MCQ

1) _________is the Activity of obtaining material which can be documents of unstructured nature

   a)Information Retrieval
   b)Boolean Retrieval
   c) Dictionary Retrieval
   d)Tolerant Retrieval

2) Three major component of Information Retrieval system are document subsystem ,Retrieval Subsystem and ______

   a) User Subsystem

   b) Query Subsystem

   c) Feedback Problem

   d) Subsystem

3) _______is the unit of information that we want to return as a result of

   a) Documents

   b) Collection

   c) Posting List

   d) Index

4) _______is the topic about which the user desires to know more
   a. Query
   b. Information Need
   c. Term
   d. Document

5) _______is the smallest unit of information in a query

   a. Term
   b. Document
   c. Word
   d. Index

6) Permuterm index is a index form of_______index

   a. Inverted
   b. Real
   c. Term
   d. Dictonary

7) A Query like mon is called as________wildcard query

   a. Trailing
   b. Heading
   c. New
   d. Real

8) _______algorithm is the algorithm for phonetic hashing

   a. Soundex
   b. Phonetic
   c. k-gram
   d. Edit distance

9) Edit distance is_______type of spelling correction
   a. Isolated term correction
   b. Context Sensitive correction
   c. k-gram correction
   d. both A and B

10) In Permuterm index______symbol is used to mark the end of a term
   a. $
   b. %
   c. &
   d.

11) ____is fraction of returned result relevant to information need

   a. Precision
   b. Recall
   c. Corpus
   d. Relavance

12) Within Documernt collection each document has unique serial number known as ____

   a. Document Identifier
   b. Document Number
   c. Document Id
   d. Both b and c

13) _______is the process of selecting how to organize the work of answering a query so that least amount of work need to be done by system

a. Query optimization
b. Query minimization
c. Query Answering
d. Both a and c

14) _______is a model for information retrieval in which we can pose any query which is in the form of Boolean Expression of terms

a. Boolean Retrieval
b. Dictonary retrieval
c. Tolerent retrieval
d. Both a and b

15) _____is a fraction of relavant documents in the collection were returned by the system

a. Precision
b. Recall
c. Corpus
d. Relavance

16) The idea to use computers for searching information was published in the article As We May Think by_____in 1945

a. Vannever Bush
b. Holmstrom
c. Gerard Salton
d. Both a and c

17) _____is the data structure for faster information retrieval which is collection of selected words and associated pointers

a. Index
b. Library
c. Metadata
d. Term

18) ______were the first to adopt Information retrival system for retriving information

a. Libraries
b. Google
c. NIST
d. Both a and c

19) In Document subsystem Abstracting contains

a. Summarizing
b. Bibliographic description
c. Acquisition
d. Both a and c

20) In Document Subsystem file organization contains term by term list of records under each term is called as_____

a. Inverted
b. Sequential
c. Combination
d. Both a and c

1) IR Stands for_______________.
a) Information Retrieval
b) Information Retired
c) Inform Retrieval
d) Information Ready

2) Each item in the list is called as_______________.
a) Items
b) Posting
c) Query
d) Information

3) etr term is called __________k-grams wildcard query.
a)3
b)4
c)1
d)2

4) To search document by ________________ in IR.
a)id
b)docID
c)number
d)#digits

5) SEO stands for ______________ .
a) Search English Optimization
b) Search Engine Optimization
c) Search Engine Operator

d) Search Engine Operation

6) Dictionary performed by _________________pair
a) Key and Value
b) Value and Number
c) Id and Number
d) Name and code

7) An advantage of a positional index is that it reduces the asymptotic complexity of a postings intersection operation.
A) True
B) False

8) _________can best be described as a programming model used to develop Hadoopbased applications that can process massive amounts of data.
A) MapReduce
B) Mahout
C) Oozie
D) All of the mentioned

9) The purpose of the inverse document frequency is to increase the weight of terms with high collection frequenc.
A) True
B) False
10) URL Stands for ______________________.
a) Uniform Ravar Location
b) Uniform Resource Locator
c) Uni Resource Locate
d) Uniform Reverse Locator

11) A data structure that maps terms back to the parts of a document in which they occur is called an
A) Postings list
B) Incidence Matrix
C) Dictionary
D) Inverted Index

12) The first large information retrieval research group was formed by____________at cornell in 1960.
a) Gerard Salton
b) Ratan Tata
c) Ramesh Bush
d) Think Roy

13) Input, Purpose and Output are the factors of _________ .
a) Summarization
b) Question Answering
c) Page Rank
d) Personalized Search

14) A deadlock can be broken down by

a) Committing one or more transactions
b) Aborting one or more transactions
c) Rolling back one or more transactions
d) Terminating one or more transactions.

15) NLTK stands for ______________ .
a) Natural Language Toolkit
b) Natural Lang Tool
c) Natural Long Tooltip
d) Nature Language Toolkit

A model of information retrieval in which we can pose any query in which search terms are combined with the operators AND, OR, and NOT:
Select one:

1. Ad Hoc Retrieval
2. Ranked Retrieval Model
3. Boolean Information Model
4. Proximity Query Model

The correct answer is: Boolean Information Model
Question **2**
A data structure that maps terms back to the parts of a document in which they occur is called an (select the best answer):
Select one:

1. Postings list
2. Incidence Matrix
3. Dictionary
4. Inverted Index

The correct answer is: Inverted Index
Question **3**
A process to efficiently intersect lists to be able to quickly find documents that contain both terms is referred to as merging postings lists.
Select one:
True
False
The correct answer is 'True'.
Question **4**
The model of information retrieval in which we can pose any query in the form of a Boolean expression is called the ranked retrieval model.
Select one:
True
False
The correct answer is 'False'.
Question **5**
The number of times that a word or term occurs in a document is called the:
Select one:

1. Proximity Operator
2. Vocabulary Lexicon
3. Term Frequency
4. Indexing Granularity

The correct answer is: Term Frequency
Question **6**

Stemming increases the size of the vocabulary.
Select one:
True
False
The correct answer is 'False'.
Question **7**
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:
Select one:

1. Stop Words
2. Tokens
3. Lemmatized Words
4. Stemmed Terms

The correct answer is: Stop Words
Question **8**
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:
Select one:

1. Lemmatization
2. Case Folding
3. True casing
4. Stemming

The correct answer is: Stemming
Question **9**
An advantage of a positional index is that it reduces the asymptotic complexity of a postings intersection operation.
Select one:
True
False
The correct answer is 'False'.
Question **10**
An index that includes sequences of words or terms of variable length that have been extracted from a source document is called a:
Select one:

1. Phrase Index
2. Biword index
3. Positional index
4. Inverted Index

The correct answer is: Phrase Index
One disadvantage, as outlined in our text, of using a permuterm index for wild card queries is:
Select one:

1. It requires complex code that is difficult to maintain
2. It has the risk of key collisions which are difficult to resolve
3. The required rotations creates a very large dictionary
4. It cannot be used to find terms that are not spelled correctly

The correct answer is: The required rotations creates a very large dictionary
Question **2**
Which of the following is a technique for context sensitive spelling correction:
Select one:

1. the Jaccard Coefficient
2. Soundex algorithms

3. k-gram indexes
4. Levenshtein distance

The correct answer is: Soundex algorithms

Question **3**

For a very large collection of books of classic literature the most appropriate indexing algorithm would be:

Select one:

