Sample Test Questions

These are designed to help answer the questions on Exam 2

**PLEASE** do not ask me to answer any of the questions
CSCI 572 Review of Concepts/Algs/Etc.  
Presented This Semester

• Algorithms
  • Porters (stemming)
  • Soundex (phonemes)
  • Block Sort-Based Indexing
  • Single-Pass In-Memory Indexing
  • Logarithmic Merge Indexing
  • Computing Cosine Scores
  • *Map/Reduce
  • *Google cloud
  • *PageRank
  • *HITS (ranking)
  • *Levenshtein (spell correction)
  • *K-Means Clustering
  • *Rocchio (classification)
  • *K-Nearest Neighbor (classification)

• Concepts (cont’d)
  – *Centroids (classification)
  – *Dendrogram (clustering)
  – *Pay-Per-Click Auctions
  – *Ontologies: WordNet, FreeBase, KnowledgeGraph

• Data Structures
  – Suffix trees (autocomplete)
  – Priority Queues (ranking)
  – *B-trees (holding the dictionary)
  – *Tries (spelling correction)

• Concepts
  – Precision (ranking)
  – Recall (ranking)
  – tp, fp, fn, tn formulation
  – Harmonic Mean (ranking)
  – F Measure (ranking)
  – Mean Average Precision
  – Discounted Cumulative Gain
  – *Ontologies
  – *WordNet: hyponym, meronym, etc

* indicates presentation in second half of the semester
CSCI 572 Review of Concepts/Algs/Etc. 
Presented This Semester

- **Techniques**
  - Crawling (distributed, techniques)
  - Crawling: robots.txt, depth vs. breadth
  - Crawling: URL representation, threading
  - Stemming, Stop words, Lemmatization
  - Inverted index, positional inverted index
  - N-grams
  - Cryptographic hashing (de-duplication)
  - Shingling (de-duplication)
  - Jaccard Similarity
  - Document Boolean Model
  - Document Vector Model
  - Cosine similarity
  - Euclidean distance
  - *Champion lists, net-score
  - *Edit distance

- **Techniques (cont’d)**
  - TF-IDF (matching)
  - *Purity Index (clustering)
  - *Rand Index (clustering)
  - *Microformats (snippets)

- **Other**
  - Zipf’s Law
  - Heaps Law
  - *Spelling Correction program (Norvig)
  - *Edit distance
  - *Auctions
Questions on PageRank

1. If PR(A) represents the PageRank of node A in directed graph G and C(A) is the number of outgoing links of A, define the PageRank formula.

2. Make sure you know how to compute the PageRank for simple graphs like the ones below.

3. What is a typical value for the damping factor in the PageRank calculation?

---

Average PR: 1.000

```
Page A
1.49

Page C
1.58

Page B
0.78

Page D
0.15

About
0.69

Product
0.69

More
0.69

Home
1.92

---
```
Questions on HITS

1. What are the names given to the two types of nodes determined by HITS?
2. Is the HITS algorithm carried out once the query is entered or after the query results have been determined?
3. What sort of graph is produced by HITS?
4. How does the HITS algorithm compute the Authority score?
5. How does the HITS algorithm compute the Hub score?
Search Engine Advertising

1. What is the difference between Google’s AdWords and AdSense programs?

2. Approximately how much money did Google take in on advertising last year: 500 million, 1 billion, 10 billion, 50 billion, 100 billion?

3. What does SEO stand for?

4. Does Google accept banner ads?

5. Describe the auction process Google uses to determine the placement of ads

6. Describe the four possible forms of keyword matching that Google offers
Map/Reduce Questions

1. What is Hadoop?
2. Name two companies offering cloud services.
3. For the word occurrences problem describe what the map phase does
4. For the word occurrences problem describe what the reduce phase does
5. Describe how Map/Reduce handles failure of a node
Knowledge Systems Questions

1. Describe the difference between a taxonomy and an ontology

2. Describe three types of knowledge typically held by KnowledgeBases

3. Is Wikipedia an ontology?

4. Is WordNet a KnowledgeBase?

5. WordNet uses the terms synset, hyponym, meronym and holonym; define them

6. The WikiMedia Foundation sponsors Wikipedia and WikiData; what is WikiData and how does it relate to Wikipedia?
Query Processing Questions

1. Five strategies were mentioned for speeding up the answering of queries; describe each one in two or three sentences.

2. Does Google get more queries from desktop computers or from mobile devices?

3. Does Google use human testers to evaluate search results?
Homework #4

1. Does Lucene come bundled with Solr?
2. What sort of inverted index is produced by Lucene?
3. What is the purpose of the Lucene tokenizers?
4. Solr uses two types of scoring to produce search results, what are they?
5. What port does Solr run on?
6. Define a database shard
7. What language is Solr written in?
Spell Checking and Correction

1. Name three types of spelling errors
2. There are two approaches for correcting a word that is not in the dictionary, what are they?
3. How are n-grams used to solve the spelling correction problem?
4. Define edit distance
5. What is the name of the dictionary used in Norvig’s spelling corrector program?
6. Does Norvig’s spelling corrector program only use corrections of edit distance 1 or does it use both edit distance 1 and 2?
7. What are the computing time and space requirements for the Levenshtein algorithm?
Snippets

1. Are snippets always taken from the body of a web page?
2. Does the location of the sentence that contains the query keywords matter when it comes to choose which sentence to display as a snippet?
3. Does Google use meta information in a web page to determine a snippet?
4. Define "rich snippets"
5. Where can you find the specification for rich snippets?
6. Is the schema.org formalism an object hierarchy?
Clustering

1. Name a search engine other than Google/Bing/Yahoo that emphasizes clustering of results
2. Describe the difference between hard clustering and soft clustering
3. What definition of similarity is used by the k-means clustering algorithm?
4. Describe the optimal k-means optimization problem; is it NP-Hard?
5. How are the k means typically chosen in the k-means algorithm?
6. What is the computing time for the k-means algorithm?
7. Describe the agglomerative clustering algorithm
8. What is a dendrogram?
Question Answering

1. Name the four “W” questions used by question/answering systems
2. Do question answering systems use search engine results to help answer questions?
3. What are the TREC conferences?
4. There are five general approaches to question answering; describe each one
5. Define the Mean Reciprocal Rank
Classification

1. How does classification differ from clustering?
2. Define a centroid of a set of n points, each point a vector of length m
3. The Rocchio algorithm has two phases, what are they?
4. In the Rocchio algorithm how are the centroids determined?
5. Describe the k-nearest neighbor algorithm
6. What is the contiguity hypothesis?
7. How is k chosen in the k-nearest neighbor algorithm?
8. What is a Voroni diagram?