IS392 Spring 2018 Midterm

This closed book exam starts at 1:00 pm and lasts until 3:55pm. Please remove all your personal belongings from the desk, including cell phones.

NJIT Honor Code is strictly enforced.

Please Print Your Name: _____

Part I. Match (*2 point each*)

A1. Accuracy	A2. Anchor Text	A3. Crawler	A4. Document Data Store
A5. Frequency	A6. Freshness	A7. Hash table	A8. Heap's Law
A9. Index	A10. Linked List	A11. Log Data	A12. Occurrence
A13. Page Title	A14. PageRank	A15. Parser	A16. Popularity
A17. Position	A18. Queue	A19. Recall	A20. Relevance Feedback
A21. Stemming	A22. Text Transformation	A23. Tokenizing	A24. Zipf's Law

1. Precision and _____ are the 2 effectiveness measures in IR; the latter cannot be determined in web retrieval environment.
2. _____ is a technique used for automatically expanding and refining users' query based on the relevant documents identified through user interaction.
3. _____ and _____ are the two major components shared by both the indexing and retrieval processes in the architecture of a search engine.
4. _____ is a software to identify and acquire documents for a search engine.
5. _____ is an appropriate data structure commonly used to store URLs in a web crawler.
6. _____ describes the relationship between vocabulary size and collection size.
7. Links can be viewed as information about the _____ of a web page.
8. _____ is usually a short text that succinctly describes the topic of the linked page, and is usually not written by the authors of the destination page.
9. _____-based index supports exact phrase match.

Part II. Questions

1. Assume a person searched for the query “indexing” on both Google and Bing. Then he/she identified the relevant pages (marked by “T”) and irrelevant ones (marked by “F”) from the top-10 results based on some relevance criteria. The results are shown in Table 1.

   What is the precision of each search engine given the query “indexing”? Which search engine performs the better? (*5 points*)

Google		Bing	
Relevance	**Page Title**	**Relevance**	**Page Title**
F	Indexing Definition | Investopedia	F	Indexing - definition of indexing by The Free Dictionary
F	Indexing Overview — FamilySearch.org	F	Indexing Definition | Investopedia
F	Get Started with Indexing — FamilySearch.org	F	FamilySearch - Official Site
T	Search engine indexing - Wikipedia	T	Search engine indexing - Wikipedia
T	Database index - Wikipedia	F	American Society for Indexing
T	Web indexing - Wikipedia	F	Get Started with Indexing — FamilySearch.org
F	Getting into Indexing - LDS.org	T	Indexing Service (Windows) - msdn.microsoft.com
F	American Society for Indexing	T	DocumentDB Indexing Policies | Microsoft Docs
T	Search Concept: Indexing | Swiftype	F	Indexing Has Moved — FamilySearch.org
T	Indexing — NumPy v1.12 Manual - Numpy and Scipy Documentation	T	Database index - Wikipedia

Table 1. Top-10 search results for “indexing” in Google and Bing

2. Explain ‘freshness’ measure in evaluating the document collection. Why is ‘freshness’ not a good measure of a collection? What alternative measure to use instead? and explain the alternative measure. (*5 points*)

3. Describe “BigTable” in detail, including logical table, tablets, row structure, and how to query it. Why is it preferred over relational DB for storing documents for search systems? (*7 points*)

4. What is Zipf's law? Please draw the Zipf's curve and explain it. (*6 points*)

5. Suppose a query has 3 words "tropical fish aquarium"; and N = total # of documents in the collection. Explain why does this formula: N x P (A) X P (B) X P(C), produces a poor estimation of the number of relevant documents? (*4 points*)

6. Please list all the over-lapping tri-grams out of the following phrase: "New Jersey Institute of Technology" (exclude the quotation marks.) (*2 points*)

7. Explain ‘stemming’ and ‘stopword removal’ and their respective pros and cons. (*6 points*)

8. Describe vocabulary mismatch, and word-sense ambiguity during query matching process. How would you solve these problems? (explain your solution). (*5 points*)

9. What is PageRank? Use Fig. 1 as an example web to calculate the PageRank score for page A, B, and C, without considering the “surprise me” button. You need to write your calculation steps in detail and calculate the PageRank values for the first 3 iteration. (*10 points*)

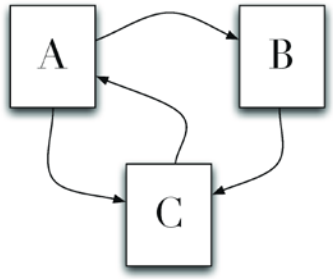

Fig. 1. An example web

Part III. Programming

1. Write a function using your preferred programming language to check whether a webpage is relevant to a certain topic. A webpage is considered as relevant if it contains **at least two different** related terms that are predefined based on the topic. Note that the relevance checking procedure should be **case-insensitive**.

