Detecting Duplication documents using LSH

1. Data description
2. IF-IDF method
3. LSH

3.1 Introduction

The idea behind locality sensitive hashing is to take the document fingerprints and chop them up into pieces, each piece being some number of minhashes. Since a single minhash (single entry in the fingerprint) has a probability equal to the Jaccard similarity of producing a collision, each chopped up portion of the fingerprint should as well. This chopped up portion is the locality in locality sensitive hashing, the hashing is just a hash function (any hash function) which produces a bin ID from the fingerprint locality being hashed. Each bin holds the entire fingerprint (with optional meta information) of the document and that of other documents that hash to the same bin.

Let's say our fingerprint has 100 minhashes in it and we chop the fingerprints into 10 pieces. Each piece of each fingerprint therefore contains 10 minhashes, we hash those again (not using minhash this time) to get a bin ID and store the whole fingerprint in every bin each of the pieces happens to land in.

When we want to know which documents are similar to a query document, we look in all the bins the query document lands in, any document in any of the bins is a potential duplicate. Comparing the full fingerprint of all documents in the bin or computing the actual Jaccard similarity between the shingle sets yields the final similarity of documents. Crucially since not all documents will land in the same bins we've reduced the number of comparisons needed to find similar or near duplicate documents.

The number of pieces to chop each fingerprint into and the size of each piece are parameters that need to be set. These should be set such that num\_pieces \* size\_of\_piece = num\_minhashes, this makes sense since having computed all the N minhashes we want to use all of them in the locality sensitive hashing part. There is however a further issue that needs to be considered when setting the parameters; the relation between the number and size of the pieces and the probability of LSH "finding" a pair of similar documents.

LSH is a probabilistic model which means that it won't always do the "right thing". Using LSH one needs to consider the similarity of a pair of documents (in this case the Jaccard similarity) and the probability that LSH will find that pair to be similar (a true positive, i.e. a correctly discovered duplicate pair). The pair of documents LSH finds to be similar should be thought of as candidate duplicates. The higher the probability, or guarantee, that LSH will find a pair of documents to be similar the falser positives the model will also produce, that is candidate duplicates that are not in fact duplicates.

* 1. Experiment

1. Result comparison