JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

TEST -2 EXAMINATION- April2024

M.Tech-II Semester (Data Science)

COURSE CODE (CREDIT): 22M11CI213(3) MAX. MARKS: 25

COURSE NAME: BIG DATA ANALYTICS

COURSE INSTRUCTORS: Er. Nitika MAX. TIME: 1 Hour 30 Minutes

Note: (a)*All questions are compulsory.*

(b)Marks are indicated against each question in square brackets.

(c) The candidate is allowed to make Suitable numeric assumptions wherever required for solving problems

---

Q1. Suppose you have a transaction database containing records of items purchased by customers. Each transaction consists of a set of items. Use Apriori algorithm, find the frequent itemsets. Assume that minimum support threshold (s = 33.33%) and minimum confident threshold (c = 60%)

Assume the following transaction database:

Transaction 1: {Hot Dogs, Buns, Ketchup}

Transaction 2: {Hot Dogs, Buns}

Transaction 3: {Hot Dogs, coke, Chips}

Transaction 4: {Chips, coke}

Transaction 5: {Chips, Ketchup}

Transaction 6: {Hot Dogs, Coke, Chips} [4]

Q2. How does the damping factor influence the computation of Page Rank? [3]

Q3. Consider stream S= 1,3,2,1,2,3,4,3,1,2,3,1 and hash function h(x)= (6x+1) mod 5. Find the count of distinct elements using Flajolet Martin Algorithm. [5]

Q4. What do Big Data filters do? Explain Bloom Filter with example. [4]

Q5 Describe the concept of market baskets and how they are utilized in data analysis? [4]

Q6. Suppose you have a stream of data representing user queries in a search engine, and you want to find the most popular queries within a decaying window of the last 5 minutes. Use a decay factor a=0.9 and a normalization constant c=0.1.

Sequence of queries over time:

t = 1 minute: "politics", "sports", "music", "technology", "politics"

t = 2 minutes: "sports", "technology", "politics", "politics", "music"

t = 3 minutes: "politics", "sports", "sports", "politics", "music"

t = 4 minutes: "technology", "politics", "sports", "sports", "sports"

t = 5 minutes: "politics", "sports", "politics", "politics", "sports"

Find the most popular queries within the last 5 minutes using Decaying window algorithm. [5]