1. Which operation has the longest latency?

a. Read 1 byte from memory
b. Read 1 byte from Solid-State Drive
*c. Read 1 byte from hard disk
d. Send 1 byte to another computer in the same datacenter

2. A search engine is given the query “used cars”, upon which it returns 500 documents. In total, there are 100 000 documents that are indexed by the search engine, 2000 of which are relevant for the given query. Out of these relevant documents, 200 are part of the returned results. What is the precision in this specific case?

a. 0.02
b. 0.10
c. 0.25
*d. 0.40

3. A search engine is given the query “used cars”, upon which it returns 500 documents. In total, there are 100 000 documents that are indexed by the search engine, 2000 of which are relevant for the given query. Out of these relevant documents, 200 are part of the returned results. What is the recall in this specific case?

a. 0.02
*b. 0.10
c. 0.25
d. 0.40

4. Given a search engine that has a precision of 0.2 and a recall of 0.3, what is its F1-Measure?

a. 0.06
b. 0.12
*c. 0.24
d. 0.27

5. What kind of things does an inverted index map to what kind of other things?

a. lists of terms are mapped to individual postings
b. lists of terms are mapped to lists of postings
c. individual terms are mapped to individual postings
*d. individual terms are mapped to lists of postings

6. What is the best strategy for processing the elements of a Boolean conjunctive query to optimize execution time?

a. Process in decreasing order of the estimated size of the respective intermediate results
*b. Process in increasing order of the estimated size of the respective intermediate results
c. Process in decreasing order of the estimated tf-idf values of the contained terms
d. Process in increasing order of the estimated tf-idf values of the contained terms

7. What can be estimated by applying Heaps' Law?

*a. The size of the dictionary
b. The length of posting lists
c. The similarity of documents
d. The query execution time

8. Which statement is true for an index and its text corpus of a realistic size?

*a. There are many more tokens than terms.
b. There are about as many tokens as terms, but not exactly.
c. There are exactly as many tokens as terms.
d. There are many more terms than tokens.

9. What is true about positional indexes?

a. Results can contain false positives for queries of more than two words
b. The index has about the same size as regular non-positional index
*c. They can be used to efficiently answer phrase queries
d. Documents are strictly treated as bag of words

10. Which of the statements below is true?

*a. A non-positional index is normally 2 or more times smaller than a positional index.
b. A non-positional index is normally about the same size as a positional index.
c. A non-positional index is normally 2 or more times bigger than a positional index.
d. A non-positional index is normally 10 or more times bigger than a positional index.

11. The tenth most frequent word in a typical text corpus is normally how frequent compared to the most frequent word?

a. about 99% as frequent as the most frequent word
b. about 90% as frequent as the most frequent word

c. about 50% as frequent as the most frequent word
*d. about 10% as frequent as the most frequent word

12. What is the edit distance between “abba” and “baab” assuming the four operations insert, delete, replace, and transpose?

a. 1
*b. 2
c. 3
d. 4

13. Calculate the Jaccard coefficient for the following pair of query and document: Document = “amsterdam berne cairo delhi”; Query = “amsterdam rotterdam”

a. 0.17
*b. 0.2
c. 0.33
d. 0.5

14. What is true about stop words and their tf and idf values?

a. Stop words have low tf and low idf values.
b. Stop words have low tf and high idf values.
*c. Stop words have high tf and low idf values.
d. Stop words have high tf and high idf values.

15. What determines the number of dimensions of a vector space model?

*a. the number of terms
b. the number of tokens
c. the number of documents
d. the number of queries

16. Which of the following is NOT a suitable data structure for the dictionary of an information retrieval system?

a. skip list
b. hash table
*c. linked list
d. binary tree

17. What should be looked up in a permuterm index for query 'pre*tion'?

a. pre$tion*
b. tion*pre$
c. pre*tion$
d. tion$pre

18. If we create a term-document incidence matrix for 50 documents, which each contains exactly 100 tokens, how many non-zero cells (that is, cells with a 1 instead of a 0) will this matrix maximally contain, given that we know that exactly 2000 distinct terms exist overall in these documents?

a. 2000
*b. 5000
c. 100 000
d. 200 000

19. The following are entries of a positional index for the terms “beer” and “month”, each showing the entries for one of the four documents 201, 202, 203, and 204. Which document could contain the phrase “beer of the month”?

