NLP Revision

Choose the correct answer:

1. What does the "Phonetical and Phonological" level focus on?

- A) Understanding sentence structure

- B) Understanding sound patterns and speeches

- C) Understanding the literal meaning of words

- D) Understanding real-world knowledge

- Answer: B) Understanding sound patterns and speeches

2. Which level involves understanding the structure of words and systematic relations?

- A) Syntactic

- B) Semantic

- C) Morphological

- D) Discourse

- Answer: C) Morphological

3. The "Lexical" level focuses on:

- A) Understanding part of speech

- B) Understanding sentence structures

- C) Understanding sound patterns

- D) Understanding real-world context

- Answer: A) Understanding part of speech

4. What is the main focus of the "Syntactic" level?

- A) Understanding word structure

- B) Understanding sentence structure

- C) Understanding larger text units

- D) Understanding literal meanings

- Answer: B) Understanding sentence structure

5. Which level addresses the literal meaning of words, phrases, and sentences?

- A) Semantic

- B) Pragmatic

- C) Lexical

- D) Morphological

- Answer: A) Semantic

6. The "Discourse" level is concerned with:

- A) Single-word meanings

- B) Units larger than a single sentence

- C) Sentence structure

- D) Sound patterns

- Answer: B) Units larger than a single sentence

7. What does the "Pragmatic" level focus on?

- A) Understanding systematic word relations

- B) Real-world knowledge and broader sentence context

- C) Part of speech identification

- D) Sentence structure

- Answer: B) Real-world knowledge and broader sentence context

8. Understanding part of speech is the main focus of which level?

- A) Syntactic

- B) Lexical

- C) Phonetical

- D) Semantic

- Answer: B) Lexical

9. Which level helps in understanding the patterns present in sound and speeches?

- A) Pragmatic

- B) Phonetical and Phonological

- C) Discourse

- D) Syntactic

- Answer: B) Phonetical and Phonological

10. Real-world context and bigger sentence meaning is analyzed at the ____ level.

- A) Discourse

- B) Morphological

- C) Pragmatic

- D) Semantic

- Answer: C) Pragmatic

11. What type of ambiguity is present in the words "write" and "right"?

- A) Syntactic ambiguity

- B) Phonological ambiguity

- C) Semantic ambiguity

- D) Pragmatic ambiguity

- Answer: B) Phonological ambiguity

12. The word "bank" being interpreted as either a financial institution or the side of a river is an example of:

- A) Morphological ambiguity

- B) Semantic ambiguity

- C) Syntactic ambiguity

- D) Phonological ambiguity

- Answer: B) Semantic ambiguity

13. The ambiguity of the word "play," which can be a noun or a verb, is an example of:

- A) Part-of-speech ambiguity

- B) Syntactic ambiguity

- C) Pragmatic ambiguity

- D) Morphological ambiguity

- Answer: A) Part-of-speech ambiguity

14. In the sentence "I can see a man with a telescope," what type of ambiguity is demonstrated?

- A) Semantic ambiguity

- B) Phonological ambiguity

- C) Syntactic ambiguity

- D) Morphological ambiguity

- Answer: C) Syntactic ambiguity

15. Which of the following types of ambiguity is caused by words having multiple meanings?

- A) Syntactic ambiguity

- B) Phonological ambiguity

- C) Semantic ambiguity

- D) Morphological ambiguity

- Answer: C) Semantic ambiguity

16. When a sentence can be interpreted in more than one grammatical way, it is an example of:

- A) Syntactic ambiguity

- B) Part-of-speech ambiguity

- C) Semantic ambiguity

- D) Phonological ambiguity

- Answer: A) Syntactic ambiguity

17. The word "bat" referring to either an animal or sports equipment is an example of:

- A) Phonological ambiguity

- B) Semantic ambiguity

- C) Syntactic ambiguity

- D) Pragmatic ambiguity

- Answer: B) Semantic ambiguity

18. In the sentence "He saw the boy with the binoculars," the ambiguity arises because:

- A) The sentence structure allows multiple interpretations

- B) The words have multiple meanings

- C) The pronunciation of words is unclear

- D) The part of speech is uncertain

- Answer: A) The sentence structure allows multiple interpretations

