Project 3

Dreycen Foiles [dfoiles2]
Olek Yardas [yardsol2]

November 27, 2022

1 Introduction

For this project, we are investigating how we can predict the sentiment of a movie review automatically by only using a relatively small vocabulary of words. This is an interesting project because it is our first attempt at natural language processing which is an incredibly important aspect of current machine learning focus. To accomplish this task, we use some of the techniques recommended by Professor Feng. We were surprised to find that highly accurate sentiment prediction models can be created without using advanced techniques such as neural networks. We were able to get results that surpassed the benchmarks using only logistic regression.

2 Methods

The project consisted of two main distinct parts. The first part where we extracted data from the `alldata.tsv` file. The second part where we used the extracted data to train a model to predict the sentiment of a movie review. In the following two sections, we will discuss the methods we used to accomplish these tasks.

2.1 Vocabulary Generation

2.2 Sentiment Prediction

3 Results

All runs were performed on a custom desktop PC with an Intel i5-9400 2.9 GHz and 16 GB of RAM.

3.1 Vocabulary Generation

Final vocabulary size: 935. Time to generate vocab 4 minutes 13 seconds. Some examples of words in our final vocab list include:

‘actors’, ‘adds’, ‘adds_to’,‘10_10’, ‘surprisingly_good’, ‘this_bad’,‘waste’, ‘wrong’, and ‘would_recommend’

3.2 Sentiment AUC on Splits

Fold	Runtime (s.)	AUC
1	15.995	0.964
2	15.392	0.964
3	18.329	0.963
4	15.387	0.964
5	15.422	0.963

4 Conclusion