Homework 1: Finding Similar Items: Textually Similar Documents

Hiranmayi Pechetti, Axel Kaliff

15 November 2021

1) Introduction

The goal of the assignment is to implement the stages of finding textually similar documents based on Jaccard similarity using the shingling, minhashing, and locality-sensitive hashing (LSH) techniques and corresponding algorithms. The implementation has been done using no framework and in Python . To test and evaluate the implementation we wrote a program that uses our implementation to find similar documents in a corpus 100 documents that were in html format.

2) Code Explanation

We wrote the code in a Jupyter notebook. For every task in the assignment we created classes. Each class is explained below:

1. Shingling: Under class Shingling we created a function called ***k_shingle.*** Given a document we get a list of hashed shinglings by going through the document character by character and hashing that character and the ***k*** number of characters after it, appending the hashed characters to an array.
2. CompareSets: Using the formula

   $$sim(set1, set2) = |set1 \cap set2| / |set1 \cup set2|$$

   We compare two sets of integers (set1, set2) and calculate their jaccard similarity (sim). The union and the intersection of the sets are calculated using the standard Python library. We deemed it appropriate to round the result to three decimal points.
3. MinHashing: We created a function ***get_signature*** under the class MinHashing. It takes a list of shingles as input and computes a list of signatures.
4. CompareSignatures: We created a function ***compare*** under the class CompareSignatures. It takes 2 minhash signature vectors as input and returns the similarity as a fraction of components they agree with.

5. LSH: We created a function locality_sensitive_hashing uses banding and hashing and finds candidate pairs of signatures that agree on at least fraction t of their components given a collection of minhash signatures (integer vectors) and a similarity threshold t.

3) How to run the code

1. Open terminal and run
2. Open the file "Homework 1.ipynb" in jupyter notebook.
3. Run the cell
4. The output will be:

```
For 100 documents the execution times are:
Shingling
Jaccard similarity between document 3 and 15 is 1.0
Jaccard similarity between document 29 and 52 is 1.0
Execution time for Shingling is 0.73 seconds
MinHashing
Jaccard similarity between document 3 and 15 is 1.0
Jaccard similarity between document 29 and 52 is 1.0
Execution time for MinHashing is 1.023 seconds
LSH
Execution time for LSH: 0.012 seconds
```

Figure 1: Final Output