Alex Jian Zheng

Proposal

Project

Task

Sentiment analysis.

Data

Restaurant reviews, like the Yelp dataset sample in Chapter 3, but in **Chinese**. Now I have three options of datasets. They all contain the variables of "comment" and "rating".

1. https://github.com/panluoluo/crawler-analysis/blob/master/%E4%B8%AD%E6%96%87%E6%96%87%E6%9C%AC%E6%83%85%E6%84%9F%E5%88%86%E6%9E%90/data1.csv
2. https://www.kesci.com/mw/dataset/5e946de7e7ec38002d02d533
3. https://tianchi.aliyun.com/dataset/dataDetail?dataId=4366

Baseline

For the 1st dataset, there is two baselines available, from https://zhuanlan.zhihu.com/p/60723550. The first one is using the "snownlp" package, and the accuracy is 0.763. The second one is using naïve bayes, and the accuracy is 0.899. I am not sure if I can do better than the second one. I have not checked existing baselines for the 2nd and 3rd datasets.

Team

Myself (and maybe some friends who would like to review my work).

Techniques

I plan to try logistic regression, multilayer perceptron, convolutional neural network, and other techniques that I will learn in the course. I will use PyTorch.

Timeline

Date	**Progress**
Oct-28	Proposal
Nov-2	Go through the Yelp example
Nov-9	Take care of the specialty of Chinese
Nov-16	Fist deliverable. Finish the logistic regression.
Nov-23	Finish MLP & CNN
Nov-30	Try different hyperparameters
Dec-7	Final presentation
Dec-11	Project writeup