Information Retrieval **End-term Examinations [6 May]**

Name: **Roll Number:**

Instructions:

1. It is a close book examination.
2. Please write name, roll number, and project group number in the answer sheet.
3. Exam duration is 2 hours.
4. All questions are mandatory.
5. Calculators are not allowed.
6. There is no negative marking.
7. Total marks are 100.

Section 1: MCQs [2 marks each]

Question 1: Which of the following is True about PageRank?

A. Higher PageRank scores imply the page is less important
B. PageRank scores of pages that have a link to a given page influence its PageRank score
C. The higher the in-degree of links of a page, the higher is its long term visit rate (PageRank score)
D. The higher the out-degree of links of a page, the higher is its long term visit rate (PageRank score)

Answer: B, C
Explanation:
Higher the PR score, the higher the importance of the page
PageRank score of voters influences the scores of voted page
Higher in-degree implies a higher PR score

Question 2: The size of the dictionary in inverted index decreases when doing stemming
A. True B. False

Answer: True. Stemming decreases the size of the dictionary as multiple words are now mapped to their root word.

Question 3: Mark all correct statements.

A. Precision quantifies the number of positive class predictions that actually belong to the positive class.
B. Precision quantifies the number of positive class predictions made out of all positive examples in the dataset.
C. Recall quantifies the number of positive class predictions that actually belong to the positive class.
D. Recall quantifies the number of positive class predictions made out of all positive examples in the dataset.

Answer: A, D

Question 4: Mark all correct statements.

A. F-Measure provides a single score that balances both the concerns of precision and recall in one number.
B. F-Measure provides a single score that gives more importance to precision than recall.
C. F-Measure provides a single score that gives more importance to recall than precision.
D. We can accurately compute F-Measure for any search engine such as Google.

Answer: A

Question 5: Select the FALSE statement(s):

A. Hub score of a page P is calculated as the sum of the authority scores of all pages that point to P.
B. Authority score of a page P is calculated as the sum of the authority scores of all pages that point to P.
C. Hub score of a page P is calculated as the sum of the authority scores of all pages that P points to.
D. Authority score of a page P is calculated as the sum of the hub scores of all pages that P points to.

Answer: A, B, D

Hub score of a page P is calculated as the sum of the authority scores of all pages that P points to. Hence option C is correct.
Explanation:

$$h(x) \leftarrow \sum_{x \mapsto y} a(y)$$

Question 6: In the pagerank algorithm, the process of jumping to a random node from a dead-end with a certain probability is called:

A. Jump
B. Transfer
C. Teleport
D. Walk

Answer: C

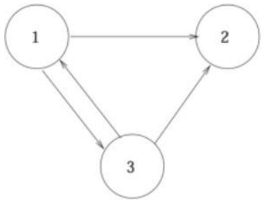
Question 7: Intersection operation of posting lists is always optimal when merged in decreasing order of list size.

A. True
B. False

Answer: False
Explanation: Consider three terms with postings list sizes s1 = 100, s2 = 105 and s3 = 110. Suppose the intersection of s1 and s2 has length 100 and the intersection of s1 and s3 length 0.

The ordering s1, s2, s3 requires 100+105+100+110=315 steps through the postings lists. The ordering s1,s3,s2 requires 100+110+0+0=210 steps through the postings lists.

Question 8: Assuming no teleportation, which node is expected to have the highest pagerank in the given graph?



A. 1
B. 2
C. 3
D. Insufficient Information

Ans: 2
Node 2 has no outgoing links due to which it will absorb all the pagerank during power iteration (2 is a dead node)

Question 9: While indexing a large set of documents, what might be a better option?

A. Add one document at a time to the index
B. Index multiple documents at once
C. Depends on the document content and length
D. Can’t say

Answer: B, Index multiple documents at once
Bulk indexing can be much faster than indexing individual documents because less number of merges (and hence disk seeks) may be required to construct the final index.

Question 10: There are N buyers competing for a single ad slot in a click-through auction. Consider the problem of maximizing welfare. Each buyer *i* is defined by their bid *v_i* that represents the value for a click, which is reported by the buyer and a click-through rate (CTR) *r_i* that specifies the probability of a click. The ideal solution for the auctioneer is to choose the ad:

A. Maximizing r_i + v_i
B. Maximizing r_i * v_i
C. Minimizing r_i + v_i
D. Minimizing r_i * v_i

Answer: B
The ideal solution for the auctioneer is to choose the ad maximizing the product r_i*v_i and then price according to a second price auction. Given on page 1 of the paper “Eligibility Mechanisms: Auctions Meet Information Retrieval”.

