BIO542: Big Data Mining in Healthcare

(1st May 2024, End-Sem Exam)

Maximum Marks: 60 **Duration: 75 Minutes**

Instructions: This question paper have two sections, A and B. Attempt any 14 questions from section A, each question carries 2 marks (Total 28 marks). Attempt any 8 questions from section B, each question carries 4 marks (Total 32 marks). Write all answers in answer sheet only.

Section A

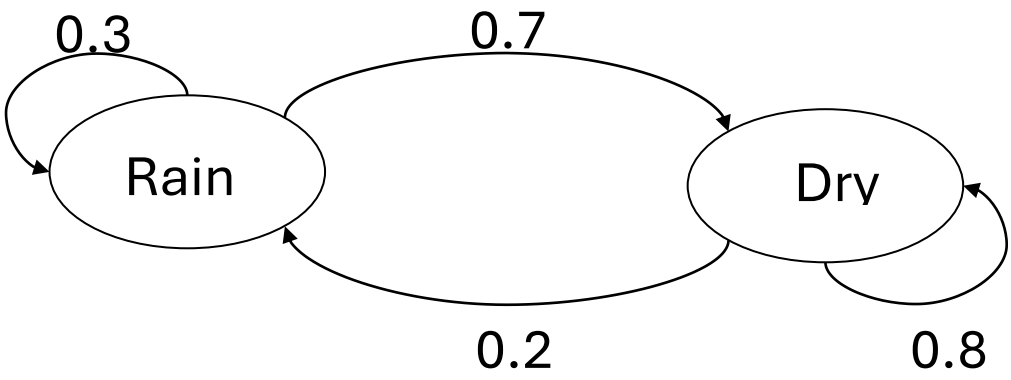
1. **Name two types of vectors implemented in Mahout software.**
   Ans - Dense Vector, Random Access Sparse Vector, Sequential Access Sparse Vector

2. **Expand following terms in 3C + FPM + O in Mahout software.**
   Ans - 3C - Collaborative Filtering, Clustering, Classification
   FPM – Frequent Pattern Mining
   O – Others

3. **List two popular methods used to generate vector embeddings of words**
   Ans - Latent Semantic Analysis/Indexing, Word2Vec, GloVe, FastText

4. **Write is the full form of LSTM and CBOW.**
   Ans - LSTM – Long Short Term Memory
   CBOW – Continuous Bag of Words

5. **Show Hidden Markov model by figure**



6. **List two large language models commonly used in computational biology**

   Ans – Any two from following

   AlphaFold, BioBERT, DNABERT, ProTrans, TAPE (Task-Agnostic Protein Embeddings), ESM (Evolutionary Scale Modeling)

7. **Compute Jaccard similarity between C_1 = [2,2,1,2,5]; C_2 = [2,4,1,2,5]**

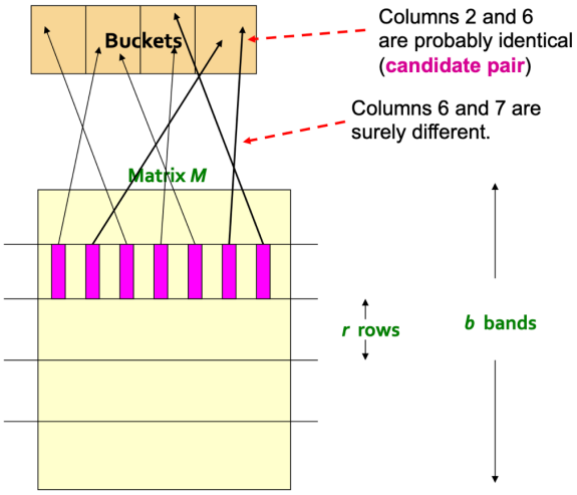
   Ans- 0.8 (4/5)

8. **What is signature and its application in Min-hashing**

   Ans - Signature - short integer vectors that represent the sets, and reflect their similarity

   Application in Min-hashing - a signature is used for fast and efficient estimation of the similarity between sets.

