Pedro Vincenty

WA 2023

Homework 3 Report

Utilizing the program I created in Homework 2, I scraped ratings and reviews from 100 movies on rotten tomatoes to build a model that predicts a movie’s ratings based on its reviews. I created a count vectorizer classifier with an accuracy near 100% (99.36%) utilizing unigrams and bypassing English stop words. My program read 5 pages of reviews for Martin Scorsese’s Silence, ran them through the same count vectorizer, and predicted the ratings of the reviews with my new logistic model at an accuracy of 89%.

The below data frame shows the first 10 records where the model incorrectly predicted the rating of the movie.

|  | **Review** | **Rating** | **Prediction** |
| --- | --- | --- | --- |
| 2 | Silence is an example that with passion a film… | 1 | 0 |
| 8 | It’s not quite the “priestsploitation" nonsens... | 0 | 1 |
| 11 | Silence is a dreadful film from beginning to end. | 0 | 1 |
| 16 | The film completes what would be Scorsese's sp... | 1 | 0 |
| 35 | Its complex message about faith and the faithf... | 0 | 1 |
| 54 | Basically, Silence looks and sounds great. But... | 0 | 1 |
| 63 | Even at its unnecessary length, Silence is sel... | 1 | 0 |
| 65 | By the end of the ordeal, the question is stal... | 0 | 1 |
| 79 | An ambitious tape that limps at times. However... | 1 | 0 |
| 93 | As one of the unarguable geniuses of American ... | 0 | 1 |
|  |  |  |  |

The bar graph below shows the prediction accuracy of the model by the number of predictions made.

A green and red bar graph

Description automatically generated