19. Which type of ambiguity occurs when the sound of two words is identical but their meanings differ?

- A) Morphological ambiguity

- B) Phonological ambiguity

- C) Syntactic ambiguity

- D) Semantic ambiguity

- Answer: B) Phonological ambiguity

20. What is the primary purpose of using regular expressions in NLP preprocessing?

- A) To calculate sentence sentiment

- B) To tokenize text into sentences

- C) To find patterns in text

- D) To translate text into another language

- Answer: C) To find patterns in text

21. Which regular expression pattern matches any sequence of digits?

- A) `\w+`

- B) `\s+`

- C) `\d+`

- D) `\D+`

- Answer: C) `\d+`

22. What is the main goal of tokenization in NLP?

- A) To remove punctuation from text

- B) To determine the sentiment of text

- C) To extract named entities from text

- D) To split text into smaller units like words or sentences

- Answer: D) To split text into smaller units like words or sentences

23. Which library is commonly used for tokenization in Python?

- A) TensorFlow

- B) NLTK

- C) Matplotlib

- D) Pandas

- Answer: B) NLTK

24. Why are stop words removed during NLP preprocessing?

- A) They carry significant semantic meaning

- B) They are irrelevant and can clutter analysis

- C) They are difficult to tokenize

- D) They increase the accuracy of sentiment analysis

- Answer: B) They are irrelevant and can clutter analysis

25. Which of the following is a common stop word?

- A) Algorithm

- B) Python

- C) And

- D) Sentence

- Answer: C) And

26. What does stemming do to words?

- A) It removes prefixes only

- B) It reduces words to their root form

- C) It replaces synonyms in the text

- D) It counts word frequency

- Answer: B) It reduces words to their root form

27. Which of these is an example of stemming?

- A) "Running" becomes "Run"

- B) "Cats" becomes "Felines"

- C) "Play" becomes "Played"

- D) "Happy" becomes "Happier"

- Answer: A) "Running" becomes "Run"

28. What is the difference between stemming and lemmatization?

- A) Lemmatization uses dictionaries to find base forms, while stemming does not

- B) Stemming finds synonyms, while lemmatization simplifies sentences

- C) Stemming works on verbs only, while lemmatization works on nouns

- D) Lemmatization converts words into uppercase letters

- Answer: A) Lemmatization uses dictionaries to find base forms, while stemming does not

29. Which library provides lemmatization in Python?

- A) Matplotlib

- B) NumPy

- C) WordNet in NLTK

- D) OpenCV

- Answer: C) WordNet in NLTK

30. What does POS tagging assign to each word in a sentence?

- A) Named entities

- B) Grammatical roles such as noun, verb, adjective

- C) Synonyms

- D) Translation equivalents

- Answer: B) Grammatical roles such as noun, verb, adjective

31. What is automatic tagging in NLP?

- A) Assigning part of speech (POS) tags to words automatically

- B) Generating topic models from text

- C) Translating text automatically

- D) Matching regular expressions

- Answer: A) Assigning part of speech (POS) tags to words automatically

32. What is typically required for automatic tagging to be effective?

- A) A trained model

- B) High-frequency words

- C) A dictionary of synonyms

- D) Translation capabilities

- Answer: A) A trained model

33. What is the purpose of Named Entity Recognition (NER)?

- A) To identify specific entities such as names, dates, and locations

- B) To identify synonyms in text

- C) To assign grammatical roles to words

- D) To create embeddings for text

- Answer: A) To identify specific entities such as names, dates, and locations

34. Which of these is an example of a named entity?

- A) London

- B) 1990

- C) Microsoft

- D) All of the above

- Answer: D) All of the above

35. Which preprocessing task is essential for reducing the dimensionality of text data?

- A) Removing stop words

- B) Lemmatization

- C) Both A and B

- D) None of the above

- Answer: C) Both A and B

36. Which preprocessing task is necessary to improve the quality of downstream NLP tasks like translation or summarization?

