CSCU9YE Assignment Report

Student no: 2519302

November 17th 2018

Contents

1 Introduction

BIG ASS TODO

2 Pre-Processing

This section will aim to cover the different aspects of pre-processing employed in this project. Within the datasets used in this project, we can see that the general steps one would need to take are as follows:

- Lemmatization
- Removal of stop words
- Removal punctuation

2.1 Lemmatization

First version: Lemmatization is a technique employed to remove inflection and/or derive the base form of a word in the dataset. As an example, the words, *cleaning*, *cleaner*, *cleanliness* can all be reduced to its base word, *clean*.

Such mapping then allows us to find all relevant documents using a specific word.

Second version: Lemmatization is a form of text normalization where the aim is to remove inflection and/or derive the base word from a family of words in the dataset. As an example:

Word		**Lemma**
Cleaning	$\rightarrow$	Clean
Cleaner	$\rightarrow$	Clean
Cleanliness	$\rightarrow$	Clean

2.2 Stop Words

//todo

2.3 Punctuation

//todo