Project Proposal

Problem

News drive stock market. I'd like to use Natural Language Processing (NLP) to extract signal from the news and predict stock shares prices. This is an interesting problem, because:

1. Amount of data is large–every day news are generated by thousands
2. It's hard to clean and to extract valuable information from the unstructured text
3. A lot of people try to predict shares pricing daily: from analysts to non-professional citizens

Data

I'm going to scrape data from the news outlets.

Solution

Predict stock prices (**regression problem**) per stock ticker using publicly available data: news, financial reports, social media. The data pipeline:

1. Scrape
2. Mingle (stats, plots, etc) and clean (leave valuable data only)
3. Extract features. E.g.:
    a. Detect company (entity)
    b. Date
    c. Location
    d. Sentiment
4. Build supervised learning model(s). Use news text and extracted features as predictors (X) and share price as predicted variable (Y).

Deliverable

A web app that would take a company name (ticker) from the SP500 list and graph prediction for its shares price.

Resources

As at this stage I'm not planning on using Deep Learning the projected resources are:

1. CPU: a laptop should be enough, any cloud platform will be enough
2. Memory: the U.S. news texts for one year is ~500MB, so a few GB for storage and regular laptop/cloud VPS memory should suffice
3. GPU: may speed up processing, will determine at modelling stage

Repository

https://github.com/yegorkryukov/stockai