CIS 563 - Intro to Data Science
HW #6
Reza Zafarani - Fall 2021

Due: Dec 6, 2021, 11:59 PM

Problem 1 - Shingling

- Assume a document has length N, what are the maximum and minimum size of the k-shingle ($k < N$) set? What if we use a bag (multiset) instead of a set?
- Using a 2-shingle set {ab, bc, cd, da}, how many kinds of length N ($N > 4$) documents can we generate?

Problem 2 - Min Hashing

Element	S_1	S_2	S_3	S_3
0	0	1	0	1
1	0	1	0	0
2	1	0	0	1
3	0	0	1	0
4	0	0	1	1
5	1	0	0	0

- (Book Exercise 3.3.3) For the above Table:
  1. Compute the minhash signature for each column if we use the following three hash functions: $h_1(x) = 2x + 1 \bmod 6$; $h_2(x) = 3x + 2 \bmod 6$; $h_3(x) = 5x + 2 \bmod 6$.
  2. Which of these hash functions are true permutations?
  3. How close are the estimated Jaccard similarities for the six pairs of columns to the true Jaccard similarities?
- Consider a general hash function $h(x) = ax + b \bmod c$, what kind of a, c can guarantee to generate a true permutation.
- (Book Exercise 3.3.5) Prove that if the Jaccard similarity of two columns is 0, then min hashing always gives a correct estimate of the Jaccard similarity.

Problem 3 - LSH

- To select document pairs by the LSH, we set b=10 and r=10. What is the expected false negative rate for documents with 90% similarity ?
- (Book Exercise 3.4.4) Suppose we wish to implement LSH using MapReduce. Specifically, assume chunks of the signature matrix consist of columns, and elements are key-value pairs where the key is the column number and the value is the signature itself (i.e., a vector of values).

  a Show how to produce the buckets for all the bands as output of a single MapReduce process. *Hint*: Remember that a Map function can produce several key-value pairs from a single element.

  b Show how another MapReduce process can convert the output of (a) to a list of pairs that need to be compared. Specifically, for each column i, there should be a list of those columns $j > i$ with which i needs to be compared.

Problem 4 - Bloom Filter

Implement the Bloom filter. Use:

`http://www.stopforumspam.com/downloads/listed_username_30.zip`

as your set S. This is a set of usernames known to be spam for the last 30 days. Select a proper hashing memory size (n) and find the optimal number of hash functions (k). Use the spam usernames for the last 365 days:

`http://www.stopforumspam.com/downloads/listed_username_365.zip`

as your stream. Submit your (1) code, (2) optimal k for your n, and (3) the percentage of false positives all on blackboard. There is no need to submit your datasets.

Note. For hashing you can use

- **murmurHash**: `https://sites.google.com/site/murmurhash/`
- **FNV**: `http://isthe.com/chongo/tech/comp/fnv/`
- **Jenkins Hash**: `http://www.burtleburtle.net/bob/hash/doobs.html`
- Or you Hash Function of choice.

Problem 5 - FM Method (Takes Time!)

Implement the Flajolet-Martin (FM) algorithm. Count the number of distinct quotes (quotes are denoted with lines that start with Q) in the MemeTracker dataset (all files):

`https://snap.stanford.edu/data/memetracker9.html`

Submit (1) the estimated number from FM and (2) your code on blackboard. In your implementation, use the method discussed in Section 4.4.3 to provide more accurate results. Do not to submit your datasets!

Note: The dataset is about 50-60 GB uncompressed. It can be easily downloaded on SU network.