   The objective function should have two input parameters shown in the Table 2, and it should return a Boolean value, i.e., "True" or "False" when a webpage is relevant or not respectively. (*7.5 points*)

Parameter	**Type**	**Meaning**
webpageContent	String	The text content of a webpage.
relatedTerms	List	A list of related terms.

Table 2. Input parameters of the objective function

2. Below is the pseudo code of a focused crawler that aims at downloading **500 unique** webpages that are relevant to a certain topic given a list of seed URLs. There is something wrong with the pseudo code. Try to find the error(s) and fix it (them). (*7.5 points*)

```
Procedure Crawl(SeedURLs)
1.      Q <- new Queue()
2.      counter <- 0
3.      For url in SeedURLs:
4.              Q.Append(url)
5.      End for
6.      While Q is not empty:
7.              url <- Q.RemoveFirstElement()
8.              page <- GetPageContent(url)
9.              counter <- counter + 1
10.             If IsRelevant(page):
11.                     SavePage(url, page)
12.             End if
13.             URLs <- ExtractURLsFrom(page)
14.             For url in URLs:
15.                     Q.Append(url)
16.             End for
17.     End while
End procedure
```

Part IV. Design

Draw a flowchart to show the procedure of the document indexing process which produces an inverted index with counts (term frequencies) shown in Fig. 2. You must use the special symbols and relationship operators shown in Fig. 3 to draw your flowchart. (*15 points*)

and	1:1			only	2:1		
aquarium	3:1			pigmented	4:1		
are	3:1	4:1		popular	3:1		
around	1:1			refer	2:1		
as	2:1			referred	2:1		
both	1:1			requiring	2:1		
bright	3:1			salt	1:1	4:1	
coloration	3:1	4:1		saltwater	2:1		
derives	4:1			species	1:1		
due	3:1			term	2:1		
environments	1:1			the	1:1	2:1	
fish	1:2	2:3	3:2	4:2	their	3:1		
fishkeepers	2:1			this	4:1		
found	1:1			those	2:1		
fresh	2:1			to	2:2	3:1	
freshwater	1:1	4:1		tropical	1:2	2:2	3:1
from	4:1			typically	4:1		
generally	4:1			use	2:1		
in	1:1	4:1		water	1:1	2:1	4:1
include	1:1			while	4:1		
including	1:1			with	2:1		
iridescence	4:1			world	1:1		
marine	2:1						
often	2:1	3:1					

Fig. 2. Inverted index with counts

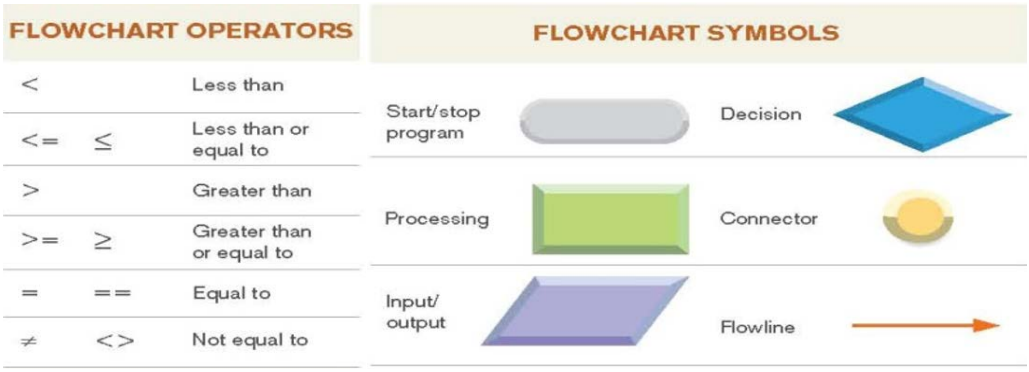

Fig. 3. Flowchart symbols and operators.