1. Block sort-based indexing algorithm
2. Single-pass in memory indexing algorithm
3. Distributed Map-Reduce indexing algorithm
4. Dynamic indexing process employing an auxiliary index

The correct answer is: Distributed Map-Reduce indexing algorithm

Question **4**

For a large collection of documents such as the internet that experience frequent change the most appropriate indexing algorithm would be:

Select one:

1. Block sort-based indexing algorithm
2. Single-pass in memory indexing algorithm
3. Distributed Map-Reduce indexing algorithm
4. Dynamic indexing process employing an auxiliary index

The correct answer is: Dynamic indexing process employing an auxiliary index

Question **5**

Given two strings s1 and s2, the edit distance between them is sometimes known as the:

Select one:

1. Levenshtein distance
2. isolated-term distance
3. k-gram overlap
4. Jaccard Coefficient

The correct answer is: Levenshtein distance

Question **6**

For a moderately large collection of static documents maintained on a single system the most appropriate indexing algorithm would be:

Select one:

1. Block sort-based indexing algorithm
2. Single-pass in memory indexing algorithm
3. Distributed Map-Reduce indexing algorithm
4. Dynamic indexing process employing an auxiliary index

The correct answer is: Single-pass in memory indexing algorithm

Question **7**

For a small collection of documents on a personal computer that don't experience any change the most appropriate indexing algorithm would be:

Select one:

1. Block sort-based indexing algorithm
2. Single-pass in memory indexing algorithm
3. Distributed Map-Reduce indexing algorithm
4. Dynamic indexing process employing an auxiliary index

The correct answer is: Block sort-based indexing algorithm

Question **8**

Hashing is a process where an item is reduced, through a mathematical process, to an integer.

Select one:

True

False
The correct answer is 'True'.
Question **9**
The size of the document collection that can be indexed by single-pass in-memory indexing algorithm is limited by the size of the disk storage the computer running the indexer process has access to.
Select one:
True
False
The correct answer is 'False'.

The formula used to estimate the vocabulary size of a collection is known as:
Select one:

1. Zipf's law
2. Power law
3. Heap's law
4. Compression ratio

The correct answer is: Heap's law
Question **2**
Which of the following is NOT a benefit of index compression?
Select one:

1. Simplified algorithm design
2. Reduction of disk space
3. Faster transfer of data from disk to memory
4. Increased Use of caching

The correct answer is: Simplified algorithm design
Question **3**
A compression algorithm that results in some loss of data is called:
Select one:

1. zipf compression
2. dictionary compression
3. lossless compression
4. lossy compression

The correct answer is: lossy compression
Question **4**
An approach to compression that takes advantage of the redundancy in the dictionary that results from common prefixes that come from sorted terms is called:
Select one:

1. Front Coding
2. Blocked storage
3. Prefix Coding
4. Variable byte encoding

The correct answer is: Front Coding
Question **5**
A disadvantage of compression is that it reduces the transfer of data from disk to memory.
Select one:
True
False
The correct answer is 'False'.
Question **6**
The 30 most common words account for 30% of the tokens in written text is known as front

coding.
Select one:
True
False
The correct answer is 'False'.
Weighted zone scoring is sometimes referred to as ranked Boolean retrieval.
Select one:
True
False
The correct answer is 'True'.
Question **2**
In the bag of words model, the exact ordering of terms within the document is both significant and relevant to processing.
Select one:
True
False
The correct answer is 'True'.
Question **3**
The purpose of the inverse document frequency is to increase the weight of terms with high collection frequency.
Select one:
True
False
The correct answer is 'False'.
Question **4**
A scheme where a weight is assigned to a term based upon the number of occurrences of the term within a document is called:
Select one:

1. Bag of Words
2. Document Frequency
3. Term Frequency
4. Optimal weight

The correct answer is: Term Frequency
Question **5**
The number of documents within a collection that contain a particular term is the collection frequency of the term.
Select one:
True
False
The correct answer is 'False'.
Question **6**
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:
Select one:

1. document frequency
2. tf-idf weight
3. collection frequency
4. inverse document frequency

The correct answer is: inverse document frequency
Question **7**
The tf-idf weight is highest when a term t occurs many times within a small number of documents.

Select one:
True
False
The correct answer is 'True'.
Question **8**
The tf-idf weight is lower when a term t occurs many times in a document or occurs in relatively few documents.
Select one:
True
False
The correct answer is 'False'.
Question **9**
A measure of similarity between two vectors which is determined by measuring the angle between them is called:
Select one:

1. cosine similarity
2. sin similarity
3. vector similarity
4. vector scoring

The correct answer is: cosine similarity
Question **10**
An index that is often supplemental to the inverted index and contains terms from only a particular field or section of a document is called a parametric index.
Select one:
True
False
The correct answer is 'True'.
An approach to retrieval in a search that is likely (but not precisely) to produce the top K scoring documents is called:
Select one:

1. Exact top K document retrieval
2. top scoring document retrieval
3. Inexact top K document retrieval
4. Imprecise top K document retrieval

Question **2**
An approach to computing scores in an IR system that pre-computes for each term in the dictionary, the set of documents with the highest weights for the term is:
Select one:

1. Champion list
2. Impact ordering
3. Cluster pruning
4. Tiered indexes

Question **3**
An approach to computing scores in an IR system that orders documents in the posting list of a term by decreasing order of term frequency is called:
Select one:

1. Champion list
2. Impact ordering
3. Cluster pruning
4. Tiered indexes

Question **4**

An approach to computing scores in an IR system that selects a sample of documents randomly from the collection as leaders which are in the index and links similar documents to it (followers) is called:
Select one:

1. Champion list
2. Impact ordering
3. Cluster pruning
4. Tiered indexes

Question **5**
Which of the following items is not a component of a complete search system?
Select one:

1. Document cache
2. Indexers
3. Spell correction
4. Horizontal index

Question **6**
Which of the following is NOT one of the types of queries in a complete search system discussed in our text?
Select one:

1. Wildcard Query
2. Boolean retrieval
3. Phrase Query
4. Ranked retrieval Query

Question **7**
Considering only documents containing terms whose idf exceeds a preset threshold is an index elimination.
Select one:
True
False

Question **8**
A scoring function that computes an aggregate of a document's relevance from multiple sources is called evidence accumulation.
Select one:
True
False

CS 3308: INFORMATION RETRIEVAL
A scheme where a weight is assigned to a term based upon the number of occurrences of the term within a document is called:
Select one:
a. Bag of Words
b. Document Frequency
c. Term Frequency
d. Optimal weight
The correct answer is: Term Frequency

Question 2

Question text
A group of related documents against which information retrieval is employed is called:
Select one:

a. Corpus
b. Text Database
c. Index Collection
d. Repository
The correct answer is: Corpus

Question 3

Question text
Weighted zone scoring is referred to as:
Select one:
a. ranked Boolean retrieval
b. Zipf retrieval
c. Ad Hoc query retrieval
d. Jaccard retrieval
The correct answer is: ranked Boolean retrieval

Question 4

Question text
An approach to compression that takes advantage of the redundancy in the dictionary that results from common prefixes that come from sorted terms is called:
Select one:
a. Front Coding
b. Blocked storage
c. Prefix Coding
d. Variable byte encoding
The correct answer is: Front Coding

Question 5

Question text
True/False: Given two strings s1 and s2, the edit distance between them is sometimes known as the Levenshtein distance.
Select one:
True
False
The correct answer is 'True'.