Question 11: Which of the following is TRUE about Recall?

A. Recall is a non-increasing function of the number of relevant docs retrieved.
B. Recall is a non-decreasing function of the number of relevant docs retrieved.
C. A system that returns all relevant docs has 100% recall.
D. A system cannot have 100% recall.

Answer: B, C

Question 12: Which of the following are TRUE?

A. Web search is an example of a precision-critical task.
B. Legal and patent search is an example of a precision-critical task.
C. Web search is an example of a recall-critical task.
D. Legal and patent search is an example of a recall-critical task.

Answer: A, D

Question 13: Which of the following are TRUE?

A. BM25 is sensitive to term frequency only
B. BM25 is sensitive to document length only
C. BM25 is sensitive to both term frequency and document length
D. BM25 is not sensitive to both term frequency and document length

Answer: C

Question 14: What's true about Webgraph?

A. It uses Adjacency lists.
B. WebGraph provides a way to store efficiently the URLs of a Web graph.
C. WebGraph is a framework that provides simple methods to manage very large graphs.
D. Algorithms For compressing Web graphs that exploit gap compression.

Answer: A, C, D

Question 15: Mark True or False for the options.

A. Hyperlink Induced Topic Search (HITS) is query independent.
B. PageRank is query independent.

Answer: A) False, B) True

Question 16: About Cumulative gain(CG) and Discounted cumulative gain(DCG) select options which are true:

A. CG is affected by changes in ordering of search results and hence, DCG is used for more accurate measure.
B. Highly relevant documents are more useful if appearing earlier in search result.
C. CG includes the position of a result in the calculation of gain of the result set.
D. DCG fluctuates a lot based on the length of the corpus.

Answer: B), D)

Question 17: Mark True or False for the options:

A. Normalized DCG metric penalizes for bad documents in the result.
B. Normalized DCG does not work well for query comparison when there are several equally good results.
C. The problem in using Normalized DCG is the unavailability of an ideal ordering of results when only partial relevance feedback is available.

Answer: A) False, B) True, C) True

Question 18: Between HITS and PageRank which of the followings are True:

A. HITS emphasizes mutual reinforcement between authority and hub webpages.
B. PageRank attempts to capture the distinction between hubs and authorities.
C. PageRank is a Link analysis algorithm based on a random surfer model.
D. PageRank and HITS only operate on a small subgraph from the web graph.

Answer: A), C)

Question 19: Mark A), B) as true or False.

A. In a Markov chain, the probability distribution of next states for a Markov chain depends only on the current state.
B. In a Markov chain, the probability distribution of next states for a Markov chain depends on how the Markov chain arrived at the current state.

Answer: A) True , B) False

Question 20: Given the options mark which are not correct about PageRank Algorithm.

A. The idea behind this algorithm is that pages visited more often in this walk are less important.
B. The idea behind this algorithm is that pages visited more often in this walk are more important.
C. In the PageRank algorithm the surfer proceeds in a random walk from node to node.
D. If the current location of the surfer has no out-links teleport operation is used.
E. If the current location of the surfer has no out-links, random walks are used.

Answer: A), E)

Section 2: Descriptive and Numerical Questions [12 marks each]

Question 21: Assume you have an IR system that aims to provide you legal documents for a given query. Assume that the system retrieved 20 documents out of which 2 were not relevant to the asked query on “SC verdict on Labor Law.”. Assume there are 1000 relevant documents for this case among 1000000000 documents which your IR system has indexed. Do the following:

A. Compute numbers for True Positive, True Negative, False Positive and False Negative
B. Compute precision, recall, and F1 score

Answer: TP: 18, FP: 2, FN: 982, TN: 1000000000 - 20 - 982
Compute precision, recall, and F1 score with respective formula.

Question 22: Assume you have an IR system (say, Snoogle) that always says there is no relevant document for the given query and returns nothing. Explain with an example which of the following evaluation metric would give best numbers (scores): precision, recall, F1-score, and accuracy. Can you trust the scores that you have got in the aforementioned situation? Justify which evaluation metric we should use in IR?

