Exam with solution

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 from the Netherlands to California and back

2. Who or what performs the assessments of a relevance benchmark when it is created?

*a. human experts
b. indexer
c. tokenizer
d. tf-idf weighting

3. A search engine is given the query “restaurants in Amsterdam”, upon which it returns 10'000 documents. In total, there are 100'000 documents that are indexed by the search engine, 4000 of which are relevant for the given query. Out of these relevant documents, 2000 are part of the returned results. What is the precision in this specific case?

a. 0.10
*b. 0.20
c. 0.40
d. 0.50

4. A search engine is given the query “restaurants in Amsterdam”, upon which it returns 1000 documents. In total, there are 100'000 documents that are indexed by the search engine, 5000 of which are relevant for the given query. Out of these relevant documents, 500 are part of the returned results. What is the recall in this specific case?

a. 0.01
*b. 0.10

c. 0.20
d. 0.50

5. Given a search engine has a precision of 0.6 and a recall of 0.4 for a given task. What is the F1-measure in this case?

a. 0.12
b. 0.24
*c. 0.48
d. 0.50

6. An inverted index maps each term to what?

a. another term
b. a posting
c. a list of terms
*d. a list of postings

7. What is used to estimate the size of the dictionary in an IR system?

a. F-measure
b. Gold Standard
c. Zipf's Law
*d. Heaps' Law

8. With the Block Sort-Based Indexing (BSBI) algorithm, for which step is it necessary to have an entire block loaded into memory?

*a. sorting
b. parsing
c. merging
d. none of the steps

9. What is true about biword indexes

*a. Results have no false positives for queries of two words
b. The index is typically smaller than a regular non-positional index

c. Documents are strictly treated as bag of words
d. They can be easily extended in practice to make them 3-word indexes, 4-word indexes, or above

10. Which step of creating an index is highly language-dependent?

a. inverting the index
*b. stemming
c. sorting
d. processing of wildcards

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

a. about 96% as frequent as the most frequent word
b. about 80% as frequent as the most frequent word
*c. about 25% as frequent as the most frequent word
d. about 12.5% as frequent as the most frequent word

12. Which type of index can efficiently handle wildcard queries, where the wildcard can appear anywhere in the query term?

a. positional index
*b. permuterm index
c. biword index
d. block sort-based index

13. What is the edit distance between “car” and “cure” assuming the operations insert, delete, and replace?

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

14. Calculate the Jaccard coefficient for the following pair of query and document: Document = “rabbit eats carrots”; Query = “Tom eats carrots”

a. 0.1667
b. 0.3333
*c. 0.5
d. 0.6667

15. What weighting scheme is used to increase the importance of infrequent words?

a. collection frequency
b. term frequency
*c. inverse document frequency
d. inverse term frequency

16. What is true about the tf and idf values of words like “the” and “of” in a typical corpus of documents in English?

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

17. How many dimensions would a vector space model have for a document collection of only two documents, where the content of document 1 is “the cat and the dog” and the content of document 2 is “the funny dog” (stopwords are NOT filtered but treated as regular words)?

a. 2
b. 4
*c. 5
d. 7

18. Which of the following operations can be performed with a dictionary that is based on a tree, but not with a dictionary that is based on hashes?

a. Fast lookup with constant time O(1)
*b. Efficiently enumerating all terms that start with a given character sequence

c. Retrieving the posting list for a term without possible false positives
d. Retrieving positional postings in the case of a positional index

19. What should be looked up in a permuterm index for query 'r*al'?

a. r$al
b. r$al*
c. $al*r
d. al$r

20. If we create a term-document incidence matrix for 200 documents, which each contains exactly 50 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 3000 distinct terms exist overall in these documents?

a. 3000
*b. 10 000
c. 150 000
d. 600 000

21. The following are entries of a positional index for the terms “a” and “pear”, each showing the entries for one of four documents. Which document could contain the phrase “I ate a pear”?

a. document 201: <a: ... 201: 1, 13; ... > <pear: ... 201: 2, 31; ... >
b. document 202: <a: ... 202: 7; ... > <pear: ... 202: 6, 43; ... >
*c. document 203: <a: ... 203: 8, 21; ... > <pear: ... 203: 9; ... >
d. document 204: <a: ... 204: 5, 15; ... > <pear: ... 204: 7, 13; ... >

22. What is the inverse document frequency (using log10) for a term that appears overall 10 000 times in 1000 documents in a text corpus of 1 000 000 documents?

a. 2
*b. 3

c. 4
d. 5

23. What is P(b|a) according to Bayes' Theorem given P(a) = 0.5, P(b) = 0.4, and P(a|b) = 0.2?

a. 0.04
*b. 0.16
c. 0.25
d. 0.32

24. 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)=0.1, P(B)=0.2, 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

25. What is maximized when training Support Vector Machines (SVM)?

a. The number of support vectors
b. The angle between the support vectors and the separating hyperplane
*c. The distance of the support vectors to the separating hyperplane
d. The distance between the hyperplanes (which are defined by the support vectors)