Question 6

Question text
True/False: Ad hoc retrieval is a model of information retrieval in which we can pose any query in which search terms are combined with the operators AND, OR, and NOT.
Select one:
True
False
The correct answer is 'False'.

Question 7

Question text
True/False: An advantage of compression is that it reduces the transfer of data from disk to memory.
Select one:
True
False
The correct answer is 'True'.

Question 8

Question text
True/False: The process where multiple lists are evaluated using AND or OR operators in a Boolean retrieval query is called an intersection operation.
Select one:
True
False
The correct answer is 'True'.

Question 9

Question text
For a small collection of documents on a personal computer that don't experience any change the most appropriate indexing algorithm would be:
Select one:
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
The correct answer is: Block sort-based indexing algorithm

Question 10

Question text
True/False: The number of documents within a collection that contain a particular term is the collection frequency of the term.
Select one:
True
False
The correct answer is 'False'.

Question 11

Question text
The number of times that a word or term occurs in a document is called the:
Select one:
a. Proximity Operator
b. Vocabulary Lexicon
c. Term Frequency
d. Indexing
e. Granularity
The correct answer is: Term Frequency

Question 12

Question text
True/False: In the bag of words model, the exact ordering of terms within the document is not relevant to processing.
Select one:
True
False
The correct answer is 'True'.

Question 13

Question text
True/False: The Jaccard algorithm is a technique for context sensitive spelling correction.
Select one:
True
False
The correct answer is 'False'.

Question 14

Question text
True/False: Precision in an information retrieval system refers to the fraction of relevant documents in the collection that were returned by the system.
Select one:
True
False
The correct answer is 'True'.

Question 15

Question text
True/False: The purpose of the inverse document frequency is to increase the weight of terms with high collection frequency.
Select one:
True
False
The correct answer is 'False'.

Question 16

Question text
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:
Select one:
a. Stop Words
b. Tokens
c. Lemmatized Words
d. Stemmed Terms
The correct answer is: Stop Words

Question 17

Question text
A process that reduces the size of a vocabulary by reducing to the 'root' of words is called:
Select one:
a. Stemming
b. Lemmatizing
c. Removal of stop words
d. Posting
e. pruning
The correct answer is: Stemming

Question 18

Question text
A compression algorithm that results in some loss of data is called:
Select one:
a. zipf compression
b. dictionary compression
c. lossless compression
d. lossy compression
The correct answer is: lossy compression

Question 19

Question text
Which of the following is NOT a benefit of index compression?
Select one:
a. Simplified algorithm design
b. Reduction of disk space
c. Faster transfer of data from disk to memory
d. Increased Use of caching
The correct answer is: Simplified algorithm design

Question 20

Question text
True/False: tf-idf weight is 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.
Select one:
True
False
The correct answer is 'False'.

Question 21

Question text
True/False: Vector similarity is a measure of similarity between two vectors which is determined by measuring the angle between them.
Select one:

True
False
The correct answer is 'False'.

Question 22

Question text
True/False: Heap's law is the formula used to estimate the vocabulary size of a collection is.
Select one:
True
False
The correct answer is 'True'.
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:
Select one:
a. internet collection
b. reference book
c. separate IR system.
d. standard test collection
The correct answer is: standard test collection

Question 2

Question text
Precision is the fraction of retrieved documents that are relevant.
Select one:
True
False
The correct answer is 'True'.

Question 3

Question text
Recall is the fraction of non relevant documents that are retrieved.
Select one:
True
False
The correct answer is 'False'.

Question 4

Question text
Accuracy is typically the most accurate measure of IR system effectiveness.
Select one:
True
False
The correct answer is 'False'.

Question 5

Question text

The F-measure is a single measure that balances precision versus recall.
Select one:
True
False
The correct answer is 'True'.

Question 6

Question text
The purpose of the inverse document frequency is to increase the weight of terms with high collection frequency.
Select one:
True
False
The correct answer is 'False'.

Question 7

Question text
The standard approach to information retrieval system evaluation involves around the notion of:
Select one:
a. Quantity of documents in the collection
b. Relevant and non relevant documents.
c. Accuracy
d. user happiness
The correct answer is: Relevant and non relevant documents.
A web server communicates with a client (browser) using which protocol:
Select one:
a. HTML
b. HTTP
c. FTP
d. Telnet
The correct answer is: HTTP

Question 2

Question text
The basic operation of a web browser is to pass a request to the web server. This request is an address for a web page and is known as the:
Select one:
a. UAL: Universal Address Locator
b. HTML: Hypertext Markup Language
c. URL: Universal Resource Locator
d. HTTP: Hypertext transfer protocol
The correct answer is: URL: Universal Resource Locator

Question 3

Question text
A web page whose content doesn't vary from one request to another is called a:
Select one:

a. Text Page
b. Dynamic Page
c. Active Server Page
d. Static Page
The correct answer is: Static Page

Question 4

Question text
A web link within a web page that references another part of the same page is called a:
Select one:
a. Out link
b. Vector
c. In link
d. Tendril
The correct answer is: In link

Question 5

Question text
In the context of web search engines the manipulation of web page content for the purpose of appearing high up in search results for selected query terms is called:
Select one:
a. Paid inclusion
b. SPAM
c. SEO
d. Link Analysis
The correct answer is: SPAM

Question 6

Question text
Results from a search engine that are based upon the retrieval of items using a method of term weighting such as cosine similarity is a form of:
Select one:
a. Sponsored Search
b. Algorithmic Search
c. Informational Search
d. Navigational Search
The correct answer is: Algorithmic Search

Question 7

Question text
A program that captures and indexes content from web pages is known as what insect:
Select one:
a. Fly
b. Centipede
c. Mosquito
d. Spider
The correct answer is: Spider

Question 8

Question text
The list of web pages that a web crawler has queued up to index is called the:
Select one:
a. Web Page Queue
b. Seed set
c. URL Filter
d. URL Frontier
The correct answer is: URL Frontier

Question 9

Question text
In order to access a particular web site in the internet, the URL must be converted into an IP address. Which service does this conversion?
Select one:
a. HTTP
b. TNS
c. DNS
d. DHCP
The correct answer is: DNS
A model of information retrieval in which we can pose any query in which search terms are combined with the operators AND, OR, and NOT:
Select one:
a. Ad Hoc Retrieval
b. Ranked Retrieval Model
c. Boolean Information Model
d. Proximity Query Model
The correct answer is: Boolean Information Model

Question 2

Question text
A data structure that maps terms back to the parts of a document in which they occur is called an (select the best answer):
Select one:
a. Postings list
b. Incidence Matrix
c. Dictionary
d. Inverted Index
The correct answer is: Inverted Index

Question 3

Question text
A process to efficiently intersect lists to be able to quickly find documents that contain both terms is referred to as merging postings lists.
Select one:
True

False
The correct answer is 'True'.

Question 4

Question text
The model of information retrieval in which we can pose any query in the form of a Boolean expression is called the ranked retrieval model.
Select one:
True
False
The correct answer is 'False'.

Question 5

Question text
The number of times that a word or term occurs in a document is called the:
Select one:
a. Proximity Operator
b. Vocabulary Lexicon
c. Term Frequency
d. Indexing Granularity
The correct answer is: Term Frequency

Question 6

Question text
Stemming increases the size of the vocabulary.
Select one:
True
False
The correct answer is 'False'.

