1. What is the main purpose of a PAT Tree?

a) To sort numbers efficiently
b) To perform fast text searching with preprocessing
c) To compress data files
d) To calculate probabilities
Answer: b

2. What does each position in the text correspond to in PAT Trees?

a) A character
b) A semi-infinite string (sistring)
c) A node
d) A word
Answer: b

3. What is the key operation performed on sistrings?

a) Concatenation
b) Lexicographical comparison
c) Arithmetic computation
d) Hashing
Answer: b

4. How many external nodes exist in a PAT tree for a text of size n?

a) n − 1
b) n
c) 2n
d) n + 1
Answer: b

5. PAT in PAT Tree stands for:

a) Pattern Analysis Tree
b) Patricia Tree
c) Pattern Attribute Tree
d) Patternized Array Tree
Answer: b

6. Which search in PAT tree finds all sistrings starting with a given prefix?

a) Range searching
b) Prefix searching
c) Regular expression searching
d) Proximity searching
Answer: b

7. Proximity searching finds:

a) Strings lexicographically close
b) Occurrences of two strings within a certain distance
c) Patterns in numeric data
d) The longest string
Answer: b

8. Range searching in PAT tree retrieves strings:

a) Within a fixed length
b) Between two lexicographic limits
c) Based on position
d) Randomly
Answer: b

9. Longest repetition in a text corresponds to:

a) Shortest internal node

b) Tallest internal node
c) Smallest external node
d) Root node
Answer: b

10. Most frequent string searching finds:
a) Shortest substring
b) Most repeated substring
c) Lexicographically smallest substring
d) Pattern with highest ASCII value
Answer: b

11. Regular expression searching in PAT trees is based on:
a) NFA conversion
b) DFA conversion
c) Hash function
d) Trie traversal
Answer: b

12. In PAT trees, what is bucketing?
a) Grouping small subtrees into one unit
b) Deleting external nodes
c) Compressing internal nodes
d) Replacing sistrings
Answer: a

13. Super-nodes are used to:
a) Increase node redundancy
b) Reduce disk access time
c) Add extra layers
d) Simplify algorithms
Answer: b

14. Average disk access reduction using super-nodes is:
a) 1/2 of height
b) 1/5 of height
c) 1/10 of height
d) Equal to height
Answer: c

15. PAT array is an implementation where:
a) Data stored sequentially in buckets
b) Tree replaced by sorted sistring array
c) Hash table used
d) Random access avoided
Answer: b

16. Time complexity for searching PAT trees as arrays:
a) $O(n^2)$
b) $O(\log_2 n)$
c) O(1)
d) O(n log n)
Answer: b

17. Where was PAT tree research primarily developed?
a) MIT
b) University of Waterloo
c) Stanford

d) Harvard
Answer: b

18. Merging large PAT arrays can be optimized using:
a) Random access
b) Sequential I/O
c) Linked lists
d) Binary trees
Answer: b

19. The Delayed Reading Paradigm helps reduce:
a) CPU usage
b) Random disk access cost
c) File size
d) Algorithm steps
Answer: b

20. Cost of searching a bucket with size b using binary search is:
a) b
b) 2 log b – 1
c) b^2
d) log b
Answer: b

21. What is the main goal of stemming in Information Retrieval?
a) To find the frequency of words
b) To reduce words to their root or base form
c) To translate text
d) To tokenize sentences
Answer: b

22. Conflation in IR refers to:
a) Mapping multiple morphological variants to a single form
b) Indexing all words in a text
c) Combining documents
d) Tokenizing text into words
Answer: a

23. Manual conflation can be done using:
a) Regular expressions
b) Neural networks
c) Stemming algorithms
d) Stopword removal
Answer: a

24. Automatic conflation is achieved using:
a) Human editing
b) Stemmers
c) Translators
d) Hashing
Answer: b

25. Over-stemming occurs when:
a) Not enough letters are removed
b) Too many letters are removed
c) Words are duplicated
d) No letters are removed
Answer: b

26. Under-stemming occurs when:
a) Words are conflated excessively
b) Too few letters are removed
c) The stemmer deletes entire words
d) Words are misspelled
Answer: b

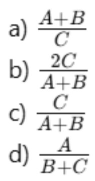
27. Which algorithm uses successor variety to find stems?
a) Lovins
b) Porter
c) Hafer & Weiss
d) Successor Variety Algorithm
Answer: d

28. In successor variety, the cutoff method segments words based on:
a) Threshold value
b) Word length
c) Prefix-suffix pair
d) Alphabetic order
Answer: a

