**BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI**

**PILANI CAMPUS**

**SECOND SEMESTER 2016 – 2017**

**Information REtriEval (CS F469)**

**MID SEMESTER EXAM**

**Date: 11.03.2017 Weightage: 30 % (60 Marks)**

**Duration: 90min. Type: Closed Book**

**Note: Answer all parts of the question together.**

**Answers must be brief.**

1. Answer the following multiple choice questions : [1 \* 10 = 10 marks]

a) Which of the following terms have the same soundex code?

A. Wainwrit and Wainwrite

B. Wainwrite and Wainwright

C. Both A and B

D. None of the above

b) If I search for term X, and term X has many synonyms, then precision is more likely to be a problem than recall.

A. True

B. False

C. Not enough information provided

c) A crawler gathers documents and sends them to an indexer, which employs the following modules,

(A) a stemmer;

(B) a language detector to detect the language of each document;

(C) a stop-word eliminator

(D) a filter that detects the format (pdf, Word, etc.) of the document.

Give the correct sequence in which the indexer should apply these modules to a document?

A. D B C A

B. A B C D

C. D C B A

D. B A C D

d) Variable-size arrays are used for building the postings lists. These have the following properties:

1. Less seek time and the corpus is dynamic
2. More seek time and the corpus is dynamic
3. Less seek time and the corpus is static
4. More seek time and the corpus is static

e) What is the query-document match score that the Jaccard coefficient computes for the (query, document) pair below?

Query: month of march

Document: John died in march

1. 1/6
2. 1/7
3. 2/6
4. 2/7

f) Consider 2 documents d and d’. d’ is generated by pasting the contents of d twice in d’. Which of the following is true?

A. Euclidean distance between (query, d) = (query, d’)

B. Cosine similarity between (query, d) = (query, d’)

C. Both A and B

D. Not enough information provided

1. Group-average agglomerative clustering (GAAC) is determined by:
2. Average similarity of all document pairs including those from the same cluster.
3. Average similarity of all document pairs including those from the same cluster but self-similarities are not included in the average.
4. Average similarity of all document pairs excluding those from the same cluster.
5. Average similarity of all document pairs excluding those from the same cluster but self-similarities are not included in the average.
6. Which is true about the Bernoulli model of text classification?
7. It estimates P(t|c) as the fraction of documents of class c that contain term t
8. It estimates P(t|c) as the fraction of tokens or fraction of positions in documents of class c that contain term t
9. It considers the number of occurrences of the term in the test document.
10. It does not consider the probability of non-occurrence of the terms of the vocabulary in the test document.
11. Consider the following sentence “ANY ANY SAT ON A BENCH” assume a unigram language model and calculate the probability of generating this sentence. Assume that your machine can generate only these words and all words probability are equal.
12. 0.000064
13. 0.000032
14. 0.000128
15. 0.000016
16. Given an English sentence e has m words e1,e2..em and Foreign sentence f has l words f1,f2..fl, the total no. of possible alignments are:
17. (l+1)m
18. (l)m
19. (l-1)m
20. (l)m-1

Q2. [5 + 2 + 3 + 4 = 14 marks]

a) Compute the Levenshtein distance between “CATS” and “NETWORKS” and mention the calculations in the matrix below. Also, state the formulae used.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **N** | **E** | **T** | **W** | **O** | **R** | **K** | **S** |
|  |  |  |  |  |  |  |  |  |  |
| **C** |  |  |  |  |  |  |  |  |  |
| **A** |  |  |  |  |  |  |  |  |  |
| **T** |  |  |  |  |  |  |  |  |  |
| **S** |  |  |  |  |  |  |  |  |  |

b) An IR system returns 8 relevant documents, and 10 non-relevant documents. There are a total of 20 relevant documents in the collection. What is the precision of the system on this search, and what is its recall?

c)What is the likely effect of ‘Stemming’ and ‘Lemmatization’ on:

* 1. Vocabulary size: Increase, Decrease, Unpredictable?
  2. Precision: Increase, Decrease, Unpredictable?
  3. Recall: Increase, Decrease, Unpredictable?

d) Consider the following documents:

Doc1: catholic church in brisbane

Doc2: garden city church brisbane

Doc3: brisbane courier garden city

Doc4: where in brisbane catholic church

1. Draw a term-document incidence matrix for this document collection.
2. Draw the positional inverted index representation for this collection.

Q3. Cluster to following documents using K-means with K=2. [8 marks]

D1: “data collection data”

D2: “data mining”

D3: “mining collection”

D4: “collection collection”

Assume D1 and D3 are chosen as initial seeds. Use tf without idf and without normalization. Use the distance measure as euclidean distance. Show the clusters and their centroids, after each iteration until convergence. Show the intermediate steps clearly.

Q4. [9 Marks]

Given below are two tables on a collection with a total of 1, 000, 000 documents, giving the tf values and the idf values for the 4 terms and 2 documents. Compute the scores and rank the given documents for the query “best car insurance”. Take the tf-idf weights with normalization.

Table: tf values

|  |  |  |
| --- | --- | --- |
| Term | Doc1 | Doc2 |
| Car | 27 | 24 |
| Auto | 3 | 0 |
| Insurance | 0 | 20 |
| Best | 4 | 7 |

Table: idf values

|  |  |
| --- | --- |
| Term | df*t* |
| Car | 10,000 |
| Auto | 5,000 |
| Insurance | 1,000 |
| Best | 50,000 |

Q5. [6 + 4 =10 marks]

a) Given below is the Table which lists the 5 documents in the training set and the appropriate class they belong. Also the test document is given.

1. Estimate the parameters of Naive Bayes classifier.
2. Apply the classifier to the test documentand classify whether it belongs to a1 or b1.

Table: Data for NB estimation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Docid | Words in document | In class a1 | In class b1 |
| Training Set | D1 | good | Yes | No |
|  | D2 | very good | Yes | No |
|  | D3 | bad | No | Yes |
|  | D4 | very bad | No | Yes |
|  | D5 | very bad very bad | No | Yes |
| Test Set | D6 | good bad very bad | ? | ? |

b) Consider the following document: “The universe contains many different universities”:

1. How many entries a bigram index would contain?
2. How do you process query “uni\*rse” by using the permuterm index? Show what terms will you search for and how?
3. Use the 2-gram index for processing the following wildcard query: tol\*. Is "tool" result for the wildcard query tol\*? If the answer is yes, how can we solve this problem ?

Q6. [2 + 7 = 9 Marks]

1. What do you mean by Query translation in Cross language information retrieval? What is the disadvantage in query translation method?
2. Given the following parallel corpus compute the IBM Model1 parameters for two iterations of EM algorithm.

|  |  |  |
| --- | --- | --- |
|  | e | f |
| S1 | this peacock | ye mor |
| S2 | blue peacock | neela mor |

END