Question 7

Question text
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:
Select one:
a. Stop Words
b. Tokens
c. Lemmatized Words
d. Stemmed Terms
The correct answer is: Stop Words

Question 8

Question text
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:
Select one:

a. Lemmatization
b. Case Folding
c. True casing
d. Stemming
The correct answer is: Stemming

Question 9

Question text
An advantage of a positional index is that it reduces the asymptotic complexity of a postings intersection operation.
Select one:
True
False
The correct answer is 'False'.

Question 10

Question text
An index that includes sequences of words or terms of variable length that have been extracted from a source document is called a:
Select one:
a. Phrase Index
b. Biword index
c. Positional index
d. Inverted Index
The correct answer is: Phrase Index
One disadvantage, as outlined in our text, of using a permuterm index for wild card queries is:
Select one:
a. It requires complex code that is difficult to maintain
b. It has the risk of key collisions which are difficult to resolve
c. The required rotations creates a very large dictionary
d. It cannot be used to find terms that are not spelled correctly
The correct answer is: The required rotations creates a very large dictionary

Question 2

Question text
Which of the following is a technique for context sensitive spelling correction:
Select one:
a. the Jaccard Coefficient
b. Soundex algorithms
c. k-gram indexes
d. Levenshtein distance
The correct answer is: Soundex algorithms

Question 3

Question text
For a very large collection of books of classic literature the most appropriate indexing algorithm would be:

Select one:
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
The correct answer is: Distributed Map-Reduce indexing algorithm

Question 4

Question text
For a large collection of documents such as the internet that experience frequent change the most appropriate indexing algorithm would be:
Select one:
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
The correct answer is: Dynamic indexing process employing an auxiliary index

Question 5

Question text
Given two strings s1 and s2, the edit distance between them is sometimes known as the:
Select one:
a. Levenshtein distance
b. isolated-term distance
c. k-gram overlap
d. Jaccard Coefficient
The correct answer is: Levenshtein distance

Question 6

Question text
For a moderately large collection of static documents maintained on a single system the most appropriate indexing algorithm would be:
Select one:
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
The correct answer is: Single-pass in memory indexing algorithm

Question 7

Question text
For a small collection of documents on a personal computer that don't experience any change the most appropriate indexing algorithm would be:
Select one:
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
The correct answer is: Block sort-based indexing algorithm

Question 8

Question text
Hashing is a process where an item is reduced, through a mathematical process, to an integer.
Select one:
True
False
The correct answer is 'True'.

Question 9

Question text
The size of the document collection that can be indexed by single-pass in-memory indexing algorithm is limited by the size of the disk storage the computer running the indexer process has access to.
Select one:
True
False
The correct answer is 'False'.
The formula used to estimate the vocabulary size of a collection is known as:
Select one:
a. Zipf's law
b. Power law
c. Heap's law
d. Compression ratio
The correct answer is: Heap's law

Question 2

Question text
Which of the following is NOT a benefit of index compression?
Select one:
a. Simplified algorithm design
b. Reduction of disk space
c. Faster transfer of data from disk to memory
d. Increased Use of caching
The correct answer is: Simplified algorithm design

Question 3

Question text
A compression algorithm that results in some loss of data is called:
Select one:
a. zipf compression
b. dictionary compression
c. lossless compression
d. lossy compression
The correct answer is: lossy compression

Question 4

Question text
An approach to compression that takes advantage of the redundancy in the dictionary that results from common prefixes that come from sorted terms is called:
Select one:
a. Front Coding
b. Blocked storage
c. Prefix Coding
d. Variable byte encoding
The correct answer is: Front Coding

Question 5

Question text
A disadvantage of compression is that it reduces the transfer of data from disk to memory.
Select one:
True
False
The correct answer is 'False'.

Question 6

Question text
The 30 most common words account for 30% of the tokens in written text is known as front coding.
Select one:
True
False
The correct answer is 'False'.
Weighted zone scoring is sometimes referred to as ranked Boolean retrieval.
Select one:
True
False
The correct answer is 'True'.

Question 2

Question text
In the bag of words model, the exact ordering of terms within the document is both significant and relevant to processing.
Select one:
True
False
The correct answer is 'True'.

Question 3

Question text
The purpose of the inverse document frequency is to increase the weight of terms with high

collection frequency.
Select one:
True
False
The correct answer is 'False'.

Question 4

Question text
A scheme where a weight is assigned to a term based upon the number of occurrences of the term within a document is called:
Select one:
a. Bag of Words
b. Document Frequency
c. Term Frequency
d. Optimal weight
The correct answer is: Term Frequency

Question 5

Question text
The number of documents within a collection that contain a particular term is the collection frequency of the term.
Select one:
True
False
The correct answer is 'False'.

Question 6

Question text
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:
Select one:
a. document frequency
b. tf-idf weight
c. collection frequency
d. inverse document frequency
The correct answer is: inverse document frequency

Question 7

Question text
The tf-idf weight is highest when a term t occurs many times within a small number of documents.
Select one:
True
False
The correct answer is 'True'.

Question 8

Question text
The tf-idf weight is lower when a term t occurs many times in a document or occurs in relatively few documents.
Select one:
True
False
The correct answer is 'False'.

Question 9

Question text
A measure of similarity between two vectors which is determined by measuring the angle between them is called:
Select one:
a. cosine similarity
b. sin similarity
c. vector similarity
d. vector scoring
The correct answer is: cosine similarity

Question 10

Question text
An index that is often supplemental to the inverted index and contains terms from only a particular field or section of a document is called a parametric index.
Select one:
True
False
The correct answer is 'True'.
An approach to retrieval in a search that is likely (but not precisely) to produce the top K scoring documents is called:
Select one:
a. Exact top K document retrieval
b. top scoring document retrieval
c. Inexact top K document retrieval
d. Imprecise top K document retrieval
The correct answer is:Inexact top K document retrieval

Question 2

Question text
An approach to computing scores in an IR system that pre-computes for each term in the dictionary,the set of documents with the highest weights for the term is:
Select one:
a. Champion list
b. Impact ordering
c. Cluster pruning
d. Tiered indexes
The correct answer is: Champion list

Question 3

Question text
An approach to computing scores in an IR system that orders documents in the posting list of a term by decreasing order of term frequency is called:
Select one:
a. Champion list
b. Impact ordering
c. Cluster pruning
d. Tiered indexes
The correct answer is: Impact ordering

Question 4

Question text
An approach to computing scores in an IR system that selects a sample of documents randomly from the collection as leaders which are in the index and links similar documents to it (followers) is called:
Select one:
a. Champion list
b. Impact ordering
c. Cluster pruning
d. Tiered indexes
The correct answer is: Cluster pruning

Question 5

Question text
Which of the following items is not a component of a complete search system?
Select one:
a. Document cache
b. Indexers
c. Spell correction
d. Horizontal index
The correct answer is: Horizontal index

Question 6

Question text
Which of the following is NOT one of the types of queries in a complete search system discussed in our text?
Select one:
a. Wildcard Query
b. Boolean retrieval
c. Phrase Query
d. Ranked retrieval Query
The correct answer is: Ranked retrieval Query

Question 7

Question text
Considering only documents containing terms whose idf exceeds a preset threshold is an index

elimination.
Select one:
True
False
The correct answer is 'True'.

