Name and Group: **Series ONE**

Email:

Information Retrieval Exam

Suppose we have the following documents:

Document	Words
D1	a b b a b b c
D2	a a b a b a
D3	b b b b b b c c

1. If my main concern is producing the fastest possibly dictionary search, which will be the best possible data structure to use?
   a) An array of words and word IDs over a contiguous chunk of memory, because this would allow very fast sequential passes through all dictionary words.
   b) A sorted linked list with skip pointers, because this would allow both very fast sequential passes and very fast look-ups.
   c) A hash table, because this would allow very fast look-ups.
   d) A B-tree, because this would allow on average the fastest look-up time for a word.

2. How will a basic inverted index for this corpus look like?
   a) a => D1 -> D2 -> D3; b => D1 -> D2 -> D3; c => D1 -> D2 -> D3
   b) a => D1 -> D1 -> D2 -> D2 -> D2 -> D2;
      b => D1 -> D1 -> D1 -> D1 -> D2 -> D2 -> D3 -> D3 -> D3 -> D3 -> D3 -> D3;
      c => D1 -> D3 -> D3
   c) a => D1 -> D2; b => D1 -> D2 -> D3; c => D1 -> D3
   d) a => D1:1,4 -> D2:1,2,4,6;
      b => D1:2,3,5,6 -> D2:3,5 -> D3:1,2,3,4,5,6;
      c => D1:7 -> D3:7,8.

3. What is the best order for processing the Boolean query << a AND b AND c >>?
   a) Intersect postings for „a" AND „c" first, then intersect result with postings for „b".
   b) Intersect postings for „a" AND „b" first, then intersect result with postings for „c".
   c) Intersect postings for „b" AND „c" first, then intersect result with postings for „a".
   d) It does not matter.

4. Suppose the corpus above is dynamic. Which of the following issues will you have to deal with?
   a) New or changed documents.
   b) Spell corrections.
   c) Deleted documents.
   d) Document access control lists, if any.

5. If I use simple TFxIDF for ranking, which will be the order generated for the regular query << a b >>?
   a) D1, D2, D3
   b) D2, D1, D3
   c) D1, D3, D2
   d) D2, D3, D1.

6. Using un-normalized cosine similarity on TFxIDF (just the dot product between the TFxIDF vectors of each document), which two of the three documents above are most similar to each other?
   a) (D1, D3)
   b) (D2, D3)
   c) All are equally similar
   d) (D1, D2)

7. Given the following sequence of gamma coded gaps, reconstruct the postings sequence: 110111011110010100.
   a) 2 -> 5 -> 9 -> 10 -> 11
   b) 4 -> 6 -> 10 -> 11
   c) 11 -> 14 -> 16400 -> 16402
   d) 7 -> 10 -> 20 -> 22.

8. How is it best to implement a basic indexing pipeline for Chinese?
   a) Just like a European one, but making sure we use UTF-8 in order to maintain the character encodings.
   b) Just like a European one, but splitting words using a specialized dictionary (in UTF-8).
   c) Just like a European one, but (1) splitting words using a specialized dictionary (in UTF-8) and (2) without removing stopwords, because they are not present in Asian languages.
   d) Just like a European one, but (1) making sure we use UTF-8 in order to maintain the character encodings and (2) indexing numbers separately, because they are used differently in Asian texts.

9. Suppose two words are considered similar if the Jaccard coefficient corresponding to their trigrams is greater than 0.5. Which of the following pairs contains similar words?
   a) (alone, alane)
   b) (alone, along)
   c) (alike, alite)
   d) (restaurant, restaurate)

10. Suppose a search returns documents D1, D2, and D3 in this order. The correct results in the system would have been D2, D1, D4, and D5 in this order. Which are the precision and recall for the engine in this case?
   a) P = 0.67; R = 0.5
   b) P = 0.5; R = 0.67
   c) P = 0.67; R = 0.4
   d) P = 0.4; R = 0.67

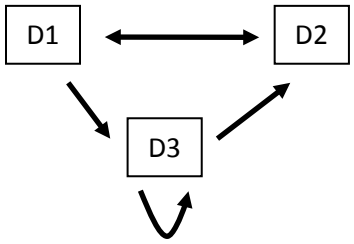
11. Which of the following are drawbacks for Mean Average Precision?
   a) It is not used in practice.
   b) There can be duplicate results in the output.
   c) There can be very similar results in the output.
   d) It does not provide information about the system itself, but about how people perceive it.

12. What is the most probable type of query for the web query “Microsoft”?
   a) Informational query
   b) Shopping query
   c) Downloads and documentation query
   d) Navigational query

13. Given the bow-tie structure of the web, which is the best place to start a crawl of all its documents?
   a) IN area
   b) OUT area
   c) SCC / CENTER area
   d) It does not matter, because the crawl will reach all pages anyway.

14. Which of the following must considered when crawling?
   a) Last modification date for each page
   b) Frequency of site accesses
   c) Robots rules
   d) Order in which pages are crawled.

Suppose documents D1, D2, D3 have the following hyperlink structure between them:



15. Which will be their scores and rankings if we order them using PageRank, computed with $\alpha = 0.1$ (probability to jump to a random page) for two iterations?
   a) D2=0.39, D3=0.32, D1=0.29
   b) D1=0.34, D2=0.34, D3=0.34
   c) D3=0.39, D2=0.32, D1=0.29
   d) D2=0.42, D3=0.40, D1=0.18.

16. Which of the following should be top priority design principles for Information Retrieval user interfaces?
   a) Accuracy
   b) Usability
   c) Feeling of user accomplishment
   d) Previous user experience.