Stonks: Project Description

• What is your theme/topic/goal/issue to be tackled - why is it important to you?
My topic is market prediction. This is a fundamentally simple and reasonable goal. However, depending on the data that I harvest from news sources for this goal it could have either a reasonable outcome that is highly helpful or an absurd, chaotic, misrepresentative one. In the first case this project would be more of a market analysis tool and in the other case it would be an absurd experimental data driven art piece detailing the chaotic nature of market reporting and decision making in our very fast paced world.

Markets have a treasure trove of data associated with them. They have existed for hundreds of years. They have millions, billions or trillions of data points created constantly, and they are still being updated at every millisecond. Furthermore, as they continue to be updated, they have more data generated because they expand exponentially. I think this data theme offers a lot of avenues for interesting and revealing outcomes.

This is important to me because analysis and advice regarding markets is usually very misleading, biased or just wrong. This is because markets are one of the most difficult things to predict. I want to attempt to solve this problem or at least show how difficult it is to solve this problem.

• What form will your project evolve into - who is your audience?
This project would evolve into a news consolidation platform. It would interest either anyone who is interested in market analysis or anyone who is interested in real time crowdsourcing and data driven art. Its audience is people who would normally scour many different news sources looking for a piece of information that would fit their predetermined belief about the direction of a stock. It would act as a tool that would replace the apophenia associated with looking for patterns in a collection of news data. It would be a scientific tool to screen out this irrationality and lead to possibly a better decision.

Everyone who participates in a market wants to predict them. Specifically, whether this value of its constituents will increase or decrease. While this is a very simplistic goal, since markets are almost completely stochastic, or random, it becomes an almost impossible problem to solve. This is the problem I will attempt to solve for my final project. I think there are quite a lot of people who would be interested in this goal. Markets are huge focal points for many jobs and careers revolve around them. My audience would be anyone who either has an interest or who works in one of these fields. My audience would be anyone who is interested in data consolidation would also be interested in this, but that would be much fewer people, I think…

• Discuss how each of the two readings listed above have inspired/motivated your current choices with regards to the project.

Apophenia, a concept from Hito Steyerl's writing, is a major problem when trying to analyze market data. It is so easy to start to see patterns that are just figments of one's imagination even when there are real takeaways to be had. My project would function as a scientific tool to replace apophenia.

Stereyl also foregrounds the concepts of signal and noise. In my project the noise would be irrelevant news regarding market movements and the signal would be news that is relevant and actually operates as predictors. Even in retrospect, it's impossible to know, on an individual basis, which of the millions of headlines predicted a movement in the market, let alone, any one of the market's constituents. However, if given a large enough dataset of daily market movements and daily news collections, I think it might be possible to screen out the noise and find which keywords, topics, or news sources are the most reliable in predicting the market and therefor constitute the signal. Furthermore, I can download archives of the history of news and markets and look at how news predicted markets in the past.

Also, Mimi Onuoha states that, "As we collect more data, we prioritize things that fit patterns of collection." Applied to my dataset, time is prioritized when collecting market data. Every datapoint has an associated a time tamp and sometimes this becomes misleading. This is because time is largely arbitrary when it comes to movements in prices. Usually price changes happen because of events, not times. So, time becomes an almost useless datapoint even though it is the most common. This is also why I am prioritizing collecting data over time. The more data I have, the more accurate my machine prediction is.

• What medium(s) do you intend to use and why?

I intend to use the medium of Data. This data would be displayed on a website so that may be considered the medium instead. I want the final product to be an accessible website that updates market predictions in real time. And/or a feed of consolidated news computationally analyzed.

It's hard to envision this project's front-end result. Despite how complicated the back end could become, the front-end would really be no more than a one-dimensional signal advising the observer to buy or sell the market as a whole.

So, along with this simplistic, capitalist application, I also want to create a front-end for the news aggregation aspect of the project. Maybe an apophenic newspaper. Or machine-readable news. Or rather, a machine digested news. One that aggregates hundreds or even thousands of news sources and spits out the most euphoric/pessimistic stories at the top and most mundane ones at the bottom. Or even a newspaper that combines stories on a given topic into one article resulting in a jumbled, schizophrenic diatribe of sorts.

