

Text Technologies for Data Science Assignment 3

s1140740

November 1, 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*.

Type 1 detection

Detecting plagiarisms of type 1 was easy. The content of each story is extracted and if it is the first time we encountered this content it is stored in a dictionary as a key with the story id as a value. However, if we have seen the exact same content before we flag the documents as duplicates.

Type 2 detection

Notes