CSCI 455: Project 1

Brayden Faulkne and Christina Hintonr

February 24, 2019

1 Introduction

The main purpose of this program is to go through a set of html files, and find how frequently each term is used in the actual text of the page. The program also handles preprocessing of this text, including stemming and stop word removal.

2 Overview

The program works by reading in each individual file, getting the html free text, writing that text to a new file, then appending the text the a string containing all the text from all previous files. Once all the text is retrieved, all characters in the string all set to lower case, and punctation is translated to whitespace. Then the string is split into the individual words using the NLTK libraries tokenize function.

References

[1] Richardson, Leonard. "Beautiful Soup." *Beautiful Soup: We Called Him Tortoise Because He Taught Us.*, www.crummy.com/software/BeautifulSoup/. Accessed 24 Feb. 2019.

[2] "Natural Language Toolkit." *NLTK*, NLTK Project, 17 Nov. 2018, www.nltk.org/.