Question 8

Question text
A scoring function that computes an aggregate of a document's relevance from multiple sources is called evidence accumulation.
Select one:
True
False
The correct answer is 'True'.
An approach to computing scores in an IR system that orders documents in the posting list of a term by decreasing order of term frequency is called:
Select one:
a. Champion list
b. Impact ordering
c. Cluster pruning
d. Tiered indexes
The correct answer is: Impact ordering

Question 2
Question text
An approach to computing scores in an IR system that selects a sample of documents randomly from the collection as leaders which are in the index and links similar documents to it (followers) is called:
Select one:
a. Champion list
b. Impact ordering
c. Cluster pruning
d. Tiered indexes
The correct answer is: Cluster pruning

Question 3
Question text
Precision is the fraction of retrieved documents that are relevant.
Select one:
True
False
The correct answer is 'True'.

Question 4
Question text
A scoring function that computes an aggregate of a document's relevance from multiple sources is called evidence accumulation.
Select one:
True
False

The correct answer is 'True'.

Question 5
Question text
The size of the document collection that can be indexed by single-pass in-memory indexing algorithm is limited by the size of the disk storage the computer running the indexer process has access to.
Select one:
True
False
The correct answer is 'False'.

Question 6
Question text
A program that captures and indexes content from web pages is known as what insect:
Select one:
a. Fly
b. Centipede
c. Mosquito
d. Spider
The correct answer is: Spider

Question 7
Question text
A web link within a web page that references another part of the same page is called a:
Select one:
a. Out link
b. Vector
c. In link
d. Tendril
The correct answer is: In link

Question 8
Question text
A data structure that maps terms back to the parts of a document in which they occur is called an (select the best answer):
Select one:
a. Postings list
b. Incidence Matrix
c. Dictionary
d. Inverted Index
The correct answer is: Inverted Index

Question 9
Question text
A process to efficiently intersect lists to be able to quickly find documents that contain both terms is referred to as merging postings lists.
Select one:
True
False
The correct answer is 'True'.

Question 10
Question text
Which of the following items is not a component of a complete search system?
Select one:
a. Document cache
b. Indexers
c. Spell correction
d. Horizontal index
The correct answer is: Horizontal index

Question 11
Question text
An index that is often supplemental to the inverted index and contains terms from only a particular field or section of a document is called a parametric index.
Select one:
True
False
The correct answer is 'True'.

Question 12
Question text
A model of information retrieval in which we can pose any query in which search terms are combined with the operators AND, OR, and NOT:
Select one:
a. Ad Hoc Retrieval
b. Ranked Retrieval Model
c. Boolean Information Model
d. Proximity Query Model
The correct answer is: Boolean Information Model

Question 13
Question text
An approach to compression that takes advantage of the redundancy in the dictionary that results from common prefixes that come from sorted terms is called:
Select one:
a. Front Coding
b. Blocked storage
c. Prefix Coding
d. Variable byte encoding
The correct answer is: Front Coding

Question 14
Question text
Results from a search engine that are based upon the retrieval of items using a method of term weighting such as cosine similarity is a form of:
Select one:
a. Sponsored Search
b. Algorithmic Search
c. Informational Search
d. Navigational Search

The correct answer is: Algorithmic Search

Question 15
Question text
A web page whose content doesn't vary from one request to another is called a:
Select one:
a. Text Page
b. Dynamic Page
c. Active Server Page
d. Static Page
The correct answer is: Static Page

Question 16
Question text
An approach to retrieval in a search that is likely (but not precisely) to produce the top K scoring documents is called:
Select one:
a. Exact top K document retrieval
b. top scoring document retrieval
c. Inexact top K document retrieval
d. Imprecise top K document retrieval
The correct answer is: Inexact top K document retrieval

Question 17
Question text
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:
Select one:
a. internet collection
b. reference book
c. separate IR system.
d. standard test collection
The correct answer is: standard test collection

Question 18
Question text
Stemming increases the size of the vocabulary.
Select one:
True
False
The correct answer is 'False'.

Question 19
Question text
Python programing language you can't learn in 2021
True
False
The correct answer is 'False'.

Question 20
Question text

The purpose of the inverse document frequency is to increase the weight of terms with high collection frequency.
Select one:
True
False
The correct answer is 'False'.

Question 21
Question text
A disadvantage of compression is that it reduces the transfer of data from disk to memory.
Select one:
True
False
The correct answer is 'False'.

Question 22
Question text
The number of documents within a collection that contain a particular term is the collection frequency of the term.
Select one:
True
False
The correct answer is 'False'.

Question 23
Question text
For a moderately large collection of static documents maintained on a single system the most appropriate indexing algorithm would be:
Select one:
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
The correct answer is: Single-pass in memory indexing algorithm

Question 24
Question text
In order to access a particular web site in the internet, the URL must be converted into an IP address. Which service does this conversion?
Select one:
a. HTTP
b. TNS
c. DNS
d. DHCP
The correct answer is: DNS

Question 25
Question text
An index that includes sequences of words or terms of variable length that have been extracted from a source document is called a:
Select one:

a. Phrase Index
b. Biword index
c. Positional index
d. Inverted Index
The correct answer is: Phrase Index

Question 26
Question text
The number of times that a word or term occurs in a document is called the:
Select one:
a. Proximity Operator
b. Vocabulary Lexicon
c. Term Frequency
d. Indexing Granularity
The correct answer is: Term Frequency

Question 27
Question text
The model of information retrieval in which we can pose any query in the form of a Boolean expression is called the ranked retrieval model.
Select one:
True
False
The correct answer is 'False'.

Question 28
Question text
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:
Select one:
a. Stop Words
b. Tokens
c. Lemmatized Words
d. Stemmed Terms
The correct answer is: Stop Words

Question 29
Question text
The formula used to estimate the vocabulary size of a collection is known as:
Select one:
a. Zipf's law
b. Power law
c. Heap's law
d. Compression ratio
The correct answer is: Heap's law

Question 30
Question text
The tf-idf weight is lower when a term t occurs many times in a document or occurs in relatively few documents.
Select one:

True
False
The correct answer is 'False'.

Question 31
Question text
Wh ign
b. Reduction of disk space
c. Faster transfer of data from disk to memory
d. Increased Use of caching
The correct answer is: Simplified algorithm design

Question 32
Question text
Given two strings s1 and s2, the edit distance between them is sometimes known as the:
Select one:
a. Levenshtein distance
b. isolated-term distance
c. k-gram overlap
d. Jaccard Coefficient
The correct answer is: Levenshtein distance

Question indexing algorithm would be:
Select one:
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
A web server communicates with a client (browser) using which protocol:
Select one:
a. HTML
b. HTTP
c. FTP
d. Telnet
The correct answer is: HTTP

Question 35
Question text
An approach to computing scores in an IR system that pre-computes for each term in the dictionary, the set of documents with the highest weights for the term is:
Select one:
a. Champion list
b. Impact ordering
c. Cluster pruning
d. Tiered indexes
The correct answer is: Champion list

Question 36
Question text
The purpose of the inverse document frequency is to increase the weight of terms with high collection frequency.

Select one:
True
False
The correct answer is 'False'.

Question 37
Question text
Accuracy is typically the most accurate measure of IR system effectiveness.
Select one:
True
False
The correct answer is 'False'.

Question 38
Question text
In the context of web search engines the manipulation of web page content for the purpose of appearing high up in search results for selected query terms is called:
Select one:
a. Paid inclusion
b. SPAM
c. SEO
d. Link Analysis
The correct answer is: SPAM

Question 39
Question text
Considering only documents containing terms whose idf exceeds a preset threshold is an index elimination.
Select one:
True
False
The correct answer is 'True'.

