BIO543: Big Data Mining Healthcare

(15th April 2024, 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. Which of the following software is best software for big data mining?
   a) Scikit learn
   b) **Mahout**
   c) Weka
   d) MySQL
2. Which component of the Hadoop ecosystem is used for scripting?
   a) Mahout
   b) Hive
   c) **Pig**
   d) Ambari
3. Which of the following is not part of search engine (Lucene)?
   a) Crawling
   b) Ranking
   c) Indexing
   d) **Embedding**
4. What type of files are not extracted or analyze by Lucene search engine?
   a) PDF
   b) XML
   c) **TIFF**
   d) MS Word
5. What is purpose of shingling?
   a) Convert nucleotides to amino acid
   b) **Convert documents to sets**
   c) Convert SQL to NOSQL
   d) Computing Jaccard similarity
6. In which technique similar documents are grouped in buckets for similarity search.
   a) Min-Hashing
   b) **LSH**
   c) Curve clustering
   d) BLAST
7. What will be Jaccard similarity between set (1,1,0,0,1,1) and (1,0,1,1,1,1)?
   a) **0.5**
   b) 0.75
   c) -0.75
   d) 1.0
8. In which technique, large sets are converted to short signature.
   a) Shingling
   b) RNA-seq
   c) Word embedding
   d) **Min-Hashing**
9. In LSH, if document C1 and C2 have similarity 0.80 or more, b=20, r=5, what is percent probability that C1 and C2 will be in same band.
   a) 100 %
   b) < 95.00%
   c) **> 99.96%**
   d) < 90.25%

10. Which of the following is not part of cancer hallmarks?
    a) Evading growth suppressors
    b) **Maintaining apoptosis**
    c) Resisting cell death
    d) Inducing angiogenesis
11. Study of all mRNA molecules in a cell is called.
    a) Genomics
    b) Epigenomics
    c) Proteomics
    d) **Transcriptomics**
12. Which technique all to measure relative gene expression (e.g. disease vs healthy)?
    a) **cDNA microarray**
    b) RNA-seq
    c) Affymetrix
    d) DNA-seq
13. Which of the following is used for prioritization of anticancer drugs in CancerDP?
    a) **Transcriptomics**
    b) Glycomics
    c) Epigenomics
    d) Metabolomics
14. Which of the following database do not contain gene expression information?
    a) GEO
    b) **Uniprot**
    c) ArrayExpress
    d) TCGA
15. In which stage you will classify a cancer patient who have single tumor that has not grown into any blood vessels?
    a) Late-stage
    b) Stage-IV
    c) **Stage-I**
    d) Stage-II
16. Which of the following is simple feature not ML/DL trained features?
    a) FastText
    b) GloVe
    c) Word2Vec
    d) **TF-IDF**
17. Which of the following software used large language model for predicting protein structures?
    a) **AlphaFold**
    b) Modeller
    c) HHpred
    d) Robetta
18. Which of the following is not part of large language models
    a) **Transition probabilities**
    b) Attention
    c) Pre-trained models
    d) Fine tuning of models
19. Which of the following approach is not used in large language model?
    a) Sentence correction
    b) Text translation
    c) Text completion
    d) **Text similarity**
20. Which company developed large language model “Go AI” ?
    a) OpenAI
    b) Meta
    c) **DeepMind**
    d) Google