• What is your data: where will you get it; will it be collected - how and why?

Half of my data will be company valuation data. The other half will be news sources. The first half of this data will be a collection of company stock tickers and their associated performances, sourced from company tracking APIs such as NPM google-finance or NPM yahoo-fin. The other half will be data from news sources. Either using

a web crawler and scraper in combination (such as NPM cheerio and NPM Crawler), or a news API (such as 'NewsAPI') that provides news articles in JSON format. The project will be in connecting the two of these datasets in a meaningful way using sentiment analysis.

Another important note is that I plan to analyze news and market data history in order to inform my predictions. If I look at the history of how certain news did or didn't act as a predictor over the course of history, I can use this information to screen for current news that will be a signal for the market. Historical market data is very easily available, news archives are more difficult to find because they may not go as far back in history, but they are still available as well.

- At a very high level: what are the algorithm(s) that will be used and implemented to achieve your intentions?

One algorithm that I know I will be using at this point is Sentiment analysis. For this I would use NPM sentiment, the same library that was shown in class. This would allow me to analyze whether a piece of news has a positive or negative sentiment. However, the other datapoint I would need to collect that sentiment analysis would not necessarily provide is which part of the market the news article could be a predictor of - whether it is a general market predictor such as a news article about a pandemic, political upheaval, or a market crash itself. OR whether it is a news article that relates to a sector of the market or even a specific stock. For this I would need to write or find an algorithm that looks for keywords within the headline or possibly even within the content of the article that would classify it as a tech related article, healthcare, automotive, etc. I wonder if there is a way to computationally analyze a news article title to detect whether it falls within a set of predetermined categories. OR whether there are enough articles daily in the world that contain the specific name of a company. Since this could end up being very coding intensive, I will start with creating an algorithm that finds articles that would be predictors of the market in general and not just one part of it.

Two other algorithms that would be fundamental are a news data endpoint and a market data endpoint. The news data endpoint could either come in the form of a web crawler combined with a web scraper (NPM cheerio, NPM crawler for example). These would allow me to access simple strings containing news headlines. The difficult part here is that then I would need to extend my crawler to be able to 'click' on the hotlinks of every news website headline per news source in order to download the content of the article it is associated with. This becomes very coding intensive very quickly. I already tried to get into programming this, and part way through I realized that there are news APIs, such as 'NewsApi,' that offer streams of news data already in JSON format. This API has options to stream from certain regions and only from news that contain certain keywords/phrases.

The Open Data handbook is a project that provides inspiration for my project. In it is described open data as data that is *can be freely used, re-used and redistributed by*

anyone - subject only, at most, to the requirement to attribute and sharealike. I am grateful that open news archives exist. Otherwise, I would have to pay in order to access this history. The open data handbook also specifies that the in order to be open, the data must be available in full. This is very important for my project because, for my algorithms to be as accurate as possible, they must have a full scope of data accessible. This is also why I plan to make my application open as well. So that just as I can build on the shoulders of this open resource others will be able to build on mine.

In the why section, the open data handbook outlines government data as being especially valuable because of its centrality of collection and quantity. Government data is usually open as legislated by law. In my case, since I am building a data application that is intended to analyze corporations, my case is very different. The kind of data that I require is almost as far away from government data as possible. As far away as corporations are from the government I suppose. There are probably various closed data sets that would be far more predictive of the market than news sources. However, since this data is highly valuable, I will not have an affordable route to accessing them.

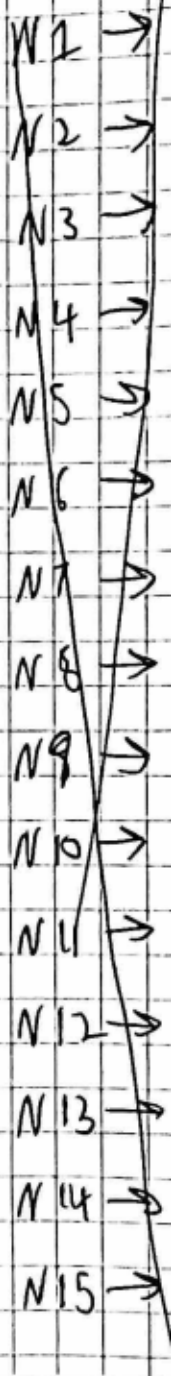
Every Stock	% Change	News Source most correlated with
AAA	0.3 %	CNBC
AAB	0.1 %	CNN
AAC	1.0 %	New York times
AAD	1.1 %	BBC
AAE	9 %	Guardian
AAF	0.8 %	Los Angeles times
AAG	2 %	BBC
AAH	3 %	
AAI	5 %	Washington Post
AAJ	7 %	Daily Mail
AAK	1 %	CBS News
AAL	2 %	ABC
AAM	3 %	Marketwatch
AAN	4 %	CNBC - Business
AAO	5 %	CNBC - politics
AAP	6 %	CNBC - Tech

