Thomas Walsh

Dr. Dligach

COMP 329 001

26 January 2020

Homework 1 Findings

This paper is a short summary of my findings while completing Assignment 1. The assignment itself was to create a program that classifies a review to be either positive or negative. I will discuss my approach, my results, and certain challenges that I have encountered.

I was given two large text files containing several reviews, one file was filled with positive reviews and the other was filled with negative ones. I knew I wanted my classification system to identify positive and negative words in a review as part of the classification process. So, I made a program that created a list of the most frequently used words for each of the text files. I then went through the lists and saved all positive/negative terms that appeared over thirty times. I then made my classification program and for each word in the review it checks to see if that word exists in the list of positive or negative words. If it is a match it adds to a counter. At the end of the review the program compares the counters and whichever one is higher determines the nature of the review.

My approach seems to work relatively well, I have tested it 15 times with a mixture of good and bad reviews and it correctly identified 12 of them. Of the reviews that it had missed it was simply an issue of adding more words to the word bank. Of course, The word bank method has its limits (which I will talk about shortly) so if I were to improve the program I would make another method that searches for simple ratings both in numerical form (9/10, 15/100, 2/5) and written form (4 out of 5 stars, twenty percent). This could be used in tandem with the word bank to avoid mistakes.

The challenges I found were pretty interesting. There are very few reviews that don’t address both the negative and the positive so very rarely is there a review that exclusively uses words from one word bank. Another interesting thing that I found is that positive reviews seem to have a bigger vocabulary than negative ones. The list of negative words was at least half the size of the positive list. So to fix this I put a weight on all negative words. One negative word counts for three positive words. This fix did seem to work out pretty well, I do wonder if three is too much of a weight though. I also found that this method is not as useful on shorter reviews. The longer the review the better.