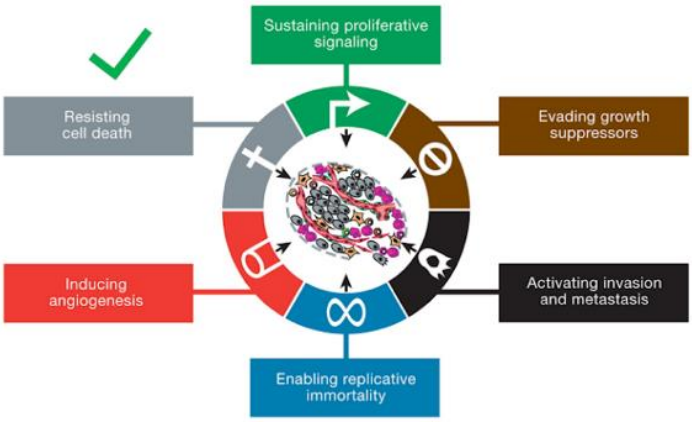
9. **Draw LSH diagram and show buckets, rows and bends**



10. **Write any four hallmarks of cancer.**

Ans – Any four from following

Cancer : Hallmarks



Hanahan *et al*, 2011

11. List application of CancerDP and CancerTSP

Ans - CancerDP - predicting priority/potency of an anticancer drug against a cancer cell line

CancerTSP – Thyroid Cancer Stage Prediction

12. Write full form of RPKM and TPM used in RNA-Seq

RPKM - Reads Per Kilobase Million

TPM – Transccripts Per Kilobase Million

13. Briefly describe partitional and hierarchical clustering

Ans-

- **Partitional algorithms:** Construct various partitions and then evaluate them by some criterion
- **Hierarchical algorithms:** Create a hierarchical decomposition of the set of objects using some criterion

14. What is difference between K-means and K-medoids clustering

Ans -

- In k-means, each cluster is represented by a centroid, which is the mean of all data points assigned to the cluster. The centroid might not necessarily be one of the data points.
- In k-medoids, each cluster is represented by a medoid, which is the most centrally located point in the cluster, i.e., the data point that minimizes the average dissimilarity to all other points in the cluster. Unlike centroids, medoids must be actual data points.

15. Write full form of UPGMA and NJ method

UPGMA - Unweighted Pair Group Method

NJ - Neighbor Joining

16. Write definition and difference between paralogs and orthologs

Paralogs: Genes that arise from a gene duplication within a species' genome.

- Having similarity in sequence but may have diverged in function overtime.
- Related but distinct functions within the same organism.

Orthologs: Genes in different species that evolved from a common ancestral

- Retain similar functions across different organisms, differences in sequence
- Play equivalent roles in different species
- Important for inferring evolutionary relationships and gene function

17. In FluSPred, how many proteins used for building prediction model

Ans - 15 types of Influenza A proteins

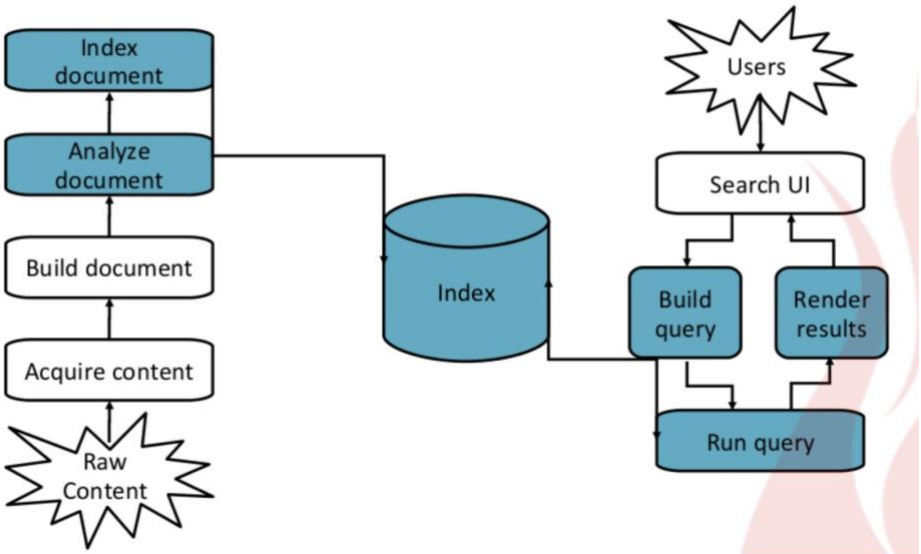
18. Write formula for weighted moving average or autocorrelation

