IR Midterm 2021

1. _____ is the class of all tokens containing the same character sequence.
    a. A token
    b. A type
    c. A term
    d. None of the above

Answer: b. A type

2. _____ reduces variant forms to base form using natural language processing techniques.
    a. A stemmer
    b. A lemmatizer
    c. Both of them
    d. None of them

Answer: b. A lemmatizer

3. Given this query: Natural **AND** language **AND** processing. Let the document frequencies for faculty, artificial and intelligence are 10, 30 and 20 respectively. The best order for query processing is _____
    a. Intersect posting lists of natural and language then intersect the intermediate result with processing
    b. Intersect posting lists of language and processing then intersect the intermediate result with language
    c. Intersect posting lists of natural and processing then intersect the intermediate result with language
    d. None of the above

Answer: c. Intersect posting lists of natural and processing then intersect the intermediate result with language

4. Reducing all letters to lower case is a common strategy to perform _____
    a. Stemming
    b. Lemmatization
    c. Case-Folding
    d. Normalization

Answer: c. Case-Folding

5. _____ is the fraction of relevant documents in collection that are retrieved.
    a. Precision
    b. Recall
    c. F-measure

Answer: b. Recall

6. _____ is a normalized type that is included in the IR system's dictionary.
    a. Token
    b. Type
    c. Term
    d. Keyword

Answer: c. Term

7. Information retrieval systems can operate at three prominent scales which are web search, personal information retrieval and enterprise search.
    a. True
    b. False

Answer: a. True

8. Index granularity is the process of determining what the document unit that will be used for indexing
    a. True
    b. False

Answer: a. True

9. Information retrieval is finding material of structured nature that satisfies the user information need
    a. True
    b. False

Answer: b. False

10. Grepping through text is a practical solution to process proximity query
    a. True
    b. False

Answer: b. False

11. The dictionary of inverted index is commonly kept in memory, while posting lists are normally kept on disk
    a. True
    b. False

Answer: a. True

12. K-grams is one of the possible approaches for segmenting character sequences.
    a. True
    b. False

Answer: a. True

13. Stemmers use language-specific rules.
    a. True
    b. False

Answer: a. True

14. The step of converting byte sequence in a document into a sequence of characters is performed after tokenization.
    a. True
    b. False

Answer: b. False

15. The general trend in IR systems has been moved over time from standard use of quite large stop lists to very small stop lists to no stop list
    a. True
    b. False

Answer: a. True

16. More skip pointers lead to long skip spans and few successful skips.
    a. True
    b. False

Answer: b. False

17. QUERY is the topic about which the user desires to know more while information need is what the user conveys to the IR system in an attempt to communication the QUERY
    a. True
    b. False

Answer: b. False

18. English stemmers improve recall for some queries but harms precision on others
    a. True
    b. False

Answer: a. True

19. Tokenization is the process of canonicalizing tokens so that matches occur despite differences in the character sequences
    a. True
    b. False

Answer: b. False

Explanation: Normalization

20. One of the advantages of term-document incidence matrix is that it usually very sparse
    a. True
    b. False

Answer: b. False