**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.** idf*car* = log10(*N*/df*t*) = log10(2000/520) = log103.85 = 0.59

idf*auto* = log10(*N*/df*t*) = log10(2000/618) = log103.24 = 0.51

idf*insurance* = log10(*N*/df*t*) = log10(2000/430) = log104.65 = 0.67

idf*best* = log10(*N*/df*t*) = log10(2000/790) = log102.53 = 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))/((5.902 + 7.652 + 3.352 + 10.002)1/2 x (17.702 + 6.122 + 5.362 + 4.002)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 x 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.

1. Find the length of a doc-id.
2. Find the size of the collection in MBs.
3. Find the size of the uncompressed posting file. **(0.5 x 3 = 1.5)**

**Soln. i.** Length of doc-id = log2500000 = log10500,000/log102 = 18.93 ≈ **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 x 109 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.** idf*clock* = log10(*N*/df*t*) = log10(4/3) = 0.12

idf*tick* = log10(*N*/df*t*) = log10(4/2) = 0.30

For *d*1, tf*clock* = 1, idf*clock* = 0.12 → tf-idf*clock* = 1 x 0.12 = 0.12

tf*tick* = 4, idf*tick* = 0.30 → tf-idf*tick* = 4 x 0.30 = 1.20

Score of *d*1 = 0.12 + 1.20 = **1.32 Ans.**

For *d*2, tf*clock* = 0, idf*clock* = 0.12 → tf-idf*clock* = 0 x 0.12 = 0.00

tf*tick* = 1, idf*tick* = 0.30 → tf-idf*tick* = 1 x 0.30 = 0.30

Score of *d*2 = 0.00 + 0.30 = **0.30 Ans.**

For *d*3, tf*clock* = 1, idf*clock* = 0.12 → tf-idf*clock* = 1 x 0.12 = 0.12

tf*tick* = 0, idf*tick* = 0.30 → tf-idf*tick* = 0 x 0.30 = 0.00

Score of *d*1 = 0.12 + 0.00 = **0.12 Ans.**

For *d*4, tf*clock* = 1, idf*clock* = 0.12 → tf-idf*clock* = 1 x 0.12 = 0.12

tf*tick* = 0, idf*tick* = 0.30 → tf-idf*tick* = 0 x 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.**