$$X_t = \sum_{i=1}^{N} W_i \times X_{t-i}$$

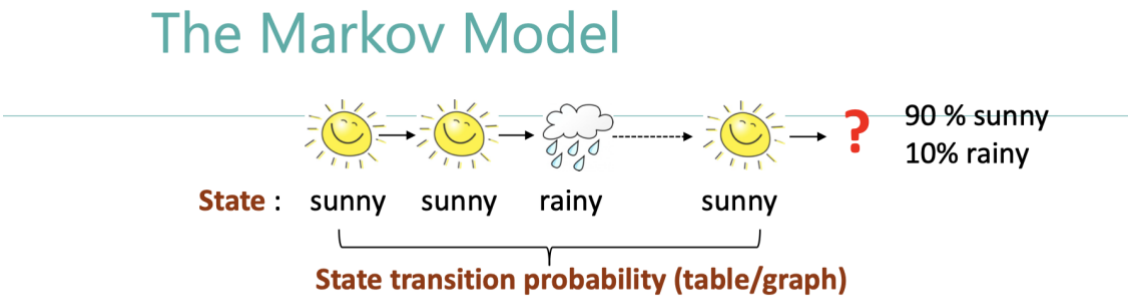
---------- **P.T.O** ----------

Section B

1. Show Lucene Architecture by diagram with labels



2. **In Hidden Markov model, show transition probabilities in all three formats.**



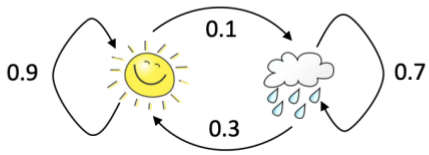
Output format 1:

Today	Tomorrow	Probability
sunny	sunny	0.9
sunny	rainy	0.1
rainy	sunny	0.3
rainy	rainy	0.7

Output format 2:

	sunny	rainy
sunny	0.9	0.1
rainy	0.3	0.7

Output format 3:



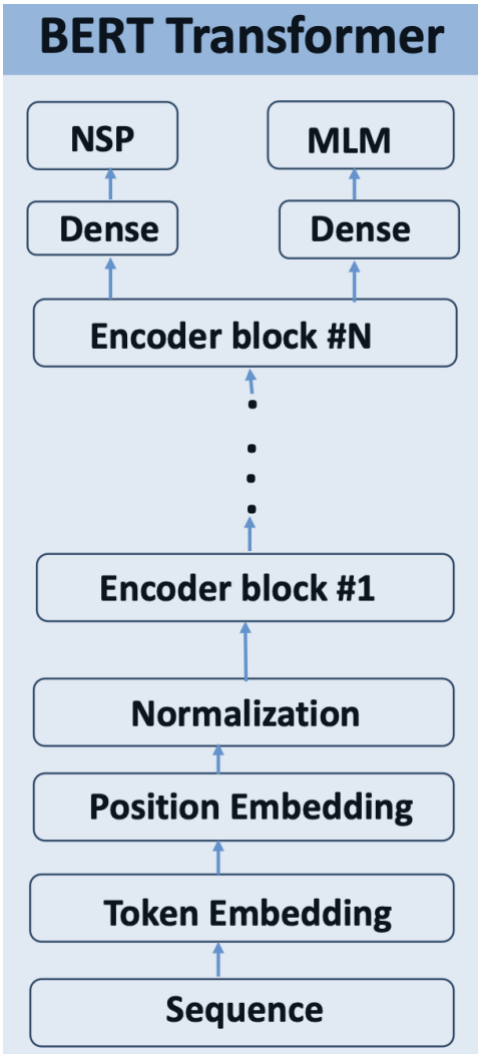
3. **List any four major building blocks of LLM with brief description of each block**