26. 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, documents are assigned to the nearest centroid

d. if run several times on the same input, k-means always returns the same result

27. Which is NOT a meaningful measure for the distance of two clusters?

a. Distance between the cluster centroids
*b. Angle between the cluster centroids
c. Minimum distance between two documents across the two clusters
d. Maximum distance between two documents across the two clusters

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

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

29. Given a corpus of documents, the number of occurrences of a term in these documents overall is called what?

*a. collection frequency
b. document frequency
c. inverse document frequency
d. term frequency

30. Agreement between raters can be calculated with what?

*a. Kappa statistics
b. Heaps' Law
c. Mean Average Precision
d. F-measure

31. In Boolean retrieval, a document is viewed as:

*a. a set of words
b. a bag of words
c. a list of characters
d. a vector of weights (e.g. tf-idf weights)

32. Which types of queries CANNOT be efficiently answered with just a positional index?

a. Boolean queries
*b. wildcard queries
c. long phrase queries
d. single-word queries

33. The vectors in the Vector Space Model represent what?

a. queries and terms
*b. documents and queries
c. postings and documents
d. terms and postings

34. What is true about the URLs in the URL frontier?

a. These URLs have been accessed (their content has been downloaded) but the content hasn't been processed yet
b. These URLs have been accessed and their content processed
*c. These URLs have been found in other pages but they have not yet been accessed themselves
d. These URLs have not yet been found in other pages

35. What is the front queue of the URL frontier responsible for?

a. eliminating duplicates
b. downloading documents
*c. managing prioritization
d. enforcing politeness

36. What is true about small-world networks?

a. they have low clustering and short average distances
b. they have low clustering and long average distances
*c. they have high clustering and short average distances
d. they have high clustering and long average distances

37. The authority score of a web page in the HITS algorithm corresponds to what?

a. the hub scores of the pages that the given page links to
*b. the hub scores of the pages that link to the given page
c. the PageRank scores of the pages that the given page links to
d. the PageRank scores of the pages that link to the given page

38. Which is NOT a property of small world networks?

a. for most pairs of nodes, their distance is greater than 1
*b. in-degree distribution is linear
c. average distance between two nodes is short
d. if two nodes are linked to the same node then they tend to be linked too

39. What does the content of the robots.txt file tell a web crawler?

a. The structure of the website, for example as a site map
b. The content of the website, for example by using Linked Data
c. Which pages of the website have changed since the last crawl
*d. Which pages of the website it is not allowed to access

40. Given the bow tie structure of the web (In, Central Core, Out), which pages can be reached by only following a sequence of links when starting from a random page in the Central Core?

a. Only some pages in the Central Core and any page in Out
b. Only some pages in the Central Core and only some pages in Out
c. Any page in the Central Core and only some pages in Out
*d. Any page in the Central Core and any page in Out

41. For a distributed crawler, URLs are normally distributed to different servers based on what?

*a. host
b. priority of the URL
c. number of duplicates
d. robots.txt

42. What is the PageRank of a page that doesn't have any incoming links?

a. teleport probability
*b. teleport probability divided by the total number of nodes
c. average PageRank of outgoing links
d. PageRank of outgoing links divided by the total number of nodes

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

a. A given initial condition always leads to exactly the same sequence of specific states
b. The states of a Markov chain correspond to the dimensions of the vector space
c. The next step depends on the current state and the complete history of previous steps
*d. Time is discrete and can therefore be represented with integers

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

*a. apple --> 3, 1
b. pear --> 3
c. orange --> 1, 3
d. the --> 2

45. Given two documents - document 1 “there we go again” and

document 2 “again and again and again and again” - what is the resulting entry for “again” for a non-positional inverted index?

*a. again --> 1, 2
b. again --> 1, 2, 2, 2, 2
c. again --> 5
d. again --> 1, 4

46. Let us assume that we have to process the query “apple AND computer” in Boolean retrieval mode, and we retrieved the posting list [4, 5, 6, 9] for apple, and [3, 4, 8] for computer. Which is then the first action of merging the lists for answering the query?

a. Add number 4 from the apple list to the result set, and then move one position forward on that list
b. Add number 3 from the computer list to the result set, and then move one position forward on that list
c. Skip number 4 of the apple list by moving one position forward on that list
*d. Skip number 3 of the computer list by moving one position forward on that list

47. What can be used to increase the recall of results?

*a. query expansion
b. index compression
c. duplicate detection
d. logarithmic weighting of term frequency

48. What is NOT 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 merging clusters at each iteration
d. The algorithm does not require us to define the number of clusters beforehand

49. What is a Spider Trap?

a. A malicious server that tries to artificially increase its PageRank
b. A malicious server that tries to break the indexer by providing infinite tf-idf values
*c. A malicious server that provides dynamic pages to trap crawlers in an infinite sequence of new pages
d. A malicious server that tries to trick users to click on pages that are not relevant and might contain spam

50. For the Information Retrieval technique called “relevance feedback”, who provides the feedback?

a. expert annotators
b. classification
c. clustering
*d. search users