Name: ____________________ Reg #: ____________________ Section: ______________

National University of Computer and Emerging Sciences, Lahore Campus

Course:	Information Retrieval and Text Mining	Course Code:	CS567
Program:	MS(Computer Science)	Semester:	Fall 2016
Duration:	180 Minutes	Total Marks:	59
Paper Date:	29-Dec-16	Weight	50%
Section:	ALL	Page(s):	10
Exam:	Final		

Instruction/Notes: Attempt the examination on the question paper and write concise answers. You can use extra sheet for rough work. Do not attach extra sheets used for rough with the question paper. Don't fill the table titled Questions/Marks.

Question	1	2	3	4	5	6	7	8	9	Total
Marks	/ 6	/ 3	/ 8	/12	/8	/6	/ 8	/ 3	/5	/ 59

Q1) Please choose either True or False for each of the following statements. For the statement you believe it is False, please give your brief explanation of it (you do not need to explain when you believe it is True) (the credit can only be granted if your explanation for the false statement is correct).[6 Marks]

1. Given a well-tuned unigram language model p(w|θ) estimated based on all the text books about the topic of "information retrieval", we can safely conclude that p("information retrieval"|θ)>p("retrieval information"|θ).

2. Assume we use Dirichlet Smoothing; duplicate the document content multiple times will not change the resulting smoothed document language model.

3. We do not use a database system to solve information retrieval problems mostly because of efficiency concern.

Q2) Please pick the most appropriate evaluation metric from Average Precision, Mean Reciprocal Rank, and Recall, for the following search tasks. [3 Marks]

a) A businessman searching for New York Time’s homepage for his breakfast reading.

b) A lawyer searching for all relevant evidence to one of his cases. The lawyer is evaluated by whether he could win the case and he bills his client by hours. Therefore he does not mind to read through all the documents that are returned by a search engine.

c) An American basketball fan searching for information and history for NBA. Some of the returned pages provide a lot of relevant details, for example, team rankings, match scores, the latest news, etc. Some pages are just marginally relevant. Others are less interesting or irrelevant.

Q3)a) Which of the following is most likely effective for increasing the PageRank score of a page: Encircle correct option. [2 Mark].

1. adding an inlink	Increase / Decrease / No effect
2. adding an outlink	Increase / Decrease / No effect
3. deleting an inlink	Increase / Decrease / No effect
4. deleting an outlink	Increase / Decrease / No effect

Q3)b) What important aspect of relevance does the NDCG metric take into account that precision, recall, and F-measure do not? [1 Mark]

Q3)c) Encircle correct option (True / False). [5 Marks]

1. Bernoulli NB classifier works better for long text documents as compared to Multinomial NB classifier. True / False
2. We should select K that maximizes RSS (Residual Sum of Squares) for KMeans clustering algorithm. True / False
3. Good initial seeds should be close to each other for KMeans clustering algorithm. True / False
4. Time Complexity of KMeans is O(n^2) where n is total number of documents. True / False
5. Macroaverage of an evaluation measure for multiclass classification problem gives equal weight to each class. True / False

Q4) Consider the following documents:

doc$_1$	phone ring person happy person
doc$_2$	dog pet happy run jump
doc$_3$	cat purr pet person happy
doc$_4$	life simple run happy
doc$_5$	life laugh walk run run

Q4) a) Smoothing is crucial in the language modelling approach to information retrieval. Why is smoothing important and how is it typically achieved? [2 Marks]

Q4) b) Construct the inverted index required for ranked retrieval for these five documents. Assume that no stemming or stop-word removal is required. (Store term frequency and term position in invereted index) [5 Marks]

Q4) c) Given the query {happy person smile}, show how a unigram language modelling approach would rank the documents outlined above. Choose a suitable form of smoothing and include all your workings. State any other assumptions made.[5 Marks]

Q5) a) Suppose that a web search engine has 100 terabytes of inverted lists. What is the total size of the inverted lists for the 3 most frequent words? Justify your answer. [3 Marks]