Question 40
Question text
The list of web pages that a web crawler has queued up to index is called the:
Select one:
a. Web Page Queue
b. Seed set
c. URL Filter
d. URL Frontier
The correct answer is: URL Frontier

Question 41
Question text
Hashing is a process where an item is reduced, through a mathematical process, to an integer.
Select one:
True
False
The correct answer is 'True'.

Question 42
Question text
Which of the following is a technique for context sensitive spelling correction:
Select one:
a. the Jaccard Coefficient
b. Soundex algorithms
c. k-gram indexes
d. Levenshtein distance
The correct answer is: Soundex algorithms

Question 43
Question text
The standard approach to information retrieval system evaluation involves around the notion of:
Select one:
a. Quantity of documents in the collection
b. Relevant and non relevant documents.
c. Accuracy
d. user happiness
The correct answer is: Relevant and non relevant documents.

Question 44
Question text
Recall is the fraction of non relevant documents that are retrieved.
Select one:
True
False
The correct answer is 'False'.

Question 45
Question text
The basic operation of a web browser is to pass a request to the web server. This request is an address for a web page and is known as the:
Select one:
a. UAL: Universal Address Locator
b. HTML: Hypertext Markup Language
c. URL: Universal Resource Locator
d. HTTP: Hypertext transfer protocol
The correct answer is: URL: Universal Resource Locator

Question 46
Question text
For a small collection of documents on a personal computer that don't experience any change the most appropriate indexing algorithm would be:
Select one:
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
The correct answer is: Block sort-based indexing algorithm

Question 47

Question text
A compression algorithm that results in some loss of data is called:
Select one:
a. zipf compression
b. dictionary compression
c. lossless compression
d. lossy compression
The correct answer is: lossy compression

Question 48
Question text
The F-measure is a single measure that balances precision versus recall.
Select one:
True
False
The correct answer is 'True'.

Question 49
Question text
Weighted zone scoring is sometimes referred to as ranked Boolean retrieval.
Select one:
True
False
The correct answer is 'True'.

Question 50
Question text
For a large collection of documents such as the internet that experience frequent change the most appropriate indexing algorithm would be:
Select one:
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
The correct answer is: Dynamic indexing process employing an auxiliary index

Question 51
Question text
The tf-idf weight is highest when a term t occurs many times within a small number of documents.
Select one:
True
False
The correct answer is 'True'.

Question 52
Question text
An advantage of a positional index is that it reduces the asymptotic complexity of a postings intersection operation.
Select one:
True
False

The correct answer is 'False'.

Question 53
Question text
One disadvantage, as outlined in our text, of using a permuterm index for wild card queries is:
Select one:
a. It requires complex code that is difficult to maintain
b. It has the risk of key collisions which are difficult to resolve
c. The required rotations creates a very large dictionary
d. It cannot be used to find terms that are not spelled correctly
The correct answer is: The required rotations creates a very large dictionary

Question 54
Question text
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:
Select one:
a. document frequency
b. tf-idf weight
c. collection frequency
d. inverse document frequency
The correct answer is: inverse document frequency

Question 55
Question text
In the bag of words model, the exact ordering of terms within the document is both significant and relevant to processing.
Select one:
True
False
The correct answer is 'True'.

Question 56
Question text
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:
Select one:
a. Lemmatization
b. Case Folding
c. True casing
d. Stemming
The correct answer is: Stemming

Question 57
Question text
A scheme where a weight is assigned to a term based upon the number of occurrences of the term within a document is called:
Select one:
a. Bag of Words
b. Document Frequency
c. Term Frequency

d. Optimal weight
The correct answer is: Term Frequency

Question 58
Question text
Which of the following is NOT one of the types of queries in a complete search system discussed in our text?
Select one:
a. Wildcard Query
b. Boolean retrieval
c. Phrase Query
d. Ranked retrieval Query
The correct answer is: Ranked retrieval Query

Question 59
Question text
A measure of similarity between two vectors which is determined by measuring the angle between them is called:
Select one:
a. cosine similarity
b. sin similarity
c. vector similarity
d. vector scoring
The correct answer is: cosine similarity
For a small collection of documents on a personal computer that don't experience any change the most appropriate indexing algorithm would be:
Select one:
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
The correct answer is: Block sort-based indexing algorithm

Question 2
Question text
Considering only documents containing terms whose idf exceeds a preset threshold is an index elimination.
Select one:
True
False
The correct answer is 'True'.

Question 3
Question text
The size of the document collection that can be indexed by single-pass in-memory indexing algorithm is limited by the size of the disk storage the computer running the indexer process has access to.
Select one:
True
False
The correct answer is 'False'.

Question 4
Question text
An index that includes sequences of words or terms of variable length that have been extracted from a source document is called a:
Select one:
a. Phrase Index
b. Biword index
c. Positional index
d. Inverted Index
The correct answer is: Phrase Index

Question 5
Question text
Recall is the fraction of non relevant documents that are retrieved.
Select one:
True
False
The correct answer is 'False'.

Question 6
Question text
In order to access a particular web site in the internet, the URL must be converted into an IP address. Which service does this conversion?
Select one:
a. HTTP
b. TNS
c. DNS
d. DHCP
The correct answer is: DNS

Question 7
Question text
A program that captures and indexes content from web pages is known as what insect:
Select one:
a. Fly
b. Centipede
c. Mosquito
d. Spider
The correct answer is: Spider

Question 8
Question text
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:
Select one:
a. Stop Words
b. Tokens
c. Lemmatized Words
d. Stemmed Terms
The correct answer is: Stop Words

Question 9
Question text
A scoring function that computes an aggregate of a document's relevance from multiple sources is called evidence accumulation.
Select one:
True
False
The correct answer is 'True'.

Question 10
Question text
Weighted zone scoring is sometimes referred to as ranked Boolean retrieval.
Select one:
True
False
The correct answer is 'True'.

Question 11
Question text
The purpose of the inverse document frequency is to increase the weight of terms with high collection frequency.
Select one:
True
False
The correct answer is 'False'.

Question 12
Question text
The model of information retrieval in which we can pose any query in the form of a Boolean expression is called the ranked retrieval model.
Select one:
True
False
The correct answer is 'False'.

Question 13
Question text
The number of times that a word or term occurs in a document is called the:
Select one:
a. Proximity Operator
b. Vocabulary Lexicon
c. Term Frequency
d. Indexing Granularity
The correct answer is: Term Frequency

Question 14
Question text
An approach to compression that takes advantage of the redundancy in the dictionary that results from common prefixes that come from sorted terms is called:
Select one:

a. Front Coding
b. Blocked storage
c. Prefix Coding
d. Variable byte encoding
The correct answer is: Front Coding

Question 15
Question text
In the context of web search engines the manipulation
Select one:
a. Paid inclusion
b. SPAM
c. SEO
d. Link Analysis
The correct answer is: SPAM

Question 16
Question text
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:
Select one:
a. Lemmatization
b. Case Folding
c. True casing
d. Stemming
The correct answer is: Stemming

Question 17
Question text
The F-measure is a single measure that balances precision versus recall.
Select one:
True
False
The correct answer is 'True'.

Question 18
Question text
Stemming increases the size of the vocabulary.
Select one:
True
False
The correct answer is 'False'.

