CS172: Information Retrieval

UC Riverside

Midterm 2 Topics Winter 2019

Midterm 2 is **cumulative** but will focus on topics covered after Midterm 1. Below are the the topics that we covered after Midterm 1. Use the quizzes and lecture notes as a study guide.

- Understand the architecture of a crawler.
- Understand the purpose of duplicate detection and how the SimHash function works.
- Know JaccardSim metric.
- Know how MapReduce works and be able to write pseudo-code to solve a particular problem using MapReduce.
- Know the design specifications of the Distributed File System. How does it handle fault tolerance? What is the replication policy?
- Understand the PageRank algorithm. Be able to generate the transition matrix. Be able to compute the first several steps of PageRank if given an example. Be able to explain the purpose of PageRank of taxation (Chapter 4).
- Understand how we can translate the Pagerank algorithm to a MapReduce job.
- Understand why we must apply hypothesis testing, and significance tests (Chapter 8).
- Understand how we can apply clustering in Information Retrieval (Chapter 9).
- Understand the KMeans clustering algorithm and some challenges we may face when applying the algorithm. Be sure to review the slides on how to handle certain issues or challenges (Chapter 9).

1. KMeans: Explain what the sum-of-squares is for k-means. Would you be able to explain this measure or apply it to a problem?

2. KMeans: Name three challenges or caveats when using KMeans and steps or strategies taken to handle/solve the problem.

3. KMeans: Consider the data set. (A, 0.1) , (B, 0.6) , (C, 0.8) , (D, 2.0), (E, 3.0)

   Apply k-Means Clustering to this data set for k=2, assuming that we choose the initial cluster centers to be A and B.

   Step 1 - Write down the cluster assignments that result. Write C, D, and E in the blanks below according to which cluster they are assigned (A and B are already assigned).
   cluster 1: A,
   cluster 2: B,

   Step 2 - After assigning examples to clusters, the next set is to recompute the cluster centroids.
   cluster 1 centroid:
   cluster 2 centroid:

   Step 3 - After recomputing the cluster centers, you reassign the examples to the clusters to which they are closest.
   cluster 1:
   cluster 2:

   Step 4- After assigning examples to clusters, you recompute the cluster centers.
   cluster 1 centroid:
   cluster 2 centroid:

   Step 5 - After recomputing the cluster centers, you reassign the examples to the clusters to which they are closest.
   cluster 1:
   cluster 2:

4. What are some of the challenges in building a crawler?
5. Describe the architecture or flow of a typical crawler.
6. Describe MapReduce architecture? Describe how a job is run in parallel and how the various tasks of the job is managed.
7. What is the purpose of a Combiner and Partitioner in a MapReduce job.
8. PageRank problem. Given the following hyperlink structure among web pages. Show the construction of the transition matrix. Show the setup of the PageRank computation with taxation using $\beta = 0.1$