- A) POS Tagging

- B) Tokenization

- C) Removing stop words

- D) All of the above

- Answer: D) All of the above

37. Why is stemming considered less accurate than lemmatization?

- A) It uses heuristic rules rather than linguistic rules

- B) It relies on dictionaries

- C) It always produces shorter words

- D) It focuses only on nouns

- Answer: A) It uses heuristic rules rather than linguistic rules

38. Which NLP preprocessing task helps identify dates and names within a document?

- A) Tokenization

- B) NER

- C) POS Tagging

- D) Regular Expressions

- Answer: B) NER

39. What does the `findall` function in regular expressions return?

- A) A single Match object

- B) A list containing all matches

- C) A boolean indicating if a match was found

- D) The first occurrence of a match

- Answer: B) A list containing all matches

40. What does the `search` function return if a match is found in the string?

- A) A string containing the matched pattern

- B) A Match object

- C) A list of all matches

- D) A boolean indicating a match

- Answer: B) A Match object

41. How does the `split` function handle matches in a string?

- A) It replaces all matches with a specified string

- B) It removes all matches

- C) It splits the string at each match and returns a list

- D) It counts the number of matches

- Answer: C) It splits the string at each match and returns a list

42. What is the primary use of the `sub` function in regular expressions?

- A) Finding all matches

- B) Splitting the string

- C) Replacing one or many matches with a string

- D) Extracting matched patterns

- Answer: C) Replacing one or many matches with a string

43. Which of the following functions would you use to check if a pattern exists anywhere in a string?

- A) `findall`

- B) `sub`

- C) `search`

- D) `split`

- Answer: C) `search`

44. What is an example of a unit in tokenization?

- A) Corpus

- B) Paragraphs

- C) Words or sentences

- D) Topics

- Answer: C) Words or sentences

45. Which stemming algorithm is not included in Python's NLTK library?

- A) Porter Stemmer

- B) Snowball Stemmer

- C) Lancaster Stemmer

- D) BERT Stemmer

- Answer: D) BERT Stemmer

46. What does a unigram tagger do when tagging tokens?

- A) It uses the previous and next tokens to tag the current token

- B) It uses only the current token in isolation to assign a tag

- C) It assigns a tag based on the entire document context

- D) It combines multiple taggers for better accuracy

- Answer: B) It uses only the current token in isolation to assign a tag

47. What is the primary difference between a unigram tagger and a bigram tagger?

- A) A unigram tagger uses one token, while a bigram tagger uses two tokens for context

- B) A unigram tagger uses words, while a bigram tagger uses sentences

- C) A unigram tagger is unsupervised, while a bigram tagger is supervised

- D) A unigram tagger is faster but less accurate than a bigram tagger

- Answer: A) A unigram tagger uses one token, while a bigram tagger uses two tokens for context

48. Which method combines different taggers for improved performance?

- A) Tokenization

- B) Default Tagging

- C) Backoff Tagging

- D) Brill’s Tagging

- Answer: C) Backoff Tagging

49. What is the purpose of Brill’s Tagger in NLP?

- A) Assign tags using unsupervised learning

- B) Combine results from multiple taggers

- C) Use transformational rules to correct tagging mistakes

- D) Use statistical models for tagging

- Answer: C) Use transformational rules to correct tagging mistakes

50. Which type of tagger assigns tags based on matching patterns in regular expressions?

- A) Bigram Tagger

- B) Default Tagger

- C) Regular Expression Tagger

- D) N-gram Tagger

- Answer: C) Regular Expression Tagger

51. What is the key limitation of training a tagger on the same data used for testing?

- A) It reduces the accuracy of the model

- B) It leads to overfitting and poor generalization

- C) It increases the time required for tagging

- D) It prevents tagging of unseen words

- Answer: B) It leads to overfitting and poor generalization

52. What is the purpose of One-Hot Encoding in NLP?

- A) Reducing the dimensionality of text data

- B) Representing categorical data as binary vectors

- C) Grouping similar words together

- D) Tokenizing text into words

- Answer: B) Representing categorical data as binary vectors

53. In One-Hot Encoding, how is the dimensionality determined?

- A) Based on the number of sentences

- B) Based on the number of unique tokens (vocabulary size)

- C) Based on the document length

- D) Based on the number of stop words removed

- Answer: B) Based on the number of unique tokens (vocabulary size)

