Chapter 1 MCQs

1. What is Natural Language Processing (NLP) a subfield of?
    A. Computer Graphics
    B. Artificial Intelligence
    C. Database Management
    D. Web Development
2. What is the primary goal of NLP?
    A. To generate random text
    B. To enable machines to understand, interpret, and answer human language meaningfully
    C. To convert text into images
    D. To process numerical data only
3. NLP is described as an engineering field that aims to construct technology to perform useful functions, diverging from which field that applies computer science to learn the rules of language?
    A. Linguistics
    B. Computational Biology
    C. Computational Linguistics
    D. Cognitive Science
4. Which two overlapping subfields comprise NLP?
    A. Natural Language Recognition (NLR) and Natural Language Display (NLD)
    B. Natural Language Understanding (NLU) and Natural Language Generation (NLG)
    C. Natural Language Input (NLI) and Natural Language Output (NLO)
    D. Natural Language Processing (NLP) and Speech Recognition (SR)
5. What does Natural Language Understanding (NLU) primarily involve?
    A. Text generation by machine
    B. Semantic analysis or identification of the intended sense of text
    C. Converting sound into text
    D. Analyzing spoken words into words
6. Which of the following is an example of a conversational agent that uses NLP?
    A. Microsoft Word
    B. Adobe Photoshop
    C. Amazon's Alexa
    D. Google Chrome
7. What kind of data does NLP require systems to have?
    A. A small, unstructured dataset
    B. A corpus, which is a vast collection of linguistic data
    C. Only numerical data
    D. Pre-processed data only
8. Before any model can learn from language, what must happen to the text data?
    A. It must be stored in a raw format
    B. It must be cleaned, made simpler, and translated into a machine-readable form
    C. It must be converted into an image
    D. It must be manually labeled by humans
9. Which of the following is NOT listed as a principal step in transforming raw text data for NLP?
    A. Tokenization
    B. Text Cleaning

C. Data Encryption
D. Lemmatization

10. What is the process of splitting text into words or parts of words called?
    A. Lemmatization
    B. Stemming
    C. Tokenization
    D. Segmentation
11. What is the purpose of Stemming and Lemmatization in data preprocessing?
    A. To increase the length of words
    B. To bring words back to their base or root form
    C. To identify grammatical errors
    D. To remove stop words
12. Which of the following is an example of a stop word?
    A. "Run"
    B. "The"
    C. "University"
    D. "Learning"
13. What is the goal of Stop Word Removal?
    A. To increase the amount of noise in the data
    B. To strip away frequent words that contribute little meaning
    C. To add more common words to the text
    D. To convert words into numbers
14. Which preprocessing step splits long text into separate sentences, which can be challenging due to abbreviations?
    A. Tokenization
    B. Stop Word Removal
    C. Sentence Segmentation
    D. Lemmatization
15. What are features in the context of NLP?
    A. Raw text data
    B. Significant attributes extracted from text, converted into numbers
    C. Random numbers assigned to words
    D. Grammatical rules of a language
16. Which feature extraction technique counts the frequency of each word in a document?
    A. TF-IDF
    B. Word Embeddings
    C. Bag-of-Words (BoW)
    D. GLOVE
17. TF-IDF measures how significant a word is in a document relative to what?
    A. Its position in the sentence
    B. All other words in the same document
    C. All documents in the corpus
    D. The length of the document
18. What do Word Embeddings (like Word2Vec and GLOVE) convert words into?
    A. Images
    B. Vectors (collections of numbers)
    C. Boolean values
    D. Alphabetic sequences
19. Word2Vec learns word meaning from which type of words?
    A. Random words

B. Adjacent words
C. Words with the same starting letter
D. Words found in other documents

20. Which of the following is considered a "Classic Model" useful for easier NLP tasks?
    A. Neural Networks
    B. Hidden Markov Models (HMM)
    C. Logistic Regression
    D. BERT
21. Which type of advanced NLP model does NOT necessarily require human-crafted features?
    A. Logistic Regression
    B. Naive Bayes
    C. Neural Networks
    D. Decision Trees
22. What is the primary function of Language Models?
    A. To summarize long texts
    B. To make predictions about what follows in a sentence
    C. To perform sentiment analysis
    D. To translate languages
23. Pretrained models like BERT, GPT-3, and GPT-4 are trained on what kind of data?
    A. Small, custom datasets
    B. Enormous collections of text such as Wikipedia
    C. Only numerical tables
    D. Only user-generated content
24. NLP plays a fundamental role in comprehending enormous quantities of which type of data?
    A. Structured numerical data
    B. Unstructured text data
    C. Image data only
    D. Audio data only
25. Which of the following is NOT explicitly mentioned as a real-world application of NLP?
    A. Virtual Assistants and Chatbots
    B. Language Localization and Translation
    C. Spreadsheet Calculation
    D. Sentiment Analysis
26. What do customer service chatbots use NLP for?
    A. To randomly generate responses
    B. To interpret the intent of the user and identify the query category
    C. To ignore user queries
    D. To only provide pre-written answers
27. NLP drives real-time language translation applications, enabling what to be overcome worldwide?
    A. Time zones
    B. Cultural differences
    C. Language barriers
    D. Network latency
28. In the medical field, how can physicians use voice dictation, powered by NLP?
    A. To search for medical images
    B. To record patient information or write prescriptions

