NLP Multiple Choice Questions Section 1 - 5

1. You are building a chatbot that answers customer questions using AI. Which field of AI primarily focuses on making this possible?

A. Computer Vision
 B. Natural Language Processing
 C. Reinforcement Learning
 D. Robotics

2. In one-hot encoding, how many dimensions would the vector have if the vocabulary size is 7?

A. 3
 B. 7
 C. Number of unique letters
 D. Depends on sentence length

3. Given a dataset with repeated words like "I, dogs, agree, agree", which representation would best capture the repetition?

A. One-hot encoding
 B. Bag of Words
 C. Word Embeddings
 D. ASCII encoding

4. Which representation method best captures semantic meaning?

A. One-hot encoding
 B. Bag of Words
 C. Embedding vectors
 D. ASCII encoding

5. Which representation requires training (unsupervised learning)?

A. One-hot encoding
 B. Bag of Words
 C. Word Embeddings
 D. Character-level counts

6. When visualizing pretrained embeddings using PCA for the following words: (man, woman, male, female, king, queen, prince), what are you likely to see?

A. Word frequency distribution
 B. Clusters of semantically related words
 C. Binary patterns of text
 D. Randomly distributed words

7. From the following, which is the most likely representation for the words "fast" and "quick"?

A. fast: [0.88, 0.45, 0.12, 0.91] / quick: [0.85, 0.41, 0.15, 0.89]
 B. fast: [0.88, 0.45, 0.12, 0.91] / quick: [-0.88, -0.45, -0.12, -0.91]
 C. fast: [1.0, 0.0, 0.0, 0.0] / quick: [0.0, 1.0, 0.0, 0.0]
 D. fast: [0.10, 0.95, -0.30, 0.05] / quick: [0.85, -0.15, 0.70, -0.50]

8. You want to find all occurrences of the word "cat" in a text. Which regex should you use?

A. cat
 B. Cat
 C. [cat]
 D. (cat)

9. What does the regex \d match?

A. Any lowercase letter
 B. Any digit
 C. Any punctuation
 D. A space

10. To find a sentence ending with "dog", which pattern should you use?

A. ^dog
 B. dog$
 C. .*dog
 D. \bdog\b

11. To remove all punctuation marks, which regex can be used with re.sub()?

A. r"\s"
 B. r"[^\w\s]"
 C. r"\d"
 D. r"."

12. Which library is known for its fast pipelines and pretrained models?

A. NLTK
 B. spaCy
 C. Stanza
 D. TextBlob

13. If your goal is research and experimentation with tokenizers, corpora, and linguistic algorithms, which should you use?

A. NLTK
 B. spaCy
 C. Stanza
 D. TextBlob

14. Which framework is most commonly associated with Transformer models in NLP?

A. NLTK
 B. spaCy
 C. Hugging Face Transformers
 D. TextBlob

15. Which statement best describes CUDA?

A. It's used for CPU parallelization
B. It enables GPU-based general computation
C. It is a visualization tool
D. It is a text-processing API

16. Given the dataset below, which class is the Naive Bayes model most likely to predict for the sentence "amazing story and acting"?

ID	Label	Text
1	Positive	love movie amazing acting
2	Positive	great story love characters
3	Positive	amazing and wonderful soundtrack
4	Negative	boring movie slow
5	Negative	terrible acting waste of time
6	Negative	boring story and bad script

A. Positive
B. Negative
C. Both equally
D. Neutral

17. What is the main goal of text classification?

A. Generating new text
B. Translating text into another language
C. Assigning labels or categories to text
D. Extracting entities from text

18. Why do modern NLP classifiers often use embeddings instead of one-hot encoding?

A. Embeddings are slower but simpler
B. Embeddings capture semantic meaning
C. One-hot encoding is more memory-efficient
D. Embeddings remove context

19. What does Laplace Smoothing solve in Naive Bayes classification?

A. It speeds up computation
B. It prevents zero probabilities for unseen words
C. It increases model complexity
D. It improves tokenization

20. What is the main advantage of neural networks over Naive Bayes in text classification?

A. They train faster
B. They capture nonlinear patterns and contextual meaning
C. They require less data
D. They use fewer parameters

21. What is the main goal of text preprocessing in NLP?

A. To visualize text data
 B. To clean and prepare raw text for modeling
 C. To train neural networks
 D. To remove all words from the text

22. Which step in preprocessing helps convert different forms of a word to a single root form?

A. Tokenization
 B. Normalization
 C. Stopword Removal
 D. Spell Checking

23. Which library is widely used for spell checking and is considered *superior* for fast correction?

A. SymSpell
 B. TextBlob
 C. SpaCy
 D. NLTK

24. Which of the following would lemmatization convert correctly, but stemming might not?

A. playing → play
 B. mice → mouse
 C. studying → study
 D. worked → work

25. What does tokenization do?

A. Converts text into numerical form called tokens
 B. Splits text into smaller units
 C. Removes stopwords
 D. Checks spelling

26. Dataset sample: "BUY CHEAP TOYS NOW!!!" and "Meeting tomorrow at 10 am."

Which cleaning step will best help you remove noise from this dataset?

A. Lowercasing
 B. Tokenization
 C. Normalization
 D. Spellchecking