Answer: Snoogle demonstrates that accuracy is not a useful measure in IR. Simple trick to maximize accuracy in IR: always say no and return nothing You then get 99.99% accuracy on most queries. Since we will get very high accuracy in the above mentioned situation, accuracy is not a good metric here. Searchers on the web (and in IR in general) want to find something and have a certain tolerance for junk. It’s better to return some bad hits as long as you return something. We use (trust) precision, recall, and F for evaluation, not accuracy. Inverse relationship between precision and recall forces general systems to go for compromise between them But some tasks particularly need good precision whereas others need good recall. We should always average over a large set of queries. There is no such thing as a “typical” or “representative” query. We need relevance judgments for information-need-document pairs – but they are expensive to produce. For alternatives to using precision/recall and having to produce relevance judgments – see end of this lecture. Thus, F1-score would be a good evaluation metric in general but precision and recall would also be a good evaluation metric for Precision-critical task and Recall-critical task, respectively.

Question 23: Consider a list of 6 documents returned by your search engine namely, D1, D2, D3, D4, D5, D6 having relevance scores as 4, 2, 8, 2, 2, 4, respectively, for query 1 and for query 2 the search engine has returned the documents in the following order: D1, D6, D5, D2, D3, D4. Assume document relevance scores for query 2 same as query 1, i.e., 4 for D1, 2 for D2, etc. Compute MAP and Normalized Discounted Cumulative Gain (NDCG) for both queries. For MAP, assume that the documents having relevance score 4 and above are relevant and less than 4 are irrelevant documents. DCG formula is $DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i}$

Answer: Compute MAP and NDCG using respective formulas and consider the answer until the first place of decimal. 6 marks for MAP and 6 marks for NDCG. If the formula is correct and relevant scores are correctly used and only make mistakes in calculation then give 4 marks in respective MAP and NDCG. If the formula is correct and relevant scores are incorrectly used in calculation then give 2 marks in respective MAP and NDCG. If the formula is also wrong then give 0 marks.

Query 1 $\quad 4, 2, 8, 2, 2, 4$

$$DCG_1 = 4 + \frac{2}{\log_2 2} + \frac{8}{\log_2 3} + \frac{2}{\log_2 4} + \frac{2}{\log_2 5} + \frac{4}{\log_2 6}$$

$$= 14.48$$

Query 2 $\quad 4, 4, 2, 2, 8, 2$

$$DCG_2 = 4 + \frac{4}{\log_2 2} + \frac{2}{\log_2 3} + \frac{2}{\log_2 4} + \frac{8}{\log_2 5} + \frac{2}{\log_2 6}$$

$$= 14.51$$

$$IDCG = 8 + \frac{4}{\log_2 2} + \frac{4}{\log_2 3} + \frac{2}{\log_2 4} + \frac{2}{\log_2 5} + \frac{2}{\log_2 6}$$

$$= 17.17$$

$$nDCG_1 = \frac{DCG_1}{IDCG} = \frac{14.48}{17.17} = 0.843$$

$$nDCG_2 = \frac{DCG_2}{IDCG} = \frac{14.51}{17.17} = 0.845$$

Query 1 $\Rightarrow$ $1, 0, 1, 0, 0, 1$

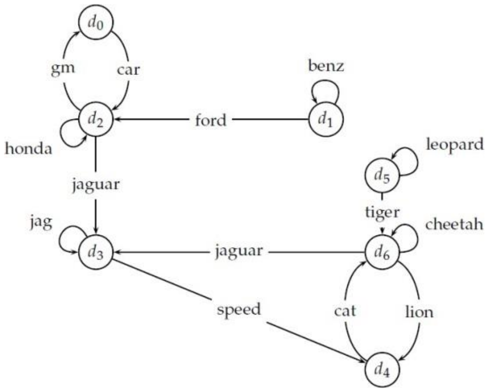
Avg Precision of Query 1 $= \frac{1}{3}\left(\frac{1}{1} + \frac{2}{3} + \frac{3}{6}\right) = 0.72$

Query 2 $\Rightarrow$ $1, 1, 0, 0, 1, 0$

Avg Precision of Query 2 $= \frac{1}{3}\left(\frac{1}{1} + \frac{2}{2} + \frac{3}{5}\right) = 0.86$

$$MAP = \frac{0.72 + 0.86}{2} = 0.79$$

Question 24: For the following web graph, arcs are annotated with the word that occurs in the anchor text of the corresponding link.



