Student Name: _____

Andrew ID: _____

Seat Number: _____

Final Exam
Search Engines (11-442 / 11-642)
December 8, 2014

Answer all of the following questions. Each answer should be thorough, complete, and relevant. **Points will be deducted for irrelevant details.** Use the back of the pages if you need more room for your answer.

Calculators, phones, and other computational devices are not permitted. If a calculation is required, write fractions, and show your work so that it is clear that you know how to do the calculation.

1. **Distributed vs. Selected Search**

   You have a collection of 100 million documents that will be searched using 10 machines. Your boss wants you to use the selective search approach to searching this collection, using 1,000 equally-sized index shards. The system uses ReDDE as the resource selection algorithm, and searches 10 shards per query. Compare the selective search system with a distributed search system that divides the collection randomly into 10 equally-sized shards and searches all shards. Thus, both configurations (distributed and selective) search 10 shards for each query.

   a. Real-world search systems often care about "latency" – the amount of a query takes to finish processing. When a system searches multiple shards in parallel, the latency is determined by the time taken to search the slowest shard plus any overheads or fixed search costs. Assume that your system can search up to 10 indexes in parallel. The typical distributed search system will have a latency equivalent to searching one 10 million document index because all 10 shards are processed in parallel. If selective search uses a 1% CSI for ReDDE, which system (selective search or distributed search) has a lower latency? What about a 10% CSI? Assume shards have a processing cost proportional to the number of documents in the index. **(12 points)**

   b. ReDDE is an unsupervised resource selection algorithm. You could also use a supervised resource selection algorithm, like the one discussed in class for federated search. Which type of resource selection algorithm – unsupervised or supervised – do you expect to be faster <u>at search time</u>? Why? Be sure to consider the cost of the features used. **[4 points]**

2. Diversification

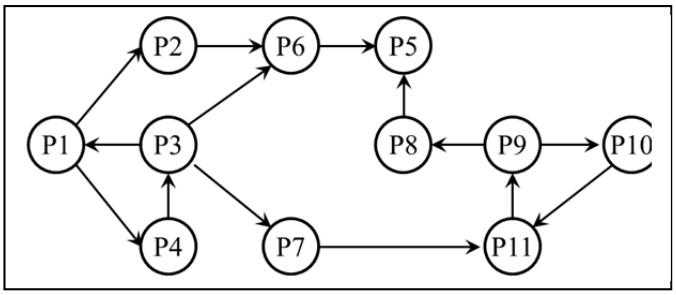
a. Describe why and when a search engine may need to diversify its search results. In class, we discussed two types of queries that benefit from diversification. Your answer must address both query types. **[6 points]**

b. Provide an example query for each type described above, and show why it needs to be diversified (it is not enough to just state that it need to be diversified). Your examples must be original. You may not use a query that was in any of the lecture notes. **[4 points]**

c. Describe the procedures of implicit search result diversification and explicit search result diversification. Compare them, and describe their differences. **[11 points]**

3. HITS

a. The root set of a query is {P1, P2, P3, P4, P5} in the web graph shown below. Provide the authority and hub scores of each page after 2 iterations; if HITS would ignore the page, write n/a. The table below is provided to help guide your work. Scores can be fractions. If you need to calculate a square root, round the result to the nearest integer to simplify your calculation; be clear about how you did this so that we will understand your answer. **[12 points]**



Authority	Initial Value	Iteration 1	After Normalization	Iteration 2	After Normalization
P1					
P2					
P3					
P4					
P5					
P6					
P7					
P8					
P9					
P10					
P11					

Hub	Initial Value	Iteration 1	After Normalization	Iteration 2	After Normalization
P1					
P2					
P3					
P4					
P5					
P6					
P7					
P8					
P9					
P10					
P11					

b. At the end of iteration, which nodes have the highest hubs and authority scores? Why do they have the highest scores? **[4 points]**

4. Personalization

a. Personalization methods assume that a person's interests are known to the search engine. Describe how the search engine acquires this information, and how it represents it (i.e., what kind of model of the user it creates). **[9 points]**

b. Describe how the search engine can use its model of a person's interests to produce a personalized ranking for a query. **[6 points]**

c. Most personalization methods assume that your interests do not change. However, analysis shows that 6-8% of a person's queries are about topics that are atypical for that individual. How can a search engine provide a personalized search experience for these queries? **[6 points]**

5. Query Suggestions

a. Describe an approach to generating query suggestions for a web search engine. You may assume that you have access to a search engine log that contains a unique user id; the user's query; information about the search results and clicked web pages; and time stamps. If your method requires additional information in the search engine log, describe what it needs and why. **[8 points]**

b. Describe an approach to generating query suggestions in an enterprise search environment. You may assume that you have access to an enterprise search log if you wish, but this cannot be your only source of information. Your method must also use information that is specific to the enterprise. Describe how these suggestions may differ from the types of suggestions produced by a web search engine. **[8 points]**

6. Page quality

Suppose that you work for a web search company. Your job is to develop a classifier that will use the page url, inlink text, and contents (i.e., not the web graph) to determine the likelihood that a web page is of low quality ("spam"). Describe 5 features that you might use in your classifier. Explain why each feature might be expected to be effective. At least 2 of your features must use the page url, and at least 2 of your features must use the page contents. Partial or no credit will be given for features that are minor variations of the same idea. **[10 points]**