Question 19
Question text
For a moderately large collection of static documents maintained on a single system the most appropriate indexing algorithm would be:
Select one:
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
The correct answer is: Single-pass in memory indexing algorithm

Question 20
Question text
Results from a search engine that are based upon the retrieval of items using a method of term weighting such as cosine similarity is a form of:
Select one:
a. Sponsored Search
b. Algorithmic Search
c. Informational Search
d. Navigational Search
The correct answer is: Algorithmic Search

Question 21
Question text
Given two strings s1 and s2, the edit distance between them is sometimes known as the:
Select one:
a. Levenshtein distance
b. isolated-term distance
c. k-gram overlap
d. Jaccard Coefficient
The correct answer is: Levenshtein distance

Question 22
Question text
A web link within a web page that references another part of the same page is called a:
Select one:
a. Out link
b. Vector
c. In link
d. Tendril
The correct answer is: In link

Question 23
Question text
In the bag of words model, the exact ordering of terms within the document is both significant and relevant to processing.
Select one:
True
False
The correct answer is 'True'.

Question 24
Question text
A data structure that maps terms back to the parts of a document in which they occur is called an (select the best answer):
Select one:
a. Postings list
b. Incidence Matrix
c. Dictionary

d. Inverted Index
The correct answer is: Inverted Index

Question 25
Question text
A disadvantage of compression is that it reduces the transfer of data from disk to memory.
Select one:
True
False
The correct answer is 'False'.

Question 26
Question text
A process to efficiently intersect lists to be able to quickly find documents that contain both terms is referred to as merging postings lists.
Select one:
True
False
The correct answer pre-computes for each term in the dictionary, the set of documents with the highest weights for the term is:
Select one:
a. Champion list
b. Impact ordering
c. Cluster pruning
d. Tiered indexes
The correct answer is: Champion list

Question 28
Question text
Accuracy is typically the most accurate measure of IR system effectiveness.
Select one:
True
False
The correct answer is 'False'.

Question 29
Question text
The list of web pages that a web crawler has queued up to index is ntier
The correct answer is: URL Frontier

Question 30
Question text
Which of the following is a technique for context sensitive spelling correction:
Select one:
a. the Jaccard Coefficient
b. Soundex algorithms
c. k-gram indexes
d. Levenshtein distance
The correct answer is: Soundex algorithms

Question 31

Question text
One disadvantage, as outlined in our text, of using a permuterm index for wild card queries is:
Select one:
a. It requires complex code that is difficult to maintain
b. It is very large dictionary
c. The required rotations creates a very large dictionary
d. It cannot be used to find terms that are not spelled correctly
The correct answer is: The required rotations creates a very large dictionary

Question 32
Question text
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:
Select one:
a. document frequency
b. tf-idf weight
c. collection frequency
d. inverse document frequency
The correct answer is: inverse document frequency

Question 33
Question text
For a very large collection of books of classic literature the most appropriate indexing algorithm would be:
Select one:
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
The correct answer is: Distributed Map-Reduce indexing algorithm

Question 34
Question text
The tf-idf weight is highest when a term t occurs many times within a small number of documents.
Select one:
True
False
The correct answer is 'True'.

Question 35
Question text
A model of information retrieval in which we can pose any query in which search terms are combined with the operators AND, OR, and NOT:
Select one:
a. Ad Hoc Retrieval
b. Ranked Retrieval Model
c. Boolean Information Model
d. Proximity Query Model
The correct answer is: Boolean Information Model

Question 36

Question text
A compression algorithm that results in some loss of data is called:
Select one:
a. zipf compression
b. dictionary compression
c. lossless compression
d. lossy compression
The correct answer is: lossy compression

Question 37
Question text
An advantage of a positional index is that it reduces the asymptotic complexity of a postings intersection operation.
Select one:
True
False
The correct answer is 'False'.

Question 38
Question text
Precision is the fraction of retrieved documents that are relevant.
Select one:
True
False
The correct answer is 'True'.

Question 39
Question text
The standard approach to information retrieval system evaluation involves around the notion of:
Select one:
a. Quantity of documents in the collection
b. Relevant and non relevant documents.
c. Accuracy
d. user happiness
The correct answer is: Relevant and non relevant documents.

Question 40
Question text
The formula used to estimate the vocabulary size of a collection is known as:
Select one:
a. Zipf's law
b. Power law
c. Heap's law
d. Compression ratio
The correct answer is: Heap's law

Question 41
Question text
The basic operation of a web browser is to pass a request to the web server. This request is an address for a web page and is known as the:
Select one:

a. UAL: Universal Address Locator
b. HTML: Hypertext Markup Language
c. URL: Universal Resource Locator
d. HTTP: Hypertext transfer protocol
The correct answer is: URL: Universal Resource Locator

Question 42
Question text
The tf-idf weight is lower when a term t occurs many times in a document or occurs in relatively few documents.
Select one:
True
False
The correct answer is 'False'.

Question 43
Question text
Which of the following is NOT a benefit of index compression?
Select one:
a. Simplified algorithm design
b. Reduction of disk space
c. Faster transfer of data from disk to memory
d. Increased Use of caching
The correct answer is: Simplified algorithm design

Question 44
Question text
An approach to computing scores in an IR system that selects a sample of documents randomly from the collection as leaders which are in the index and links similar documents to it (followers) is called:
Select one:
a. Champion list
b. Impact ordering
c. Cluster pruning
d. Tiered indexes
The correct answer is: Cluster pruning

Question 45
Question text
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:
Select one:
a. internet collection
b. reference book
c. separate IR system.
d. standard test collection
The correct answer is: standard test collection

Question 46
Question text
The 30 most common words account for 30% of the tokens in written text is known as front

coding.
Select one:
True
False
The correct answer is 'False'.

Question 47
Question text
The purpose of the inverse document frequency is to increase the weight of terms with high collection frequency.
Select one:
True
False
The correct answer is 'False'.

Question 48
Question text
The number of documents within a collection that contain a particular term is the collection frequency of the term.
Select one:
True
False
The correct answer is 'False'.

Question 49
Question text
A scheme where a weight is assigned to a term based upon the number of occurrences of the term within a document is called:
Select one:
a. Bag of Words
b. Document Frequency
c. Term Frequency
d. Optimal weight
The correct answer is: Term Frequency

Question 50
Question text
Hashing is a process where an item is reduced, through a mathematical process, to an integer.
Select one:
True
False
The correct answer is 'True'.

Question 51
Question text
An index that is often supplemental to the inverted index and contains terms from only a particular field or section of a document is called a parametric index.
Select one:
True
False
The correct answer is 'True'.

