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

**PILANI CAMPUS**

**SECOND SEMESTER 2016 – 2017**

**Information REtriEval (CS F469)**

**COMPREHENSIVE EXAMINATION**

**Date: 15.05.2017 Weightage: 25 % (50 Marks)**

**Duration: 2 hours Type: Closed Book**

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

**Answers must be brief.**

Q1. Fill in the blanks: [5 marks]

1. K-means clustering algorithm is \_\_\_\_\_\_\_\_\_ (hierarchical and soft/ flat and soft / flat and hard/ hierarchical and hard) clustering algorithm.
2. The two ways of handling phrase queries are \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (permuterm index / positional index) and \_\_\_\_\_\_\_\_\_\_\_\_ (n–gram index/ hash trees).

c) Limitations of Boolean retrieval model are absence of \_\_\_\_\_\_\_\_\_\_\_ (term frequency / document frequency) \_\_\_\_\_\_\_\_\_\_\_\_\_ (ranking/ indexing).

d) Kullback - Leibler divergence expresses the degree of \_\_\_\_\_\_\_\_\_\_\_\_\_ (similarity / discrepancy) between two probability distributions and is \_\_\_\_\_\_\_\_\_\_\_\_\_\_ (symmetric / asymmetric).

e) The three tests that a URL must pass before it gets added to the URL frontier list are \_\_\_\_\_\_\_\_\_\_\_\_\_\_, \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_, and \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

Q2. Given thefollowing phrase alignment matrix in Table 1 with English words as rows and Hindi words as columns for the given parallel corpus, answer the questions below:

 [2 + 1 + 2 + 1 + 2 +1+1= 10 marks]



Table 1

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  | C:\Users\Lavika\Desktop\Untitled.jpg | C:\Users\Lavika\Desktop\Untitled.jpg |  |
| Tomorrow | X |  |  |  |  |  |  |  |  |
| I’m |  | X |  |  |  |  |  |  |  |
| going |  |  |  |  |  |  |  |  | X |
| to |  |  |  |  |  |  |  |  | X |
| a |  |  |  |  | X |  |  |  |  |
| conference |  |  |  |  |  | X |  |  |  |
| in |  |  |  | X |  |  |  |  |  |
| Canada |  |  | X |  |  |  |  |  |  |

|  |  |
| --- | --- |
| 1. Assuming that the alignment matrix from question A is the intersection of P(f|e) and P(e|f),

identify whether the following phrase pairs are consistent with the alignment:  |  |
| 1. (Tomorrow I’mgoing, **)**
2. (a conference in Canada, कनाडा में एक सम्मेलन)
 |  |
| 1. Which is the longest phrase pair that is consistent with the alignment?
 |  |
| 1. Compute the reordering distance between the following 2 phrase pairs
 |  |
| i.(going to ,) |  |

ii. (in Canada, कनाडा में)

1. In the mathematical model of the phrase based translation why is the reordering distance not directly used but an exponentially decaying cost function is used?
2. If a spurious phrase translation pair occurs only once, how will you compute the phrase translation values. Show with the help of an example**.**
3. In the exponentially decaying cost function d = α |starti−endi−1−1)| , What should be the value of α if the movement of the phrases have to be penalized?
4. If a spurious phrase pair occurs only once in the whole parallel corpus, what will the value of Ф(f,e) and Ф(e,f)?

Q3. Consider the problem of classifying a name as being Food or Beverage. Assume the following training set in Table 2:

 [6 marks]

Table 2

|  |  |
| --- | --- |
| Document | Class |
| D1: “egg stuffing” | Food |
| D2: “chicken wings” |
| D3: “cream soda” | Beverage |
| D4: “Lemon soda” |

Apply the kNN algorithm with k=3, to classify the following test document D5: “egg soda” into its correct class? Use tf without idf and cosine similarity with length normalization for computation. Show the intermediate steps clearly.

