Topics in Quantitative Finance: Machine Learning for Finance

Final Project Proposal

Wang Yu　1601213621

Project definition:

This project will use a data set consisting of Chinese hotel reviews from www.ctrip.com to build a sentiment classifier that classifies a review as positive or negative.

Dataset:

http://www.datatang.com/data/11936

ChnSentiCorp-Htl-ba-2000: balanced corpus，positive(1000 reviews) /negative(1000 reviews)

ChnSentiCorp-Htl-ba-4000: balanced corpus，positive(2000 reviews) /negative（2000 reviews）

ChnSentiCorp-Htl-ba-6000: balanced corpus，positive(3000 reviews) /negative（3000 reviews）

Step:

1. Chinese Natural Language Processing. The reviews are segmented to words and converted to a sequence of part-of-speech tags.（stanford parser for python）
2. Build a vocabulary of unique words and create a bag-of-words feature representation using feature vector for each review.
3. The project will mainly use SVM and Decision tree to do the classify.
4. Compare and analyze the results of different methods
5. Tune parameters to improve the methods.

Reference:

[1] 苏金树，张博锋，徐昕. 基于机器学习的文本分类技术研究进展[J]. 软件学报, 2006, 17(9):1848-1859.

[2] Jon D Mcauliffe and David M Blei. Supervised topic models. In Advances in neural information processing systems, pages 121–128, 2008.