29. Which method in successor variety looks for peaks and plateaus in successor count?
a) Entropy method
b) Complete word method
c) Peak and Plateau method
d) Cutoff method
Answer: c

30. N-gram stemmers break text into:
a) Prefixes and suffixes
b) Fixed-length character sequences
c) Entire words
d) Tokens and symbols
Answer: b

Dice's coefficient in N-gram similarity is defined as:



a) $\frac{A+B}{C}$
b) $\frac{2C}{A+B}$
c) $\frac{C}{A+B}$
d) $\frac{A}{B+C}$

Answer: b

31. The Porter stemmer is based on:
a) Neural modeling
b) Iterative longest match rules
c) Dictionary lookup
d) Context-free grammar
Answer: b

32. In the Porter algorithm, measure 'm' represents:
a) Number of vowels
b) Number of vowel-consonant sequences
c) Number of letters removed
d) Number of words in text

Answer: b

33. A stem ending with a double consonant is represented by:

a) v
b) *d
c) *o
d) x

Answer: b

34. Porter algorithm transforms words using:

a) DFA rules
b) Rewrite rules (old_suffix → new_suffix)
c) Regular expressions only
d) Random replacement

Answer: b

35. Which of the following is not a step in the Porter algorithm?

a) Step 1a – Plural removal
b) Step 2 – Suffix replacement
c) Step 3 – Prefix removal
d) Step 5a – Final e removal

Answer: c

36. Thesauri in IR systems are used to:

a) Store documents
b) Provide controlled vocabulary for indexing and retrieval
c) Compress data
d) Rank documents

Answer: b

37. Equivalence relationships in a thesaurus represent:

a) Synonyms or quasi-synonyms
b) Parent-child hierarchy
c) Cause-effect relations
d) Co-occurring words only

Answer: a

38. Pre-coordination in thesauri means:

a) Phrases are constructed during retrieval
b) Phrases are predefined and stored in thesaurus
c) Words are merged automatically
d) Queries are decomposed

Answer: b

39. Discrimination Value (DV) measures:

a) Frequency of a term
b) A term's ability to distinguish between documents
c) Length of a document
d) Rank of a query

Answer: b

40. What is the main goal of string searching algorithms?

a) Compressing data
b) Finding occurrences of a pattern within a text
c) Sorting strings alphabetically
d) Translating text

Answer: b

41. If a text has length n and a pattern has length m, the number of possible alignments is:

a) n + m
b) n – m + 1
c) n × m
d) n/m
Answer: b

42. The naive algorithm works by:
a) Matching the pattern only once
b) Checking every possible position in the text
c) Using hashing
d) Skipping unmatched positions
Answer: b

43. The expected number of comparisons in the naive algorithm depends on:
a) Alphabet size
b) Word frequency
c) Sentence length
d) Random seed
Answer: a

44. The Knuth–Morris–Pratt (KMP) algorithm avoids rechecking characters by using:
a) A hash table
b) An LPS array (Longest Prefix Suffix)
c) A stack
d) A queue
Answer: b

45. In KMP, the LPS array represents:
a) Longest proper prefix which is also a suffix
b) Longest palindrome substring
c) Last position shift
d) Left prefix sum
Answer: a

46. The time complexity of the KMP algorithm is:
a) $O(n^2)$
b) O(n + m)
c) O(log n)
d) O(1)
Answer: b

47. The Boyer–Moore algorithm starts matching from:
a) The first character of the pattern
b) The middle of the pattern
c) The last character of the pattern
d) A random position
Answer: c

48. Which two heuristics are used in Boyer–Moore?
a) Prefix and suffix heuristics
b) Bad character and good suffix heuristics
c) Rolling hash and prefix sum
d) Forward and backward shifts
Answer: b

49. In the bad character heuristic, if the mismatched character is not in the pattern:
a) Shift pattern one step
b) Shift pattern past the mismatched character

c) Restart search
d) Reduce window size
Answer: b

50. In the good suffix heuristic, if substring t has another occurrence in P:
a) Pattern is shifted to align that occurrence with t in text
b) Pattern moves backward
c) Prefix is compared
d) Matching stops
Answer: a

51. If no prefix or substring matches in good suffix rule:
a) Pattern is shifted by one
b) Pattern is shifted past t
c) Pattern restarts
d) Pattern is reversed
Answer: b

