1. Which of the following techniques can be used for keyword normalization in NLP, the process of converting a keyword into its base form?

a. Lemmatization

b. Soundex

c. Cosine Similarity

d. N-grams

2. Which of the following techniques can be used to compute the distance between two word vectors in NLP?

a. Lemmatization

b. Euclidean distance

c. Cosine Similarity

d. N-grams

3. What are the possible features of a text corpus in NLP?

a. Count of the word in a document

b. Vector notation of the word

c. Part of Speech Tag

d. Basic Dependency Grammar

e. All of the above

4. You created a document term matrix on the input data of 20K documents for a Machine learning model. Which of the following can be used to reduce the dimensions of data?

1. Keyword Normalization
2. Latent Semantic Indexing
3. Latent Dirichlet Allocation

a. only 1

b. 2, 3

c. 1, 3

d. 1, 2, 3

5. Which of the text parsing techniques can be used for noun phrase detection, verb phrase detection, subject detection, and object detection in NLP.

a. Part of speech tagging

b. Skip Gram and N-Gram extraction

c. Continuous Bag of Words

d. Dependency Parsing and Constituency Parsing

6. Dissimilarity between words expressed using cosine similarity will have values significantly higher than 0.5

a. True

b. False

7. Which one of the following are keyword Normalization techniques in NLP

a. Stemming

b. Part of Speech

c. Named entity recognition

d. Lemmatization

8. Which of the below are NLP use cases?

a. Detecting objects from an image

b. Facial Recognition

c. Speech Biometric

d. Text Summarization

9. In a corpus of N documents, one randomly chosen document contains a total of T terms and the term "hello" appears K times.

What is the correct value for the product of TF (term frequency) and IDF (inverse-document-frequency), if the term "hello" appears in approximately one-third of the total documents?

a. KT * Log(3)

b. T * Log(3) / K

c. K * Log(3) / T

d. Log(3) / KT

10. In NLP, The algorithm decreases the weight for commonly used words and increases the weight for words that are not used very much in a collection of documents

a. Term Frequency (TF)

b. Inverse Document Frequency (IDF)

c. Word2Vec

d. Latent Dirichlet Allocation (LDA)

11. In NLP, The process of removing words like "and", "is", "a", "an", "the" from a sentence is called as

a. Stemming

b. Lemmatization

c. Stop word

d. All of the above

12. In NLP, The process of converting a sentence or paragraph into tokens is referred to as Stemming

a. True

b. False

13. In NLP, Tokens are converted into numbers before giving to any Neural Network

a. True

b. False

14. identify the odd one out

a. nltk

b. scikit learn

c. SpaCy

d. BERT

15. TF-IDF helps you to establish?

a. most frequently occurring word in the document

b. most important word in the document

16. In NLP, The process of identifying people, an organization from a given sentence, paragraph is called

a. Stemming

b. Lemmatization

c. Stop word removal

d. Named entity recognition

17. Which one of the following is not a pre-processing technique in NLP

a. Stemming and Lemmatization

b. converting to lowercase

c. removing punctuations

d. removal of stop words

e. Sentiment analysis

18. In text mining, converting text into tokens and then converting them into an integer or floating-point vectors can be done using

a. CountVectorizer

b. TF-IDF

c. Bag of Words

d. NERs

19. In NLP, Words represented as vectors are called as Neural Word Embeddings

a. True

b. False

20. In NLP, Context modeling is supported with which one of the following Word embeddings ?

a. Word2Vec

b) GloVe

c) BERT

d) All of the above