Q5) b) Let D be a document in a text collection. Suppose we add a copy of D to the collection. How would this affect the IDF values of all the words in the collection? Why? [3 Marks]

Q5) c) In what situation a system's Mean Average Precision performance will be equal to its Mean Reciprocal Rank performance? [2 Marks]

Q6) The goal of a retrieval model is to score and rank documents for a query. Different retrieval models make different assumptions about what makes a document more (or less) relevant than another. Suppose you issue the query "lemur" to a search engine. And, suppose that documents D101 and D123 both contain the term "lemur" twice . Answer the following questions. [6 Marks]

a) Would the ranked Boolean retrieval model necessarily give both documents the same score? If not, what information would determine which document is scored higher?

b) Would the cosine similarity necessarily give both documents the same score? If not, what would determine which document is scored higher?

c) Would the query-likelihood model (without linear interpolation) necessarily give both documents the same score? If not, what would determine which document is scored higher?

Q7) Suppose the PageRank algorithm is run on the graph in Figure 1 with all pages starting with the same rank.

a) Which page or pages will have the highest page rank in the network in Figure 1? [2 Marks]

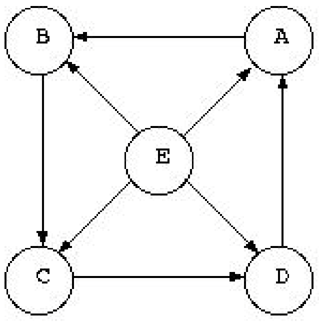


Figure 1

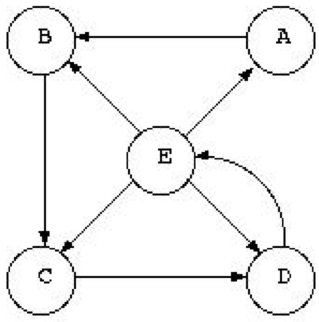
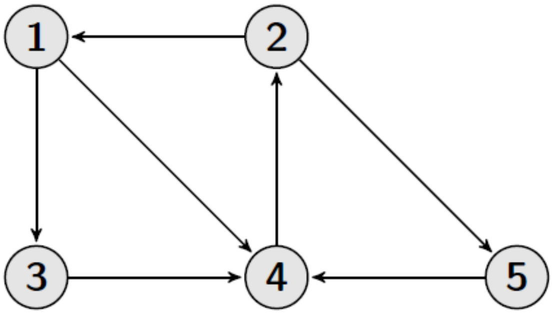


Figure 2

b) Suppose the network in Figure 1 is modified (by removing the link DA and introducing the new link DE) to produce Figure 2. Which page will now have the lowest page rank in Figure 2? Why? [2 Marks]

Q7) c) Consider a small web with 5 pages as shown below. Determine the transition probability matrix P of the Markov chain induced by PageRank for teleportation probability of 0.15 (we teleport to a random page with probability 0.15, with a uniform distribution over which particular page we teleport to). Compute the vector $\pi^{(1)}$ obtained after the first iteration of the power method, when using $\pi^{(0)} = 1/5 . [1\ 1\ 1\ 1\ 1\ ]$ as an initial state probability distribution. [4 marks]



Q8) a) Encode 14 using Elias Gamma Encoding [3 Marks]

b) Decode following number or numbers using Elias Gamma Decoding

111010011000

Q9) Based on the data below, estimate a Naive Bayes classifier using Laplace (add one) smoothing and apply the classifier to the test document. Estimate probabilities using **Bernoulli** method. Calculate the probability that the classifier assigns the test document to F = fruit or N = not fruit. [5 Marks]

	docID	**Words in document**	**class**
Training Set	1	Apple Orange Grapes	F
	2	Vitamin Apple	F
	3	Grapes Apple	F
	4	Computer Company	N
Test Set	5	Apple Apple Computer	?