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).

False and Explain unigram language model cannot capture the order of words

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

False and Explain duplicate the document will increase the document length, which affects Dirichlet Prior smoothing's results.

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

False and Explain: the major concern is that a database system cannot deal with unstructured text content.

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.

MRR

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.

Recall

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.

Average Precision

Q3) 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]

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]

Ans: It is need to avoid zero probability problem and it is achieved by giving some small non-zero probability to unseen query words in documents.

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]

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

Phone	doc$_1$ (1,1)
ring	doc$_1$ (1,2)
person	doc$_1$ (2,3,5) doc$_3$ (1,4)
happy	doc$_1$ (1,4) doc$_2$ (1,3) doc$_3$ (1,5) doc$_4$ (1,4)
Dog	Doc$_2$ (1,1)
pet	Doc$_2$ (1,2) doc$_3$ (1,3)
run	Doc$_2$ (1,4) doc$_4$ (1,3) doc$_5$ (2,4,5)
jump	Doc$_2$ (1,5)
cat	Doc$_3$ (1,1)
purr	Doc$_3$ (1,1)
life	Doc$_4$ (1,1) doc$_5$ (1,1)
simple	Doc$_4$ (1,2)
laugh	Doc$_5$ (1,2)
walk	Doc$_5$ (1,3)

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]

Laplace smoothing:

doc1 score = (2+1)/(5+14) * (1+1)/(5+14) *(0+1)/(5+14) = 0.00087 = 8.7 * 10^{-4}

doc2 score = (0+1)/(5+14) * (1+1)/(5+14) *(0+1)/(5+14) = 0.00029 = 2.9 * 10^{-4}

doc3 score = (1+1)/(5+14) * (1+1)/(5+14) *(0+1)/(5+14) 0.000583 = 5.8 * 10^{-4}

doc4 score = (0+1)/(4+14) * (1+1)/(4+14) *(0+1)/(4+14) = 0.000342 = 3.4 * 10^{-4}

doc5 score = (0+1)/(5+14) * (0+1)/(5+14) *(0+1)/(5+14) = 0.000145 = 1.4 * 10^{-4}

Ranking

doc1

doc3

doc4

doc2

doc5

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]

pr = A/rank = 0.1/rank

0.1/1 + 0.1/2 + 0.1/3 = 0.1 + 0.05 + 0.03 = 0.18 = 18% = 18 terabytes

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]

$$IDF(w) = \begin{cases} 1 + log(\frac{N+1}{DF(w)+1}) & \text{if } w \in D \\ 1 + log(\frac{N+1}{DF(w)}) & \text{otherwise} \end{cases}$$

where N is the original collection size, DF(w) is the original document frequency of word w. Therefore, when w occurs in D, its new IDF decreases (since N ≥ DF(w)); otherwise, its new IDF increases.

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

Any of the following situations will lead the same MRR and MAP performance:
_ There is only one relevant document in every query in the test collection.
_ There is no relevant document in every query in the test collection.
_ All documents associated with each query in the test collection are relevant.
_ We have perfect ranking under each query in the test collection.

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?

Ans: The ranked Boolean model scores documents based on the number of ways the document redundantly satisfies the query. In this case, we have a single-term query which happens to occur twice in each document. Therefore, each document would obtain a score of two. So, both documents would necessarily have the same score.

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

Ans: The cosine similarity is basically the inner-product divided by the vector length of the query times the vector length of the document. The vector length of the document is the square root of the number of unique terms. So, the scores given to both documents could be different, if the number of unique terms in both documents were different. The document with fewer unique terms would get a higher score.

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?

Ans: The query-likelihood model scores documents based on the probability of the query given the document language model. For a single-term query and assuming no linear interpolation, this results in the proportion of the text associated with the query term. In other words, the number of times the term occurs divided by the number of term-occurrences in the document. Because we don't know the number term-of term-occurrences in each document, we cannot say for sure that both documents would get the same score. The document with fewer term occurrences would get a higher score.

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]

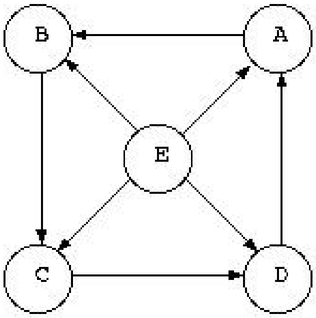


Figure 1

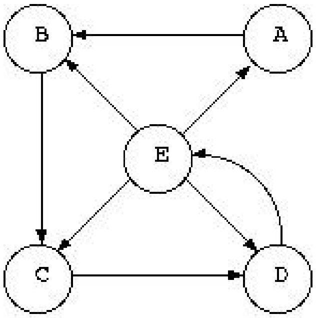


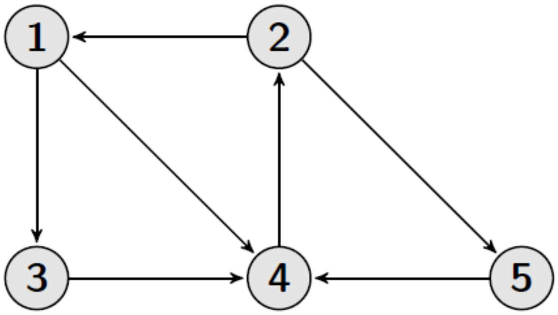
Figure 2

A, B, C, D

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]

A gets one fourth of the page rank of E. In addition to this contribution, B, C and D have additional PR inputs. Moreover, A receives one fourth the PR of E and so A can't have higher PR than E. Hence, A has the lowest page rank

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]



Ans: 0.115 , 0.2 , 0.115, 0.455, 0.115

Q8) a) Encode 14 using Elias Gamma Encoding [5 Marks]

1110110

b) Decode following number or numbers using Elias Gamma Decoding

1110100 11000

Ans: 1110100 11000 = 1100 100 = 12, 4

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	?

Solution: P(F) = 3/4, P(N) = 1/4

P (F | d5) = P(F) P(Apple |F) P(Computer | F) (1- P(Vitamin | F)) (1- P(Grapes | F)) (1- P(Orange | F)) (1- P(Company | F))

= 3/4 * 4/5 *4/5 * 1/5 * 4/5 * 3/5 * 2/5 * 3/5 = 0.0137

P N | d5) = P(N) (Apple |N) P(Computer | N) (1- P(Vitamin | N)) (1- P(Grapes | N)) (1- P(Orange | N)) (1- P(Company | N))

= 1/4 * 1/3 * 1/3 * 2/3 * 1/3 * 2/3 * 2/3 * 2/3 = 0.0054

Document 5 belongs to class Fruit since P (F | d5) > P(N | d5)