Q4. [3 + 6 = 9 marks]

1. State the properties of SVD decomposition.
2. Given the following students to courses preference at BITS Pilani in Table 3 in the form of a ratingmatrix M where each row of M represents the given student’s set of ratings and the columns represent the courses:

Table 3



From the SVD it is evident that there are two concepts of courses here: the first three are Computer Science courses while the last two are Mechanical Engineering courses. Answer the questions:

1. After we decomposed M using SVD into U, S and VT , a new student named Rachat gave the following ratings: 4 for ML,5 for PPL, and 2 for Dynamics. Rachat can be represented as vector [0 4 5 2 0]. What is the representation of Rachat in the concept space?
2. Explain in one sentence what these values indicated about Rachat’s choice.
3. Another student named Vinay has the following reviews: 5 for IR, 2 for ML, 4 for Dynamics, and 5 for Mechanics. What is the representation of Vinay in the concept space?
4. Calculate the cosine similarity between Rachat and Vinay using their concept space vectors.

Q5.[2 + 4 = 6 marks]

 a) Given a grey scale image of size 5x5 pixels in Table 4 with the intensity range K=0,1,2,.., 255. Sketch the histogram to represent the image. [Note in the Y axis you may just show the value of intensities present in the image.]

Table 4

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **4** | **100** | **250** | **4** | **200** |
|  | **3** | **6** | **35** | **6** | **5** |
|  | **5** | **4** | **30** | **35** | **20** |
|  | **6** | **3** | **5** | **30** | **10** |
|  | **200** | **3** | **4** | **2** | **100** |

b) Given the color histograms for the query and the three images named a, b and c with each histogram having four colors: red, blue, purple, and yellow where the first bin shows number of red pixels, second bin shows blue, third bin shows purple and fourth bin shows yellow. Compute the Bray Curtis dissimilarity and Squared chord and rank the images based on both the distances.





Q6**.** Consider a sample corpus consisting of three documents given below: [9 marks]

D1: cat dog difference

D2: cat dog similarity

D3: horse rabbit difference

Query: cat cat difference

a) What are the similarity scores between the query (Q) with each document given above using Jaccard coefficient?

b) Compute the tf-idf score for each term in the document.

c) Compute the cosine score between the query (Q) and each document using the tf-idf of the terms computed in question (b) and rank the documents. Use only tf’s instead of tf-idf for calculating the tf-idf score for the query.

Q7. [5 Marks]

Consider the following web pages and the set of web pages they link to

Page A points to pages D.

Page B points to pages A and C.

Page C points to pages A and D.

Page D points to pages A, B and C.

Trace the page rank algorithm for two iterations and calculate P(A), P(B), P(C) and P(D) (in this order only). What is the order of the pages after the two iterations? Use initial PR values 1 for all nodes, use d=0.85.

Remember that the way to describe the algorithm is:

PR(A) = (1-d) + d(PR(T1)/L(T1) + … + PR(Tn)/L(Tn))

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

**PILANI CAMPUS**

**SECOND SEMESTER 2016 – 2017**

**Information REtriEval (CS F469)**

**COMPREHENSIVE EXAMINATION**

**Date: 15.05.2017 Weightage: 15 % (30 Marks)**

**Duration: 1 hour Type: Open Book**

|  |  |
| --- | --- |
| **NAME:** | **ID:** |
| **Total marks:**  | **Rechecks (if any):** |

**Instructions:**

* **Answers should be written on the question paper itself.**
* **All questions are multiple choice questions with one correct answer. Each question carries 3 marks. You are required to tick the correct option.**
* **No partial marks will be awarded.**

 [10 \* 3 = 30 marks]

Q1. Consider the following documents,

Doc1: Data analysis deals with data sets

Doc2: Data mining in various fields

Doc3: Data analysis techniques and algorithms

Doc4: Data sets with temporal attributes

When the Term Document incidence matrix is constructed for the above documents, what will be the total number of entries in the matrix?

A. 52

B. 50

C. 56

D. 64

Q2. Consider the following term frequency matrix for 2 documents. If we do not use idf -weighting, what will be the cosine score between the 2 documents?

|  |  |  |
| --- | --- | --- |
| Term | Document 1 | Document 2 |
| Land | 115 | 58  |
| Brazil | 10 | 7 |
| Cricket | 2 | 0 |

1. 0.94
2. 0.89
3. 0.79
4. 0.69

Q3. Use the Query likelihood model to calculate the probability of generating the query “frog likes toad”. Which of the following is true?