54. What is a major disadvantage of One-Hot Encoding in NLP?

- A) It cannot handle numerical data

- B) It introduces a bias in text representation

- C) It creates sparse, high-dimensional vectors

- D) It fails to tokenize text properly

- Answer: C) It creates sparse, high-dimensional vectors

55. What does the Bag of Words model represent?

- A) The sequence of words in a document

- B) The frequency of each word in a document, ignoring word order

- C) Semantic relationships between words

- D) Word embeddings for each word

- Answer: B) The frequency of each word in a document, ignoring word order

56. Which of the following is a limitation of the Bag of Words model?

- A) It considers word order

- B) It is difficult to implement

- C) It fails to capture semantic meaning and context

- D) It requires a labeled dataset

- Answer: C) It fails to capture semantic meaning and context

57. How is the Bag of Words representation typically stored?

- A) As dense vectors

- B) As sparse matrices

- C) As word embeddings

- D) As CSV files

- Answer: B) As sparse matrices

58. What does the Count Vectorizer do in NLP?

- A) It tokenizes text and creates vectors of word counts

- B) It assigns probabilities to each word in a document

- C) It reduces the dimensionality of vectors

- D) It creates embeddings for words

- Answer: A) It tokenizes text and creates vectors of word counts

59. Which of the following can be specified in Count Vectorizer?

- A) Minimum and maximum word frequency thresholds

- B) Semantic relationships between words

- C) Pre-trained word embeddings

- D) Stop word removal algorithms

- Answer: A) Minimum and maximum word frequency thresholds

60. What is the main output of a Count Vectorizer?

- A) A dense embedding matrix

- B) A sparse matrix of word frequencies

- C) A semantic graph of words

- D) A probabilistic distribution of words

- Answer: B) A sparse matrix of word frequencies

61. What does TF-IDF stand for?

- A) Term Frequency – Inverse Document Frequency

- B) Text Frequency – Inverse Data Frequency

- C) Token Frequency – Indexed Document Frequency

- D) Term Factor – Indexed Data Factor

- Answer: A) Term Frequency – Inverse Document Frequency

62. What is the purpose of TF-IDF?

- A) To calculate word embeddings

- B) To weigh words based on their importance in a document relative to the corpus

- C) To remove stop words from text

- D) To capture semantic relationships between words

- Answer: B) To weigh words based on their importance in a document relative to the corpus

63. Which of the following words is likely to have a low TF-IDF score in most corpora?

- A) "the"

- B) "machine learning"

- C) "data"

- D) "algorithm"

- Answer: A) "the"

64. How does the IDF component of TF-IDF affect word weighting?

- A) Increases the weight of frequent words

- B) Decreases the weight of rare words

- C) Increases the weight of rare words

- D) Ignores word frequency altogether

- Answer: C) Increases the weight of rare words

65. What does an N-Gram represent in NLP?

- A) The meaning of a single word

- B) A sequence of N consecutive tokens in a text

- C) A mathematical formula for word embedding

- D) The frequency of a single token in a document

- Answer: B) A sequence of N consecutive tokens in a text

66. Why are N-Grams used in text processing?

- A) To improve text tokenization

- B) To capture local context and word sequences

- C) To represent words as vectors

- D) To visualize text data

- Answer: B) To capture local context and word sequences

67. What is a disadvantage of using large N-Grams (e.g., N > 3)?

- A) They reduce the dimensionality of data

- B) They require more computational resources

- C) They lose contextual information

- D) They cannot be used with sparse matrices

- Answer: B) They require more computational resources

68. What does an Occurrence Matrix represent in text analysis?

- A) The semantic relationships between words

- B) The frequency of each word in each document

- C) The embedding of each token in a vector space

- D) The similarity between two tokens

- Answer: B) The frequency of each word in each document

69. What does a Co-Occurrence Matrix measure?

- A) The frequency of individual words in a document

- B) The occurrence of pairs of words appearing together in a context window

- C) The probability distribution of tokens in a corpus

- D) The similarity between two documents

- Answer: B) The occurrence of pairs of words appearing together in a context window

70. Why are Co-Occurrence Matrices useful in NLP?

- A) They capture the global context of words

- B) They reduce dimensionality of word vectors

- C) They help in understanding word relationships and associations

- D) They tokenize text data into N-Grams

- Answer: C) They help in understanding word relationships and associations

71. Consider a document containing 200 words, where the word “machine” appears 8 times. The term frequency (TF) for “machine” is:

- A) 0.02

- B) 0.04

- C) 0.08

- D) 0.16

- Answer: B) 0.04

Explanation:

TF = (Number of times the word appears) / (Total number of words)

TF = 8 / 200 = 0.04

72- Assume there are 1 million documents, and the word “learning” appears in 10,000 of them. What is the inverse document frequency (IDF) for “learning”?

- A) log(100)

- B) log(1,000)

- C) log(10,000)

- D) log(100,000)

- Answer: A) log(100)

Explanation:

IDF = log(Total number of documents / Number of documents containing the word)

IDF = log(1,000,000 / 10,000) = log(100)

73. Given the following information:

- The word “data” appears 15 times in a document containing 500 words (TF = 15 / 500 = 0.03).

- There are 2 million documents, and “data” appears in 50,000 of them (IDF = log(2,000,000 / 50,000) = log(40) ≈ 1.6).

What is the TF-IDF score for “data”?

