Web Mining

Assignment 2 (30 points, 5% weightage)

Date Posted: Aug 29, 2013
Submission Deadline: Sep 9, 2013. 9pm.

Goal: This assignment deals with the course content covered in class until August 22.

[Basic Difficulty: Information] To assess the understanding of the course content.
[Medium Difficulty: Observation, Comparison and Generalization] To test if the students can apply that understanding to new scenarios.
[High Difficulty: Intuition] To find if the students can apply intuitive thinking in solving a new problem in that area.

Submission Instructions: Submit a single file with the answers. You can use MS Word or Latex to typeset your answers. The file name should be <rollno>_as2.pdf

Questions

1. Ranking
    a. [Basic – 1 point] Given a document D and a query Q, how do you compute the Okapi BM25 similarity between D and Q? Provide the formula and explain its components.
    b. [Medium – 2 points] The following is a retrieval formula for scoring a document D with respect to a query Q.

$$score(D,Q) = \sum_{t \in Q, t \in D} c(t,D) \times \log\left(\frac{df(t)}{N+1}\right)$$

    where c(t, D) is the count of term t in D, N is the total number of documents in the collection, and df(t) is the document frequency of t. Point out two most important reasons why this formula is unlikely performing as well as a modern retrieval formula such as BM25 (i.e., Okapi).
    c. [High – 3 points] Okapi BM25 fails when the document is very very long. Why? Suggest a slight modification to Okapi BM25 to take care of this problem.
2. Similarity Search
    a. [Basic – 1 point] In the banding technique, if the Jaccard similarity between a pair of documents is 0.8 and we are using 20 bands for minhash signatures of size 200, what is the probability that the banding technique will output the pair of documents as similar?
    b. [Medium – 2 points]
    Analysis of the Banding Technique: Suppose we use b bands of r rows each, and suppose that a particular pair of documents have Jaccard similarity s.
    The probability that the signatures agree in all rows of one particular band is ____.
    The probability that the signatures do not agree in at least one row of a particular band is ____.
    The probability that the signatures do not agree in all rows of any of the bands is _____.

The probability that the signatures agree in all the rows of at least one band, and therefore become a candidate pair, is _____.

c. [High – 3 points]
Show that 1 - Jaccard(x, y) satisfies the triangle inequality.

3. Link Analysis Algorithms
   a. [Basic – 1 point]
   Mention 2 ways of handling dead ends when performing pagerank iterations.
   b. [Medium – 2 points]
   We studied link spam in the context of PageRank. Does HITS also suffer from spam? If so, explain how spammers can spam HITS.
   c. [High – 3 points]
   Write a short pseudo-code for implementing PageRank using MapReduce.
4. LSI and EM
   a. [Basic – 1 point]
   Compare the output of clustering using KMeans and EM where clusters in EM are modeled using Gaussians.
   b. [Medium – 2 points] In a Gaussian Mixture Model, what are the latent variables and what are the parameters of the model?
   c. [High – 3 points]
   EM is guaranteed to find a local optimum of the Likelihood function. Give one way of escaping this local optima trap and obtaining a result closer to the global optimum.
5. Social Recommender Systems
   a. [Basic – 1 point]
   For the simplest collaborative filtering, predicted rating by user u for product i is ANSWER1+$\gamma(\sum_{v=1}^{n} ANSWER2(ANSWER3 - ANSWER4))$ where γ is a normalization constant.
   b. [Medium – 2 points]
   Given the list of predicted ratings and the actual ratings, mention 2 metrics to compute the accuracy of the rating prediction algorithm.
   c. [High – 3 points]
   In the class, we discussed various applications of recommendation systems. Auto-complete feature in software IDEs (like Eclipse or Visual Studio) can also be considered as recommendation systems. Write in brief about issues in recommending a method name to be called for an object.