_1 = 0.12 + 0.00 = **0.12 Ans.**

For d_4, $\text{tf}_{clock} = 1$, $\text{idf}_{clock} = 0.12 \longrightarrow \text{tf-idf}_{clock} = 1 \times 0.12 = 0.12$

$\text{tf}_{tick} = 0$, $\text{idf}_{tick} = 0.30 \longrightarrow \text{tf-idf}_{tick} = 0 \times 0.30 = 0.00$

Score of d_1 = 0.12 + 0.00 = **0.12 Ans.**

Resultant document ranking: ***d*₁, *d*₂, *d*₃, *d*₄** OR ***d*₁, *d*₂, *d*₄, *d*₃ Ans.**

P.T.O.

Q. 4> Let the top ten documents returned by a search engine for three queries be graded for relevance as:

q_1: 0, 1, 1, 0, 0, 1, 1, 0, 0, 0
q_2: 1, 1, 1, 1, 0, 0, 0, 0, 1, 0
q_3: 1, 0, 1, 0, 0, 0, 1, 1, 1, 0

where 0 implies non-relevant and 1 implies relevant. The numbers of relevant documents for the three queries are 15, 20 and 25 respectively. Find the MAP for this result set. **(5)**

Soln. AP for q_1 = (1/2 + 2/3 + 3/6 + 4/7)/15 = 0.15
AP for q_2 = (1/1 + 2/2 + 3/3 + 4/4 + 5/9)/20 = 0.23
AP for q_3 = (1/1 + 2/3 + 3/7 + 4/8 + 5/9)/25 = 0.13
Thus, MAP = (0.15 + 0.23 + 0.13)/3 = **0.17 Ans.**