COMP9313: Big Data Management

Sample Exam Questions

Question 1 HDFS

- Explain the difference between NameNode and DataNode.

- Given a file of 500MB, let block size be 150MB, and replication factor=3. How much space do we need to store this file in HDFS? Why?

Question 2 Spark

- Given a large text file, your task is to find out the top-k most frequent co-occurring term pairs. The co-occurrence of (w, u) is defined as: u and w appear in the same line (this also means that (w, u) and (u, w) are treated equally). Your Spark program should generate a list of ***k*** key-value pairs ranked in descending order according to the frequencies, where the keys are the pair of terms and the values are the co-occurring frequencies (**Hint:** you need to define a function which takes an array of terms as input and generate all possible pairs).

```
textFile = sc.textFile(inputFile)
words = textFile.map(lambda x: x.lower().split())

// fill your code here, and store the result in a pair RDD avgLen

avgLen.collect()
```

Question 3 Finding Similar Items

Suppose we wish to find similar sets, and we do so by min-hashing the sets 10 times and then applying locality-sensitive hashing with k=5 and l=2.

If two sets had Jaccard similarity 0.6, what is the probability that they will be identified in the locality-sensitive hashing as candidates (i.e. they hash at least once to the same super-hash)? You may assume that there are no coincidences, where two unequal values hash to the same hash value.

Question 4 Mining Data Streams

Suppose we are maintaining a count of 1s using the DGIM method. We represent a bucket by (i, t), where i is the number of 1s in the bucket and t is the bucket timestamp (time of the most recent 1).

Consider that the current time is 200, window size is 60, and the current list of buckets is: (16, 148) (8, 162) (8, 177) (4, 183) (2, 192) (1, 197) (1, 200). At the next ten clocks, 201 through 210, the stream has 0101010101. What will the sequence of buckets be at the end of these ten inputs?

Question 5 Recommender Systems

Consider three users u_1, u_2, and u_3, and four movies m_1, m_2, m_3, and m_4. The users rated the movies using a 4-point scale: -1: bad, 1: fair, 2: good, and 3: great. A rating of 0 means that the user did not rate the movie. The three users' ratings for the four movies are: $u_1 = (3, 0, 0, -1)$, $u_2 = (2, -1, 0, 3)$, $u_3 = (3, 0, 3, 1)$

- Which user has more similar taste to u_1 based on cosine similarity, u_2 or u_3? Show detailed calculation process.
- User u_1 has not yet watched movies m_2 and m_3. Which movie(s) are you going to recommend to user u_1, based on the user-based collaborative filtering approach? Justify your answer.