- A) 0.24

- B) 0.48

- C) 0.048

- D) 0.16

- Answer: C) 0.048

Explanation:

TF-IDF = TF × IDF

TF-IDF = 0.03 × 1.6 = 0.048

74. What is the main goal of Word2Vec?

- A) To generate synonyms for words
- B) To find the frequency of words in a document
- C) To represent words as dense vectors in a continuous vector space
- D) To cluster words into groups

Answer: C) To represent words as dense vectors in a continuous vector space

75. Word2Vec embeddings are primarily used to capture:

- A) Syntax only
- B) Semantic and syntactic relationships between words
- C) Frequency of word occurrences
- D) POS tags of words

Answer: B) Semantic and syntactic relationships between words

76. In the CBOW model, the objective is to:

- A) Predict the center word using its context words

- B) Predict the context words using the center word

- C) Find word frequency in a corpus

- D) Cluster similar words together

Answer: A) Predict the center word using its context words

77. The CBOW model uses:

- A) The context words to predict a missing word

- B) A single word to predict the sequence of context words

- C) Word frequencies to build embeddings

- D) A co-occurrence matrix

Answer: **A) The context words to predict a missing word**

78. What is the key difference between CBOW and Skip-gram in Word2Vec?

- A) CBOW predicts the target word from context words, while Skip-gram predicts context words from the target word

- B) CBOW uses bigram statistics, while Skip-gram uses unigram statistics

- C) Skip-gram is unsupervised, but CBOW is supervised

- D) CBOW is slower than Skip-gram

Answer: **A) CBOW predicts the target word from context words, while Skip-gram predicts context words from the target word**

79. How does Word2Vec capture analogies like "Paris - France + Italy = Rome"?

- A) By clustering similar word frequencies

- B) By using vector arithmetic on word embeddings

- C) By generating word co-occurrence matrices

- D) By using skip connections in neural networks

Answer: **B) By using vector arithmetic on word embeddings**

80. What is the purpose of negative sampling in Word2Vec?

- A) To reduce the size of the vocabulary

- B) To improve training speed by approximating the softmax function

- C) To penalize incorrect predictions

- D) To normalize word vectors

Answer: **B) To improve training speed by approximating the softmax function**

81. Negative sampling selects:

- A) Words that are semantically similar to the target word

- B) Random noise words that are unrelated to the context

- C) Rare words from the corpus

- D) Words from a fixed stop-word list

Answer: **B) Random noise words that are unrelated to the context**

82. During Word2Vec training, the embedding layer is:

- A) Fixed throughout the process

- B) Randomly initialized and updated as part of training

- C) Pretrained and not updated during training

- D) Derived from a co-occurrence matrix

Answer: **B) Randomly initialized and updated as part of training**

83. Word2Vec can be considered a simplified version of:

- A) A language modeling task
- B) A machine translation model
- C) A clustering algorithm
- D) A sentiment analysis task

Answer: **A) A language modeling task**

84. How does Word2Vec differ from traditional N-gram language models?

- A) Word2Vec uses dense embeddings instead of sparse representations
- B) Word2Vec generates probabilities for entire sentences
- C) Word2Vec requires labeled data for training
- D) Word2Vec does not use any context during training

Answer: **A) Word2Vec uses dense embeddings instead of sparse representations**

85. In language modeling, the main task is:

- A) Syntactic parsing
- B) Contextual prediction
- C) Next-word prediction
- D) Token segmentation

Answer: **C) Next-word prediction**

86. What is the dimensionality of Word2Vec embeddings typically?

- A) Fixed at 300 dimensions
- B) Depends on the model's configuration (e.g., 50, 100, 300 dimensions)

- C) Always equal to the size of the vocabulary

- D) Depends on the size of the input corpus

Answer: **B) Depends on the model's configuration (e.g., 50, 100, 300 dimensions)**

87. What is the key idea behind the word representation approach introduced by Mikolov?

- A) Count word occurrences

- B) Use Deep Learning to classify sentences

- C) Predict surrounding words for each word

- D) Use rule-based models for language processing

- Answer: **C) Predict surrounding words for each word**

88. What are the two architectures proposed for word representation?

- A) Neural Network and Deep Learning models

- B) Word2Vec and GloVe models

- C) Continuous Bag-of-Words and Continuous Skip-gram models

- D) Transformer and Attention models

- **Answer: C) Continuous Bag-of-Words and Continuous Skip-gram models**

89. Which of the following statements is true about this word representation approach?

- A) It relies on Deep Learning methods

- B) It encodes word meanings spatially in a vector space

- C) It cannot easily incorporate new words or documents

- D) It was first proposed by Google in 2017

- Answer: B) It encodes word meanings spatially in a vector space

90. What is the advantage of using low-dimensional vectors to represent words?

- A) It improves word prediction speed and allows easy addition of new words or documents.

- B) It ensures exact word counts across large datasets.

- C) It eliminates the need for context in language processing.

- D) It uses rule-based language models for better accuracy.

- Answer: A) It improves word prediction speed and allows easy addition of new words or documents.