Question 52
Question text
An approach to retrieval in a search that is likely (but not precisely) to produce the top K scoring documents is called:
Select one:
a. Exact top K document retrieval
b. top scoring document retrieval
c. Inexact top K document retrieval
d. Imprecise top K document retrieval
The correct answer is: Inexact top K document retrieval

Question 53
Question text
An approach to computing scores in an IR system that orders documents in the posting list of a term by decreasing order of term frequency is called:
Select one:
a. Champion list
b. Impact ordering
c. Cluster pruning
d. Tiered indexes
The correct answer is: Impact ordering

Question 54
Question text
For a large collection of documents such as the internet that experience frequent change the most appropriate indexing algorithm would be:
Select one:
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
The correct answer is: Dynamic indexing process employing an auxiliary index

Question 55
Question text
Which of the following is NOT one of the types of queries in a complete search system discussed in our text?
Select one:
a. Wildcard Query
b. Boolean retrieval
c. Phrase Query
d. Ranked retrieval Query
The correct answer is: Ranked retrieval Query

Question 56
Question text
A web server communicates with a client (browser) using which protocol:
Select one:
a. HTML
b. HTTP

c. FTP
d. Telnet
The correct answer is: HTTP

Question 57
Question text
A measure of similarity between two vectors which is determined by measuring the angle between them is called:
Select one:
a. cosine similarity
b. sin similarity
c. vector similarity
d. vector scoring
The correct answer is: cosine similarity

Question 58
Question text
A web page whose content doesn't vary from one request to another is called a:
Select one:
a. Text Page
b. Dynamic Page
c. Active Server Page
d. Static Page
The correct answer is: Static Page

Question 59
Question text
Which of the following items is not a component of a complete search system?
Select one:
a. Document cache
b. Indexers
c. Spell correction
d. Horizontal index
The correct answer is: Horizontal index
A web page whose content doesn't vary from one request to another is called a:
Select one:
a. Text Page
b. Dynamic Page
c. Active Server Page
d. Static Page
The correct answer is: Static Page

Question 2
Question text
The size of the document collection that can be indexed by single-pass in-memory indexing algorithm is limited by the size of the disk storage the computer running the indexer process has access to.
Select one:
True
False
The correct answer is 'False'.

Question 3
Question text
An approach to computing scores in an IR system that pre-computes for each term in the dictionary, the set of documents with the highest weights for the term is:
Select one:
a. Champion list
b. Impact ordering
c. Cluster pruning
d. Tiered indexes
The correct answer is: Champion list

Question 4
Question text
The basic operation of a web browser is to pass a request to the web server. This request is an address for a web page and is known as the:
Select one:
a. UAL: Universal Address Locator
b. HTML: Hypertext Markup Language
c. URL: Universal Resource Locator
d. HTTP: Hypertext transfer protocol
The correct answer is: URL: Universal Resource Locator

Question 5
Question text
A program that captures and indexes content from web pages is known as what insect:
Select one:
a. Fly
b. Centipede
c. Mosquito
d. Spider
The correct answer is: Spider

Question 6
Question text
For a moderately large collection of static documents maintained on a single system the most appropriate indexing algorithm would be:
Select one:
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
The correct answer is: Single-pass in memory indexing algorithm

Question 7
Question text
A measure of similarity between two vectors which is determined by measuring the angle between them is called:
Select one:
a. cosine similarity
b. sin similarity

c. vector similarity
d. vector scoring
The correct answer is: cosine similarity

Question 8
Question text
An index that is often supplemental to the inverted index and contains terms from only a particular field or section of a document is called a parametric index.
Select one:
True
False
The correct answer is 'True'.

Question 9
Question text
Which of the following is NOT one of the types of queries in a complete search system discussed in our text?
Select one:
a. Wildcard Query
b. Boolean retrieval
c. Phrase Query
d. Ranked retrieval Query
The correct answer is: Ranked retrieval Query

Question 10
Question text
In the context of web search engines the manipulation of web page content for the purpose of appearing high up in search results for selected query terms is called:
Select one:
a. Paid inclusion
b. SPAM
c. SEO
d. Link Analysis
The correct answer is: SPAM

Question 11
Question text
Hashing is a process where an item is reduced, through a mathematical process, to an integer.
Select one:
True
False
The correct answer is 'True'.

Question 12
Question text
Considering only documents containing terms whose idf exceeds a preset threshold is an index elimination.
Select one:
True
False
The correct answer is 'True'.

Question 13
Question text
The tf-idf weight is lower when a term t occurs many times in a document or occurs in relatively few documents.
Select one:
True
False
The correct answer is 'False'.

Question 14
Question text
A process to efficiently intersect lists to be able to quickly find documents that contain both terms is referred to as merging postings lists.
Select one:
True
False
The correct answer is 'True'.

Question 15
Question text
Accuracy is typically the most accurate measure of IR system effectiveness.
Select one:
True
False
The correct answer is 'False'.

Question 16
Question text
Weighted zone scoring is sometimes referred to as ranked Boolean retrieval.
Select one:
True
False
The correct answer is 'True'.

Question 17
Question text
Precision is the fraction of retrieved documents that are relevant.
Select one:
True
False
The correct answer is 'True'.

Question 18
Question text
The number of times that a word or term occurs in a document is called the:
Select one:
a. Proximity Operator
b. Vocabulary Lexicon
c. Term Frequency
d. Indexing Granularity

The correct answer is: Term Frequency

Question 19
Question text
An approach to computing scores in an IR system that selects a sample of documents randomly from the collection as leaders which are in the index and links similar documents to it (followers) is called:
Select one:
a. Champion list
b. Impact ordering
c. Cluster pruning
d. Tiered indexes
The correct answer is: Cluster pruning

Question 20
Question text
An approach to compression that takes advantage of the redundancy in the dictionary that results from common prefixes that come from sorted terms is called:
Select one:
a. Front Coding
b. Blocked storage
c. Prefix Coding
d. Variable byte encoding
The correct answer is: Front Coding

Question 21
Question text
For a large collection of documents such as the internet that experience frequent change the most appropriate indexing algorithm would be:
Select one:
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
The correct answer is: Dynamic indexing process employing an auxiliary index

Question 22
Question text
In order to access a particular web site in the internet, the URL must be converted into an IP address. Which service does this conversion?
Select one:
a. HTTP
b. TNS
c. DNS
d. DHCP
The correct answer is: DNS

Question 23
Question text
The 30 most common words account for 30% of the tokens in written text is known as front coding.

Select one:
True
False
The correct answer is 'False'.

Question 24
Question text
In the bag of words model, the exact ordering of terms within the document is both significant and relevant to processing.
Select one:
True
False
The correct answer is 'True'.

Question 25
Question text
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:
Select one:
a. internet collection
b. reference book
c. separate IR system.
d. standard test collection
The correct answer is: standard test collection

Question 26
Question text
Which of the following items is not a component of a complete search system?
Select one:
a. Document cache
b. Indexers
c. Spell correction
d. Horizontal index
The correct answer is: Horizontal index

Question 27
Question text
An approach to computing scores in an IR system that orders documents in the posting list of a term by decreasing order of term frequency is called:
Select one:
a. Champion list
b. Impact ordering
c. Cluster pruning
d. Tiered indexes
The correct answer is: Impact ordering

Question 28
Question text
Results from a search engine that are based upon the retrieval of items using a method of term weighting such as cosine similarity is a form of:
Select one:

a. Sponsored Search
b. Algorithmic Search
c. Informational Search
d. Navigational Search
The correct answer is: Algorithmic Search

Question 29
Question text
For a small collection of documents on a personal computer that don't experience any change the most appropriate indexing algorithm would be:
Select one:
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
The correct answer is: Block sort-based indexing algorithm

Question 30
Question text
An index that includes sequences of words or terms of variable length that have been extracted from a source document is called a:
Select one:
a. Phrase Index
b. Biword index
c. Positional index
d. Inverted Index
The correct answer is: Phrase Index

Question 31
Question text
The purpose of the inverse document frequency is to increase the weight of terms with high collection frequency.
Select one:
True
False
The correct answer is 'False'.

Question 32
Question text
A scoring function that computes an aggregate of a document's relevance from multiple sources is called evidence accumulation.
Select one:
True
False
The correct answer is 'True'.

Question 33
Question text
A compression algorithm that results in some loss of data is called:
Select one:
a. zipf compression