Ans – Any four from following

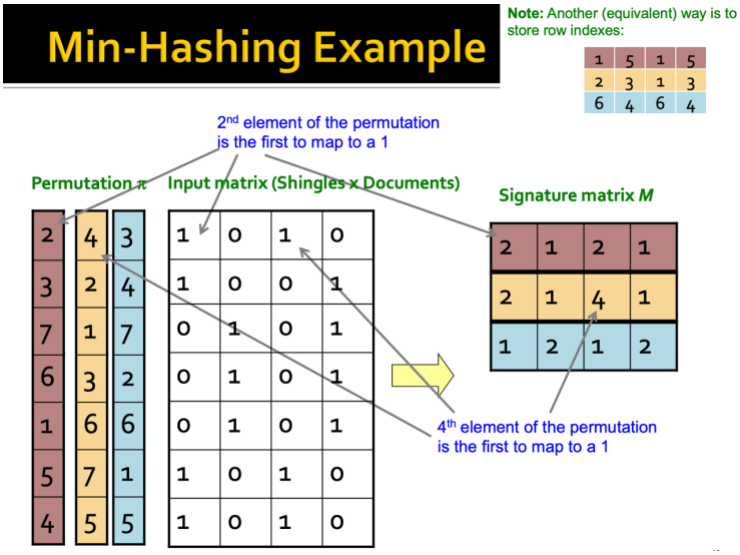
- **Pre-Training:** Learn general language and relationships between words.
- **Transfer Learning:** Knowledge gained from one task is used to improve performance on another related task.

- **Fine Tuning:** LLM pre-trained on one corpus of text data, then fine-tune its parameters for specific text classification, to improve its accuracy on domain-specific tasks.
- **Attention:** Assigns different weights to different parts of the input, allowing the model to prioritize and emphasize the most important information.
- **Embeddings:** Mathematical representations of words, phrases, or tokens, their semantic meaning and relationships in a large-dimensional space
- **Tokenization:** Breaking text down into the smallest unit of understanding

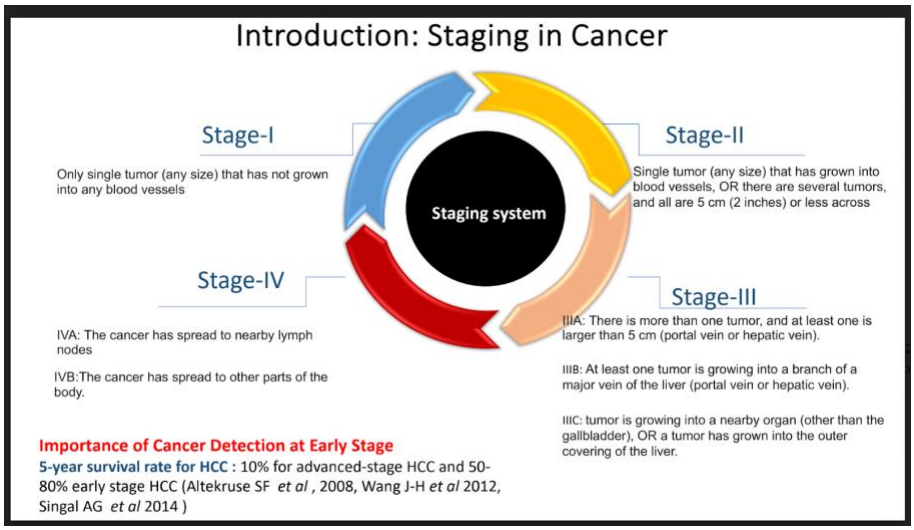
4. **Show BERT transformer by flowchart.**



