Data Mining Lab 1

Akhil Yerrapragada
akhily@kth.se

Gibson Chikafa
chikafa@kth.se

1 Assignment Goals

In this homework, we implement the stages of finding textually similar documents based on Jaccard similarity using the shingling, minhashing, and Locality-Sensitive Hashing (LSH) techniques. The implementation is done using Scala.

2 Solution overview

Our solution comprises of the following steps:

- Clean the document text. This includes removing punctuation marks, transform all letters to lower cases, convert extra white space as a single blank, and remove white spaces up front and at the end, and Split the text string into separate words.
- Perform shingling of the documents by creating a set of unique shingles of length k (k-shingles) from the content of the document.
- Then we map each shingle to an integer value using hashShigles method that uses the *MurmurHash3* algorithm.
- Compute the union of all shingles from all documents.
- We compute the characteristic matrix that have 1 in row i and column j if and only if document j contains the shingle i and 0 otherwise. Shingle i is taken from the union. For small documents, we can compute the Jaccard similarity from the characteristic matrix using the function *jaccardSimilarities*. For bigger and more documents we proceed to use minHashing
- Randomly generate n hash functions of the form ((ax + b) mod c) by randomly generating n values for coefficients a and b. c is chosen to be a constant prime number
- For a given shingle set, minhash signature of size n is generated as follows. Each of n hash functions are applied to each element of given shingle set. And then minhash value which is the minimum hash value among all elements for a given hash function is selected. Thus, n minhash value set is generated for the given shingle set.

- Finally, similarity between two documents is computed by calculating the ratio of identical minhash values among total n minhash values.

3 How to run

- We use HDFS to store our documents. Move the documents in folder *documents* to HDFS base dir.
- Compile using *sbt compile*
- Run using *sbt run*