C. To analyze patient emotions
D. To manage hospital finances

29. What is Sentiment Analysis the process of applying NLP methods to identify?
    A. The author of a text
    B. The emotional tone of a given text
    C. The grammatical correctness of a text
    D. The length of a text
30. In what area is sentiment analysis also being put to good use for social good?
    A. Financial trading
    B. Mental health, to identify signs of anxiety, depression, or stress
    C. Sports analytics
    D. Weather forecasting
31. How is NLP revolutionizing healthcare?
    A. By completely replacing human doctors
    B. By allowing systems to analyze and process unstructured medical text
    C. By automating surgical procedures
    D. By manufacturing medical equipment
32. NLP-powered tools can extract what kind of data from medical text?
    A. Patient demographic data only
    B. Medical terms (diseases, medications, procedures) and summarize results
    C. Financial records only
    D. Hospital staff schedules
33. NLP is at the core of recommendation engines for platforms like YouTube and Netflix through what mechanism?
    A. Manually tagging content
    B. Analyzing user activity history and tastes
    C. Randomly suggesting content
    D. Relying on user feedback alone
34. In the banking and legal industries, NLP helps to automatically scan and summarize long documents such as:
    A. Personal emails
    B. Compliance reports, contracts, and loan agreements
    C. News articles
    D. Social media posts
35. What is one of the most fundamental challenges in NLP, referring to the vastness and variability of human languages?
    A. Computational Efficiency
    B. Data Redundancy
    C. Complexity and Diversity of Human Languages
    D. Model Interpretability
36. Why is creating high-quality and representative training data a challenge in NLP?
    A. It is inexpensive and quick to create
    B. It is time-consuming, costly, and often subject to biases
    C. It is readily available for all languages and industries
    D. It is not necessary for effective NLP models
37. What can significantly affect NLP model performance if the training data is incomplete, biased, or inaccurately labeled?
    A. The model's speed
    B. The model's accuracy and reliability
    C. The model's complexity

D. The model's interpretability

38. What is a major limitation regarding computational power for deep learning models used in NLP?
    A. They require only basic CPUs
    B. They require substantial computational power, including GPUs or TPUs
    C. They can run on any consumer device
    D. They do not require any specialized hardware
39. What is a natural characteristic of human language where the same phrase can have different meanings depending on context?
    A. Consistency
    B. Clarity
    C. Ambiguity
    D. Redundancy
40. How does human communication often involve incomplete sentences, implied meanings, and non-standard structures, which pose difficulties for NLP models?
    A. It makes NLP simpler
    B. It poses difficulties for NLP models trying to achieve true comprehension
    C. It speeds up the processing time
    D. It reduces the need for large datasets
41. What can confuse NLP systems and lead to inaccurate interpretations when present in real-world text data?
    A. Perfectly formatted text
    B. Spelling mistakes, typographical errors, slang, and improper grammar
    C. Only formal language
    D. Standardized vocabulary
42. What is the primary concern when models learn from non-representative or prejudiced data in NLP systems?
    A. Decreased computational cost
    B. Increased transparency
    C. Bias in outputs, leading to unfair or discriminatory results
    D. Faster training times
43. What is Polysemy in the context of NLP?
    A. When a single word has only one meaning
    B. When a single word has multiple meanings depending on its usage
    C. When multiple words have the same meaning
    D. When words have no meaning
44. What is a major hurdle for supporting multiple languages in NLP systems?
    A. All languages have the same structure and vocabulary
    B. Training models that perform equally well across languages with different structures, vocabularies, and scripts
    C. It reduces the need for diverse data
    D. It simplifies model development
45. What technique involves generating new data by altering existing data, such as paraphrasing a sentence or word substitution?
    A. Data Synthesis
    B. Crowdsourcing
    C. Data Augmentation
    D. Data Cleaning
46. What method involves employing artificial techniques to create data that simulates actual use of language?

A. Data Augmentation
B. Data Synthesis
C. Crowdsourcing
D. Data Validation

47. What is a method for gathering, labeling, and validating language data for wider coverage and increased quality, often by employing large groups of people?
    A. Data Synthesis
    B. Data Augmentation
    C. Crowdsourcing
    D. Manual Annotation
48. What is crucial for maintaining trust in NLP applications by helping systems express confidence levels in their outputs?
    A. Increasing uncertainty
    B. Reducing uncertainty and minimizing false positives
    C. Generating more false positives
    D. Ignoring confidence levels
49. Implementing real-time NLP pipelines, dialogue management strategies, and intent recognition models is essential to achieve what in conversations?
    A. Random interactions
    B. Disconnected turns
    C. Natural, human-like interactions over extended conversations
    D. One-way communication
50. What is an effective NLP system dependent on, beyond just high-capacity algorithms?
    A. Only basic data handling
    B. Techniques for data handling, coping with linguistic vagueness, and adjusting to variability in linguistic conditions
    C. Only highly complex algorithms
    D. Ignoring linguistic nuances