5. **Create a signature matrix from raw matrix in Min-hashing.**



6. **Briefly describe all four stages of cancer**



7. **Write two example reads/sequences in FASTAQ format**

   Ans -

```
@HWI-ST508:210:C0EDTACXX:1:1101:1872:1227 1:N:0:
AATTGTGAAAACCCAAAAGGTGGAGCAGCCATTNTTATACATTGCAGAAGGGNGANNNANCNTTATGAAATTTAGCACCTGCCTTCCTGAAT
+
@CCFFEFFHHHHHJJJJIJJCGHEIIIJIJJJJ#1BFHIJJJJJJIJJIJJI#-;###-#-#-5?BFFFFEEEEEECCDDDDDDDDDCCDI
@HWI-ST508:210:C0EDTACXX:1:1101:1895:1233 1:N:0:
TGACATAAGCTTGCATTTGAAAAGCACCTCCGAAAGCTTCCCAGCCTCAAAGNCANNATCGNCTTCTGATGCAGTTAGGCACCACAAGAGCT
+
CCCFFFFFHHHHGJJJJJIIJJIIIHJJIJJJJJJJJJJJJJJJJJJGHJJ#.;##--;C#-5CEFFFFFEEECCEEDDDDDDDDDBDDDI
```

8. **Describe major steps/pass used in CURE algorithm**

Starting CURE

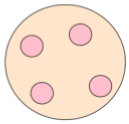
2 Pass algorithm. Pass 1:

- **0) Pick a random sample of points that fit in main memory**
- **1) Initial clusters:**
  - Cluster these points hierarchically – group nearest points/clusters
- **2) Pick representative points:**
  - For each cluster, pick a sample of points, as dispersed as possible
  - From the sample, pick representatives by moving them (say) 20% toward the centroid of the cluster

Finishing CURE

Pass 2:

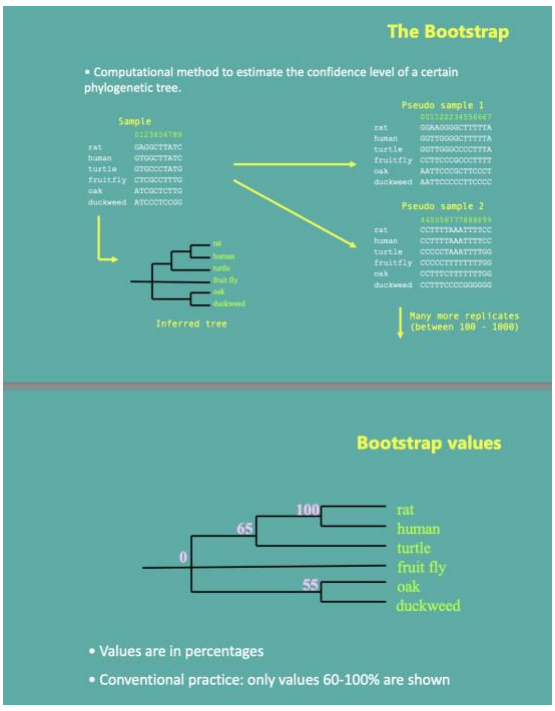
- Now, rescan the whole dataset and visit each point ***p*** in the data set
- **Place it in the "closest cluster"**
  - Normal definition of "closest": Find the closest representative to ***p*** and assign it to representative's cluster



9. Describe bootstrap algorithm and show concept by figure

***Bootstrapping*:** how dependent is the tree on the dataset

1. Randomly choose *n* objects from your dataset of *n*, *with replacement (picking columns from the alignment at random with replacement)*
2. Rebuild the tree based on the subset of the data
3. Repeat many (1,000 – 10,000) times



10. Show exponential smoothing method by example table, write formula

Univariate time series

Week	Temp	Fore
1	25	25
2	27	25
3	30	25.40
4	32	26.32
5	33	27.45
6	35	28.56
7	36	29.85
8	36	31.08
9	?	32.06
10	?	

Exponential smoothing method

$X_t = F_t = F_{t-1} + \alpha (X_{t-1} - F_{t-1})$
OR
$F_t = \alpha X_{t-1} + (1- \alpha) F_{t-1}$

where is α is smoothing function ($0 < \alpha < 1$). F_t is forecasted value for even t .

If α *is 0.2*

$F_1 = X_1 = 25$

$F_2 = 0.2 * X_1 + (1 - 0.2) * F_1 = 0.2 * 25 + 0.8 * 25 = 25 = X_1$

$F_3 = 0.2 * 27 + 0.8 * 25 = 25.40$

$F_{10} = 0.2 * F_9 + 0.8 * F_9 = F_9$