For a teleportation rate of 0.14 its (stochastic) transition probability matrix A is:

$$\begin{matrix} 0.02 & 0.02 & 0.88 & 0.02 & 0.02 & 0.02 & 0.02 \\ 0.02 & 0.45 & 0.45 & 0.02 & 0.02 & 0.02 & 0.02 \\ 0.31 & 0.02 & 0.31 & 0.31 & 0.02 & 0.02 & 0.02 \\ 0.02 & 0.02 & 0.02 & 0.45 & 0.45 & 0.02 & 0.02 \\ 0.02 & 0.02 & 0.02 & 0.02 & 0.02 & 0.02 & 0.88 \\ 0.02 & 0.02 & 0.02 & 0.02 & 0.02 & 0.45 & 0.45 \\ 0.02 & 0.02 & 0.02 & 0.31 & 0.31 & 0.02 & 0.31 \end{matrix}$$

The PageRank vector of this matrix is: $\vec{x} = (0.05 \quad 0.04 \quad 0.11 \quad 0.25 \quad 0.21 \quad 0.04 \quad 0.31)$

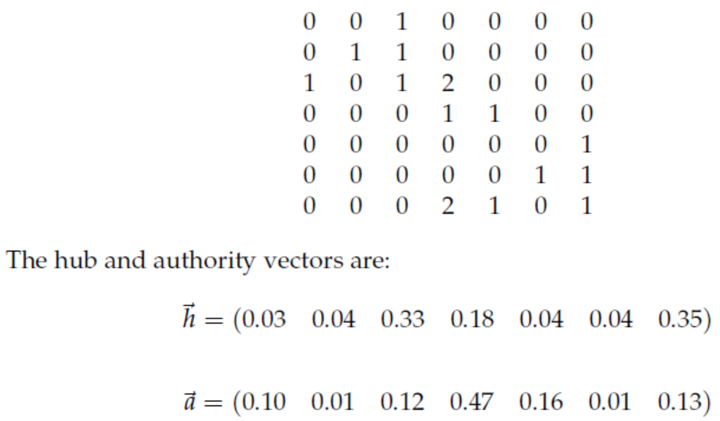
(i) What would be the PageRank vector in the next two steps. (4 marks)

Answer: Vector after next steps would be xA and xAA. If you used the correct formula and did not do the computation then give 2 marks. If used the correct formula and the computation for xA only then give 3 marks. If used the correct formula and the computation for xA and xAA both then give 4 marks. Ignore the mistake in computation here.

(ii) Which of the nodes have at least two in-links? Of these, which of the nodes has the lowest PageRank with a justification? (4 marks)

Answer: Observe that in the above Figure, q2, q3, q4 and q6 are the nodes with at least two in-links. Of these, q2 has the lowest PageRank since the random walk tends to drift out of the top part of the graph – the walker can only return there through teleportation.

(iii) Assuming the query *jaguar* and double-weighting of links whose anchors contain the query word, the matrix A for above Figure is as follows: (4 marks)



$$\begin{matrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 2 & 1 & 0 & 1 \end{matrix}$$

The hub and authority vectors are:

$$\vec{h} = (0.03 \quad 0.04 \quad 0.33 \quad 0.18 \quad 0.04 \quad 0.04 \quad 0.35)$$

$$\vec{a} = (0.10 \quad 0.01 \quad 0.12 \quad 0.47 \quad 0.16 \quad 0.01 \quad 0.13)$$

Which are the main authority and hubs in this scenario?

Answer: Here, q3 is the main authority – two hubs (q2 and q6) are pointing to it via highly weighted jaguar links.

Question 25: Poisson distribution is defined by $p(k) = \frac{\lambda^k}{k!} e^{-\lambda}$ where λ is the average rate. How is Poisson distribution applicable in IR and how is it helpful? Which of the following Poisson models is more appropriate in IR and why: (i) 1-Poisson or (ii) 2-Poisson? Which of the following is better and why: BM25 or VSM tf-idf? BM25 is defined by the following formula:

$$RSV^{BM25} = \sum_{i \in q} \log \frac{N}{df_i} \cdot \frac{(k_1 + 1)tf_i}{k_1((1-b) + b\frac{dl}{avdl}) + tf_i}$$

Answer: 2-Possion model is better. BM25 is better. See Lecture ProbIR for details. Students should provide some justification here around the discussion here.