Practice Questions

Subject:- Information Retrieval
Class: BE
Department:- Information Technology
Semester:- VIII

Note:- Answer key for each question is shown in **BOLD.**

1. Given a document collection which has 50 relevant documents, if an IR system retrieves 20 relevant and 23 irrelevant documents, what is the precision value of the system?
   A. 0.46
   B. 0.40
   C. 0.86
   D. 0.50
2. Steps of Indexing are performed in following order
   A. Stop-word Elimination, Tokenization, Stemming
   B. Tokenization, Stemming, Stop-word Elimination
   C. Tokenization, Stop-word Elimination, Stemming
   D. Stemming, Tokenization, Stop-word Elimination
3. In Information retrieval most common words such as articles, prepositions etc. are removed from tokens by using
   A. Stemming
   B. Stop-word elimination
   C. Indexing
   D. Ranking
4. Following is not a type of Rank-Based Measures
   A. Precision @ K
   B. Mean Average Precision
   C. Mean Reciprocal Rank
   D. Discounted cumulative gain
5. For a small collection of documents on a personal computer that don't experience any
   A. Block sort-based indexing algorithm
   B. Single-pass in memory indexing algorithm
   C. Distributed Map-Reduce indexing algorithm
   D. Dynamic indexing process employing an auxiliary index
6. Data stored in a table is a form of____
   A. Unstructured Data
   B. Structured Data
   C. Semi-Structured Data
   D. None of the Above
7. Recall is a

A. Fraction of retrieved documents that are relevant
B. Fraction of relevant documents that are retrieved
C. Both A and B
D. None of the above

8. Following are the example of classical models of IR
   A. The Boolean Model
   B. The vector Model
   C. Set-Based Model
   D. **All options are correct**
9. Which of the following are components of IR Model
   A. Crawling
   B. Indexing
   C. Query
   D. **All options are correct**
10. The IR problem can be defined as a
    A. 4-tuple
    B. 3-tuple
    C. 5-tuple
    D. 6-tuple
11. Inverted Index Dictionary is sorted by
    A. Term frequency
    B. Document Frequency
    C. Term/term ID
    D. DocID
12. Given a document containing the sentence **"I left my left bag at my home"** the number of tokens in the sentence is
    A. 8
    B. 4
    C. 6
    D. 1
13. Yahoo search engine uses stemming for its Index generation
    A. True
    B. False
14. Boolean Retrieval model maintains the **term frequency.**
    A. True
    B. **False**
15. The number of times that a word or term occurs in a document is called the:
    A. Proximity Operator
    B. Vocabulary Lexicon
    C. Term Frequency
    D. Indexing Granularity

16. A crude heuristic process that chops off the ends of the words to reduce inflectional forms of words and reduce the size of the vocabulary is called:
    A. Lemmatization
    B. Case Folding
    C. True casing
    D. Stemming

17. The formula used to estimate the vocabulary size of a collection is known as:
    A. Zipf's law
    B. Power law
    C. Heap's law
    D. Compression ratio

18. A compression algorithm that results in some loss of data is called:
    A. zipf compression
    B. dictionary compression
    C. lossless compression
    D. lossy compression

19. A metric derived by taking the log of N divided by the document frequency where N is the total number of documents in a collection is called:
    A. document frequency
    B. tf-idf weight
    C. collection frequency
    D. inverse document frequency

20. The tf-idf weight is highest when a term t occurs many times within a small number of documents.
    Select one:
    A. True
    B. False

21. Stemming increases the size of the vocabulary. Select one:
    A. True
    B. False

22. In information retrieval, extremely common words which would appear to be of little value in helping select documents that are excluded from the index vocabulary are called:
    A. **Stop Words**
    B. Tokens
    C. Lemmatized Words
    D. Stemmed Terms

23. The list of web pages that a web crawler has queued up to index is called the:
    A. Web Page Queue
    B. Seed set

C. URL Filter
D. **URL Frontier**

24. Recall is the fraction of non relevant documents that are retrieved.
    A. True
    B. False

25. What is the value of tf(computer, doc1) and tf(network, doc1) for following document Doc1: Computer network is a basic subject in Computer Engineering Branch.
    A. 1,1
    B. 2,1
    C. 2,2
    D. 1,2

26. The proportion of non-relevant documents that are retrieved, out of all non-relevant documents is nothing but
    A. Precision
    B. Recall
    C. Fall-out
    D. F-measure

27. Following metrics balances between precision and recall value
    A. R-precision
    B. F-score
    C. Fall-out
    D. DCG

28. Which of the following features can be used for accuracy improvement of a classification model?
    A. Frequency count of terms
    B. Vector Notation of sentence
    C. Part of Speech Tag
    D. All of these

29. To evaluate the effectiveness of an IR system the output from a standard query executed against the test IR system is compared with the known output from a:
    A. internet collection
    B. reference book
    C. Separate IR system.
    D. standard test collection

30. For a very large collection of books of classic literature the most appropriate indexing algorithm would be:
    A. Block sort-based indexing algorithm

B. Single-pass in memory indexing algorithm
C. **Distributed Map-Reduce indexing algorithm**
D. Dynamic indexing process employing an auxiliary index