BIO543: Big Data Mining Healthcare

(3rd April 2025, Quiz3)

Maximum Marks: 20 **Duration: 20 Minutes**

Name:_____ **Admission No: _____**

Instructions: Attempt all questions, each question carry one mark. Please tick only best possible option.

1. The main limitation of the Bag of Words (BOW) approach is:
   a) High computational cost
   c) Sparse representation
   b) Ignoring word order
   d) Complex implementation
2. Large models are subset of _____
   a) Large language models
   c) **Natural language processing**
   b) Hidden Markov models
   d) Deep learning techniques
3. Which of the following techniques is not developed for natural language processing
   a) Recurrent neural network
   c) LSTM
   b) LLM
   d) **SVM**
4. Which of the following is not an ANN based technique for computing word embedding
   a) **One hot encoding**
   c) Word2Vec
   b) GloVe
   d) FastText
5. Which of the following uses deep transformer-based models for computing embeddings
   a) Word2Vec
   c) **LLM**
   b) FastText
   d) GloVe
6. Which of the following uses "Viterbi Algorithm"
   a) **HMM**
   c) LSTM
   b) RNN
   d) LLM
7. Which of the following is not a protein language model
   a) ESM2
   c) ProtGPT-2
   b) **Proteomics**
   d) AlphaFold
8. Which of the following technique is not used for "Screening and Diagnosis" of cancer
   a) CT Scan
   c) MRI
   b) PET
   d) **HbA1c**
9. DNA and chemical modification come under which omics
   a) Genomics
   c) **Epigenomics**
   b) Proteomics
   c) Transcriptomics

10. Which technique is used for sequencing of genes and their expression
    a) **RNA-Seq** c) Proteomics
    b) Affymetrix d) cDNA microarray
11. In FASTAQ format which line is meaningless
    a) Line 1 c) Line2
    b) **Line 3** d) Line 4
12. Which software is used for spliced mapping of RNA-seq data
    a) Cuffdiff c) EdgeR
    b) **TopHat** d) RNA-SeQC
13. Which of the following database provides information on vaccine candidates for cancer
    a) CancerDR c) TCGA
    b) ArrayExpress d) **CancerTope**
14. What will be stage of cancer if it has spread to other parts of body
    a) Stage I c) Stage II
    b) Stage III d) **Stage IV**
15. What will be Jaccard similarity between set (1,1,0,0,1,1) and (1,1,0,0,1,1)?
    a) 0.5 c) -0.75
    b) 0.75 d) **1.0**
16. What will be Jaccard similarity between two signatures (2,2,1) and (2,4,1)
    a) **0.67** c) 0.5
    b) 0 d) 0.34
17. What is shingles for nucleotide sequence "ATGTGAC" for K=1
    a) {AT,TG,GT,GA,AC) c) **{A,T,G,C}**
    b) {A,T,G,T,G,A,C} d) {AT,GC,AC, GT}
18. If signature of two documents is identical or nearly identical than documents will be
    a) **Highly similar** c) Poorly similar
    b) No similarity d) None
19. In Min-hashing hash function is used for ____
    a) Reducing search space c) **Compress shingles**
    b) Reducing size of documents d) Compress documents
20. In LSH buckets are used for _____
    a) **Reducing search space** c) Compress shingles
    b) Reducing size of documents d) Compress documents