Text Technologies for Data Science Assignment 3

s1140740

November 2, 2014

Introduction

The purpose of this report is to provide an overview of decisions made in designing a plagiarism detection tool and cover different types of plagiarisms detected.

Plagiarism types

Type 1 Plagiarisms of this type are exact duplicates, e.g. documents `t104` and `t4172` which are classified as type 1 are exactly the same, word for word.

Type 2 Plagiarisms of this type are near duplicates, e.g. documents `t1088` and `t5015` which are classified as type 2 are almost the same, except document `t1088` is missing the word *room*.

Types 1 and 2 detection

The Simhash algorithm was used to detect these types of duplicates. For each document an md5 hash is computed and the document is then tokenized on white-space and non-alphanumeric characters with stop words removed.*

Furthermore, a 128-bit fingerprint is computed using the Simhash algorithm with md5 as a hashing function. This fingerprint is then split into l chunks of size k and each of these chunks is then stored as a key in a hash-table with the document as a value.

* List of English stop words obtained from `https://github.com/Alir3z4/stop-words` by Alireza Savand. No changes made.

Type 1 detection

To detect plagiarisms of type 1 a hash of a document is compared to every document in the same buckets. If the hashes are identical then we flag the stories as exact duplicates. Note, that two stories that are identical are guaranteed to end up in the same buckets as their vector space representation and thus their Simhash fingerprints are the same.

100% precision and recall compared to the file `type1.truth` are achieved using this technique.

Type 2 detection

To detect plagiarisms of type 2 a cosine similarity measure is computed for the document and all other documents in the same buckets. If this measure is above a certain threshold, in this case 0.8 we flag the documents as near duplicates.

Experiments were ran on different values of l. With l equal to 8, 100% precision and recall compared to the file `type2.truth` are achieved.

Notes