**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)**

**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 |

**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)**

**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)**

**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)**

**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)**