Homework 1: Sentiment Analysis

Joshua Wallace
Princeton University, Dept. of Astrophysical Sciences
joshuajw@princeton.edu

Abstract

1 Introduction

2 Related Work

[2]

3 Methods

3.1 Naive Bayes

I first used the Naive Bayes classifier from the `nltk` package [1]. I tried different minimum number of words to be counted as vocab. For 3 I got training: 86.4% accurate and test 78.8% accurate, for 4 I got training: 85% accurate and test 78.3% accurate, for 5 I got training: 84.1% accurate and test: 77% accurate, and for 6 82.8% accurate for training and 77.3% accurate for training.

I also tried to do a little bit of "feature selection" on my own. I identified some words I thought would have no impact on the sentiment of a review such as brand names (e.g., "samsung" and "bluetooth") and food items (e.g., "potato" and "taco"). I then ran the `nltk` Naive Bayes analysis on these. I chose these words because it seemed to me that these words were not positive or negative in themselves but rather were focused on the product being reviewed itself. This had a marginal effect on the percentages: some went up 0.1–0.2%, some went down 0.1–0.2%, some remain unchanged. This version of feature selection did not seem to offer any improvement (especially since it made some accuracies higher and some lower), and our feature set isn't so large as making it smaller by a few words to speed things up, so I decided to not use this "feature selection".

Implementing the `sklearn` [2] Gaussian Naive Bayes had worse results overall: for 3, 81.8% training and 71.7% test; for 4, 79.3% training and 71.8% test; for 5, 77.9% training and 70.5% test; for 6, training 76.5% and test 69.7% accurate.

4 Results

5 Discussion and Conclusion

Acknowledgments

References

[1] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python*. " O'Reilly Media, Inc.", 2009.

[2] Jun Zhu, Amr Ahmed, and Eric P Xing. Medlda: maximum margin supervised topic models for regression and classification. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1257–1264. ACM, 2009.