*a. document 201: <beer: ... 201: 4, 30; ... > <month: ... 201: 8, 33; ... >
b. document 202: <beer: ... 202: 26; ... > <month: ... 202: 23, 42; ... >
c. document 203: <beer: ... 203: 8, 10; ... > <month: ... 203: 9; ... >
d. document 204: <beer: ... 204: 5, 15; ... > <month: ... 204: 1, 19; ... >

20. What is the inverse document frequency (using log10) for a term that appears in 10 documents – with 100 occurrences in these documents – in a text corpus of 1000 documents?

a. 1
*b. 2
c. 3
d. 4

21. What is P(b|a) according to Bayes' Theorem given P(a) = 0.5, P(b) = 0.7, and P(a|b) = 0.1?

a. 0.035
b. 0.071
*c. 0.14
d. 0.35

22. Given four classes A, B, C, and D, and a trained Naive Bayes model P(red|A)=0.5,

P(red|B)=0.1, P(red|C)=0.3, P(red|D)=0.2, P(A)=0.2, P(B)=0.1, P(C)=0.3, P(D)=0.4, which class is assigned to a document with content “red”?

*a. A
b. B
c. C
d. D

23. What is true about the distance of the support vectors to the separating hyperplane in the case of Support Vector Machines (SVM)?

*a. it is maximized during training
b. it is maximized when applied to classify new data
c. it is minimized during training
d. it is minimized when applied to classify new data

24. What is true about k-means?

a. k-means is a classification method
b. the algorithm calculates an optimal value for k
c. in each iteration, each centroid is assigned to its nearest document
*d. if run several times on the same input, k-means can return different results

25. Which is a meaningful measure for the distance between two clusters?

*a. Distance between the cluster centroids
b. Angle between cluster centroids
c. Average distance of documents to their centroid
d. Average angle between two documents of the same cluster as seen from their centroid

26. Documents that are retrieved but not relevant for a given query are called what?

a. true positives
*b. false positives
c. false negatives
d. true negatives

27. Given a corpus of documents, the number of documents in which a given term occurs at least once is called what?

a. collection frequency
*b. document frequency

c. inverse document frequency
d. term frequency

28. Agreement between raters can be calculated with what?

a. Mean Average Precision
b. F-measure
*c. Kappa statistics
d. Moore's law

29. A positional index is specifically useful for supporting what type of queries?

a. wildcard queries
*b. phrase queries
c. ad-hoc queries
d. conjunctive queries

30. A query is represented by what in the Vector Space Model?

*a. a vector
b. a dimension
c. a set of vectors
d. a set of dimensions

31. What kind of URLs are in the URL frontier?

*a. URLs that have been found in other pages but have not yet been fetched themselves
b. URLs that have been fetched but have not yet been found in other pages
c. Duplicates of URLs that have already been fetched
d. URLs from the robots.txt file that have already been fetched

32. What is the back queue of the URL frontier responsible for?

a. letting high-priority URLs pass through faster
*b. making sure that there is enough time between requests to the same server
c. eliminating URLs who have identical or very similar content
d. downloading the documents by making the HTTP requests and handing DNS resolution

33. What kind of network structure does the World Wide Web have?

a. long average distance and sparse
*b. short average distance and sparse

c. long average distance and dense
d. short average distance and dense

34. The hub-score of a web page in the HITS algorithm corresponds to what?

*a. the authority scores of the pages that the given page links to
b. the authority scores of the pages that link to the given page
c. the probability that a random surfer visits the page
d. the inverse probability that a random surfer visits the page

35. What is clustering in the context of information retrieval?

a. Assigning previously unseen documents to a given set of classes
b. Assigning previously unseen queries to a given set of documents
*c. Discovering classes for a given set of documents
d. Discovering documents for a given set of classes

36. Which is a property of small world networks?

a. most node pairs have a direct link
b. in-degree distribution is linear
*c. average distance between two nodes is short
d. two nodes that are linked to the same node tend not to be linked directly

37. Given the bow tie structure of the web (In, Central Core, Out) and selecting a random page in the Out part, from which pages in the In and Central Core parts can a user reach our selected random page by just following a sequence of links?