52. The Shift-Or algorithm is also known as:
a) Rabin–Karp algorithm
b) Bitap or Shift-And algorithm
c) Naive algorithm
d) Boyer–Moore algorithm
Answer: b

53. Shift-Or algorithm represents patterns using:
a) Arrays
b) Bitmasks
c) Hash tables
d) Trees
Answer: b

54. In Shift-Or, a match is detected when:
a) The most significant bit becomes 0
b) All bits are 1
c) The bitmask equals pattern
d) The least significant bit is 1
Answer: a

55. The Karp–Rabin algorithm uses what concept for pattern matching?
a) Prefix-suffix
b) Rolling hash
c) Trie tree
d) Binary search
Answer: b

56. The modulo operation with a prime number in Karp–Rabin helps to:
a) Ensure smaller text
b) Avoid overflow and reduce collisions
c) Increase text length
d) Sort patterns
Answer: b

57. If two hash values in Rabin–Karp are equal, then:
a) They always match
b) A character-by-character check is required
c) They are discarded
d) It means collision

Answer: b

58. The time complexity of Karp–Rabin in average case is:

a) O(n + m)
b) O(nm)
c) O(n log m)
d) O(1)

Answer: a

59. In string searching, 'pattern' refers to:

a) The full text
b) The substring we are looking for
c) The character set
d) The alphabet

Answer: b

60. The process of mapping morphological variants of words to a common base form is called:

a) Indexing
b) Conflation
c) Clustering
d) Tokenization

Answer: b

61. Which of the following is not a type of stemming algorithm?

a) Affix removal
b) Successor variety
c) Table lookup
d) Syntax analysis

Answer: d

62. Which of the following evaluates how well stemmers perform?

a) Correct usage and retrieval effectiveness
b) File compression speed
c) File size
d) Hash collision rate

Answer: a

63. The main drawback of the table lookup approach to stemming is:

a) High speed
b) High storage overhead
c) Accuracy
d) Lack of stemming rules

Answer: b

64. The successor variety of a string refers to:

a) The number of words following it in a text
b) The number of different characters that follow it
c) The number of vowels in it
d) Its frequency of occurrence

Answer: b

65. When the successor variety reaches a minimum and then increases sharply, it indicates:

a) Start of a new sentence
b) End of a segment boundary
c) Spelling error
d) Word repetition

Answer: b

66. The Complete Word method in successor variety identifies a segment break when:
a) The segment is a complete word in the corpus
b) The character frequency is minimum
c) A vowel appears
d) The prefix matches a suffix
Answer: a

67. The Entropy method in successor variety uses:
a) Statistical probability distribution
b) Grammar rules
c) Syntax tree
d) Word embeddings
Answer: a

68. In N-gram models, 'n' represents:
a) Number of documents
b) Number of characters in each substring
c) Number of sentences
d) Number of words in a paragraph
Answer: b

69. The advantage of N-gram stemmers is that they:
a) Capture semantic meaning
b) Ignore semantics but work language-independently
c) Require human rules
d) Depend on dictionary size
Answer: b

70. Dice's coefficient measures in n-grams:
a) Word frequency
b) Similarity between terms based on shared n-grams
c) Document ranking
d) Query performance
Answer: b

71. In the formula for Dice's coefficient in n-grams, C represents:
a) Common unique n-grams between two terms
b) Total characters in the term
c) Term frequency in corpus
d) Cluster index
Answer: a

72. Lovins stemmer and Porter stemmer are both:
a) Statistical stemmers
b) Iterative longest-match stemmers
c) Neural stemmers
d) Linguistic parsers
Answer: b

73. The Porter stemmer was proposed in:
a) 1968
b) 1980
c) 1974
d) 1985
Answer: b

74. The main benefit of stemming in IR systems is:

a) Increasing recall by matching word variants
b) Reducing disk memory
c) Improving grammar
d) Making words longer
Answer: a

75. Which of the following is a major feature of a thesaurus in IR?
a) Controlled vocabulary
b) Large number of stopwords
c) Random word collection
d) Unrelated word mapping
Answer: a

76. The three types of term relationships in a thesaurus are:
a) Hierarchical, equivalence, and non-hierarchical
b) Semantic, syntactic, and pragmatic
c) Prefix, suffix, and infix
d) Symbolic, numeric, and linguistic
Answer: a

77. In automatic thesaurus construction, the Poisson model helps to:
a) Identify trivial and non-trivial words statistically
b) Sort terms alphabetically
c) Measure synonym frequency
d) Evaluate compression ratio
Answer: a