IMPORTANT

This data would be saved every day, so if a stock correlates with a specific news source often, that would be reflect

NEWS Source (Includes headline news as well as market centric news and subcategories)

		Sentiment
CNBC	- Business	100% (Positive) - 0% (Negative)
	- Tech	100% (Positive) - 0% (Negative)
	- Politics	100% (Positive) - 0% (Negative)
MarketWatch		100% (Positive) - 0% (Negative)
CNN		" "
NBC		" "
CBS News		" "
New York Times		" "
Daily Mail		" "
Washington Post		" "
BBC		" "
Guardian		" "
Los Angeles Times		" "
New York Post		" "
ETC.		" "



Alternatively, Instead of checking correlations with specific stocks, ~~or~~ Indecies or sectors could be checked instead. This could be more effective because News Sources Report with regards to the entire Nation or the world, not specific companies. Also there are only a few Indecies: Dow Jones, S&P 500, Nazdaq 100, S&P/TSX, ETC. AND sectors could be linked to News source Sub-categories or ~~for~~ News Sources that are focused on specific sectors like Tech, Investing, Social Media, Health, ETC. In this case, we would situate the NEWS DATA First/primarily, and then see ~~which~~ whether Indices correlate with it or which sectors correlate with which news sources

News Source	General sentiment	Top keywords	Running correlation with Index
CNBC	0 - 1	"Job"	0 - 100%
Marketwatch	0 - 1	"Bank"	0 - 100%
CNN	0 - 1	"Car"	0 - 100%
NBC	0 - 1	"Bitcoin"	0 - 100%
CBS	0 - 1	"Fed"	0 - 100%
ABC	0 - 1	"Job"	0 - 100%
BBC	0 - 1	"Amazon"	0 - 100%
New York Times	0 - 1	"Rocket"	0 - 100%
Los Angeles Times	0 - 1	"Satellite"	0 - 100%
Washington Post	0 - 1	"Carbon"	0 - 100%
Guardian	0 - 1	"Oil"	0 - 100%
New York post	0 - 1	"Kardashian"	0 - 100%
Huffington post	0 - 1	etc	0 - 100%
Fox News	0 - 1	etc	0 - 100%
USA Today	0 - 1	etc	0 - 100%
politico	0 - 1	etc	0 - 100%
Yahoo news	0 - 1	etc	0 - 100%
NPR News	0 - 1	etc	0 - 100%
Breitbart	0 - 1	etc	0 - 100%
News Week	0 - 1	etc	0 - 100%
Chicago Tribune	0 - 1	etc	0 - 100%
Salon	0 - 1	etc	0 - 100%
Boston.com	0 - 1	etc	0 - 100%
The Seattle Times	0 - 1	etc	0 - 100%

Note

Keeping track of top keywords (common words) could be useful to manually check on specific stocks by seeing if there are words that are appearing on many news sources all at once and whether they could relate to a specific type of stock and its performance

Alternatively keywords could be the focus. Top keywords would be words that occur across different news sources. They would indicate significant events in the world.

Significant keywords	Stock ticker/Index/Sector	Correlation
Job	AAA	0-1
Bank	AAE	0-1
Car	AAB	0-1
BitCoin	Tech	0-1
Fed	DowJones	0-1
Job	Health	0-1
Amazon	S&P 500	0-1
Rocket	AAZ	0-1
Satellite	Nazdaq 100	0-1
Carbon	AAT	0-1
Oil		
Television		