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | frog | Said | That | toad | likes | dog |
| Doc 1 | 0.01 | 0.03 | 0.04 | 0.01 | 0.02 | 0.005 |
| Doc 2 | 0.0002 | 0.03 | 0.04 | 0.0001 | 0.04 | 0.01 |

1. Doc 1 is more likely to generate query q
2. Doc 2 is more likely to generate query q
3. Both Doc 1 and Doc 2 are equally likely to generate query q
4. None of the Doc 1 and Doc 2 can generate the query q

Q4. Given a two-word query. The postings list of one term consists of the following 16 entries: [3,5,9,11,13,15,17,19,21,31,46,80,119,121,156,179] and for the other it is the one entry postings list: [31]. How many comparisons would be done to intersect the two postings lists with the following two strategies using postings lists stored with skip pointers, with a skip length as discussed in the class.

1. 4
2. 5
3. 6
4. 7

Q5. Consider an information need for which there are 4 relevant documents in the collection. Two queries are run on this collection and their results are shown below:

 *Query1: R N N N N N N N N R*

 *Query2: R R N N R R R N N R*

 Compute the Average Precision of each query and Mean Average Precision (MAP)?

1. 0.782
2. 0.682
3. 0.764
4. 0.6

Q6. Given the following Term Incidence matrix, list the documents retrieved for the Boolean query.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Bits  | Pilani | Students | Rock |
| D1 | 0 | 0 | 0 | 0 |
| D2 | 1 | 0 | 0 | 1 |
| D3 | 0 | 0 | 1 | 0 |
| D4 | 0 | 0 | 1 | 1 |
| D5 | 1 | 1 | 0 | 0 |
| D6 | 0 | 0 | 1 | 1 |
| D7 | 1 | 0 | 1 | 0 |
| D8 | 0 | 1 | 1 | 1 |

 | Query: (Bits AND Students) OR((NOT Students) AND (Pilani OR Rock)) |

1. D7, D5, D2
2. D7, D5, D1
3. D5, D2, D1
4. None of the above

Q7. Find the matrix L after the LU decomposition of the given matrix A using the shortcut method as discussed in the class:

$$\left[\begin{matrix}1&2&-3\\-2&1&5\\3&4&7\end{matrix}\right]$$

1. $\left[\begin{matrix}1&0&0\\2&1&0\\-3&2/5&1\end{matrix}\right]$
2. $\left[\begin{matrix}1&0&0\\-2&1&0\\3&-2/5&1\end{matrix}\right] $
3. $\left[\begin{matrix}1&0&0\\2&1&0\\-3&-2/3&1\end{matrix}\right]$
4. $\left[\begin{matrix}1&0&0\\-2&1&0\\-3&1/4&1\end{matrix}\right]$

Q8. Consider the following Dataset. Apply K-means algorithm with k=2 and the initial centroids values as C1= 9.3 and C2 = 15.7. What are the clusters generated after two iterations?

Dataset for Question

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X | 16.0 | 17.5 | 15 | 18 | 15.5 | 20 | 9 | 10 | 11.5 | 14.5 |

1. C1 = (9,10,11.5, 14.5), C2 =(16, 17.5,18,15.5,15,20)
2. C1 = (9,10), C2 =(11.5, 16, 17.5,18,15.5,15,20,14.5)
3. C1 = (9,10,11.5), C2 =(16, 17.5,18,15.5,15,20,14.5)
4. C1 = (9,10,11.5, 14.5, 15), C2 =(16, 17.5,18,15.5, 20)

Q9. Given the following alignment matrix, what will be the total no. of consistent phrase pairs that can be extracted?



1. 20
2. 23
3. 19
4. 24

Q10. Compute the rating of user E on item D using the ratings of user A and user D as the nearest neighbors to E?



1. 3.5
2. 2.5
3. 3
4. 4.5