Text Mining System for Predicting Market Response to News

Proposed Supervisors:
Luis Torgo (Econ, UP) and
José João Almeida (DI, UM)

Kewwords: information extraction in texts, predicting markets, machine learning, natural language processing

Introduction

It is assumed that financial markets are influenced by news. Prices of shares, options, etc react to news events which may have an effect the underlying item which is represented by the share or option etc.

For example, poor weather during harvest season may effect the future price of bread. The market will react by increasing the price of a future for the delivery of wheat at a fixed time in the future and possibly the share price of supermarkets and bakers.

An automated text mining / news monitoring application may be able to detect or infer the likely effect of a news event upon the specific part of the market or the market in general.

The intention of this PhD is to investigate if such systems are viable and can whether they can be used in a real world situation.

Proposal

The aim of this PhD is to extract and analyze appropriate news stories and evaluate their potential effect on the relevant market.

The PhD will attempt to use the work of Stevenson and Yangarbe to identify the requisite phrases or n-grams within the news story and the work of Fernando Pereira to provide the linguistic analysis of the story itself.

Financial markets are complex and in the first instance it may be unreasonable to analyze the effect of news on the markets. It is proposed that the PhD will initially address the effect of news on prediction markets. This is because prediction markets are simpler than traditional markets and it should be simpler to identify the effect of news. It may be possible to use traditional financial analysis to gauge when and if a news story has an effect on the prediction market (for example, through time series analysis). The rules from the prediction markets may then be applied to the more complex financial markets.

The final application will have two stages, which can be broken down to training and operational. The training phase will attempt to derive rules from which the operational stage can either make an automated trading decision or provide information to a trader. It is proposed that the training phase will use a corpus of information which will be derived from respectable journals (for example, The Financial Times) which are freely available on the Internet.