*a. From any page in Central Core and from any page in In
b. From any page in Central Core, but only from some pages in In
c. From any page in Central Core, but from none of the pages in In
d. Only from some pages in Central Core and only from some pages in In

38. Which is NOT a part of a normal crawl architecture?

a. Parsing
*b. Calculating page rankings
c. Duplicate elimination
d. URL filtering

39. What is the PageRank of a page that has 10 outgoing links but doesn't have any incoming links?

a. teleport probability
*b. teleport probability divided by the total number of nodes
c. teleport probability divided by 10
d. teleport probability divided by (10 times the total number of nodes)

40. What is true about Markov chains (of the type used to model PageRank)?

a. Which state is next depends on the history of previous states
b. The sequence of states is deterministic
c. Time is continuous
*d. States are linked by the probabilities of moving from one to the other

41. What set of pages is the HITS algorithm normally calculated on?

a. All pages containing the query string
b. All pages containing the query string and all pages that link to them
c. All pages containing the query string and all pages that are linked from them
*d. All pages containing the query string, all pages that link to them, and all pages that are linked from them

42. Which of the following is NOT a typical feature of Information Retrieval approaches following the vector space model?

*a. Well-defined queries with strict interpretation
b. False items in the results are tolerable
c. Unstructured data as main input
d. The answer to a query is a list of documents

43. Which of the following is NOT a well-formed entry for a non-positional inverted index?

a. amsterdam --> 1, 3
b. be --> 2
*c. cairo --> 3, 1
d. denver --> 3

44. Given three documents - document 1 “red red”, document 2 “red blue”, and document 3 “blue blue” - what is the resulting entry for “blue” for a non-positional inverted index?

a. blue --> 0, 1, 2
b. blue --> 1, 2

c. blue --> 2, 3, 3
*d. blue --> 2, 3

45. Let us assume that we have to process the query “jaguar AND car” in Boolean retrieval mode, and we retrieved the posting list [4, 5, 6] for jaguar, and [5, 7, 9] for car. Which is then the first action of merging the lists for answering the query?

a. Add number 4 from the jaguar list to the result set, and then move one position forward on that list
b. Add number 5 from the car list to the result set, and then move one position forward on that list
*c. Skip number 4 of the jaguar list by moving one position forward on that list
d. Skip number 5 of the car list by moving one position forward on that list

46. Different variants of tf-idf exist to rank documents according to a query. Which of the following is NOT a typical element of such variants?

a. plain (non-logarithmic) term frequency
b. logarithmically weighted term frequency
c. normalization via cosine similarity
*d. normalization via logarithmically weighted cosine similarity

47. If we treat the citation network of scientific publications as an undirected network, which other type of network is probably most similar to such a citation network in terms of its basic network structure?

a. Random network
b. Road network
*c. Airport network
d. Electricity grid network

48. What effect does Query Expansion typically have?

a. It reduces the time needed to process a query
b. It reduces the amount of memory needed to process a query
*c. It increases the recall of the results
d. It increases the precision of the results

49. What is true about Hierarchical Agglomerative Clustering (HAC)

a. The algorithm starts with a single cluster containing all documents
*b. Different distance metrics, such as single link or complete link, can be applied

c. The algorithm works by splitting a cluster into two clusters at each iteration
d. The number of clusters has to be specified beforehand

50. Which of the following is NOT a potential benefit of applying index compression techniques?

a. overall index size is smaller (so it might fit into memory)
b. memory consumption can be reduced (so more other things fit into memory)
*c. the posting lists contain fewer elements (and are therefore faster to iterate over)
d. data transfer from disk to memory is faster