2018-06-20
Jefferson Richards

Cortext.io : The Re-Search Engine
Our Nascent Aim

Purpose: Our intention in writing this document is to model an alternative approach for generating portfolios of assets. Our new method involves studying fundamental concepts in the news to determine industry risks and market risks and then to utilize subjects in the news to determine which companies, organizations, or people to pursue fundamental analysis with. Because 2019 is the year of international languages, this paper is relevant.

Preliminary: The traditional method for determining which assets to include in a portfolio is to perform top-down analysis. This process involves: economic analysis, industry analysis, and fundamental analysis. Economic analysis involves looking at political and legal policies. Specifically, investors pay attention to government spending. A keen investor will determine statistics regarding the masses and forecast where cash will flow. More modern methods involve behavioral analysis. We are keen on the idea that the laws of physics are much better reasons for why the masses act the way they do. Both tracking government policy and intuiting “laws of nature” are naive approaches to economic analysis because we are now in a global economy. National policy and demographic statistics do not have such a foundation in economic trends as they used to have before the advent of the internet. Also, foreign investment (increased/decreased?) in U.S. affairs creates a global interdependence. This can be directly deduced from the effects of the U.S. Recession of 2008 on the world financial markets.

A better approach at determining economic policy and laws of human nature is to use a concept and subject frequency based approach at analyzing the news. Global financial news repeatedly publishes articles on the concepts and subjects which are pivotal to the global economy - this can be taken as a fuzzy *a priori* axiom.

We have developed a novel approach at parsing and categorizing Natural Language with the purpose of isolating subjects and concepts with a universal algorithm. The algorithm has potential to apply to 26 languages but has only been implemented in English and Spanish. The algorithm leverages Wierzbicka’s Semantic Primes and attempts to ride the grammatical line between syntax and semantics.

> Once documents have been reduced to sentences, by parsing by period, sentences are reduced to words, by parsing by the space character. This method is standard to all of English and should not be considered novel. While words are passing through the data pipeline, we flag the words that are synonyms of the semantic primes. We do this in preparation for appending semantic metadata to the sentences. If we were to report the sentences, we would add the primes to the attributes of the html tags:
>
> We propose that primes are added to schema.org so that universal primes can be identified in all natural languages. By doing so, universal meaning can be easily extracted from text and be used to gain back some of the meaning lost by traditional translations. THe english language, especially, needs to go through a contractionary period as the recent age of language and branding has degraded the English language. Because 2019 is the year of international languages, this paper is relevant.

Once primes and capitalized words are identified, they are actually used to parse the remaining text. The remaining text is considered to be “concepts” because they are phenomena of language that have not yet been classified. We actually consider capitalized words to also be a phenomenon of language because, although they are known to be significant subjects, they are without classification. For short, we call subjects and concepts: “Phen.”

> The highest frequency phen in global financial news are assumed to be primary indicators as to the state of global economies. Of course, the representative power of phen is limited by the source of information. We cut through the bullshit and stuck with Reuters and Associated Press because they are Newswires. We placed more trust in Reuters than AP because AP had a “sensational” aspect that seemed to accentuate anomalies in the United States (tiger ate 4 babies type of shit). Our scraping cycle was around 12AM pacific time every day. We pulled around 250 articles a day. We now have 400,000 articles and 1,000,000 sentences. One of our perspectives is that the quality control of AP and Reuters allowed for us to work under the notion that each sentence is complete in necessary qualities for the layman to understand an event. In other words, we considered each sentence to be a headline. Although we drew a distinction between headlines and inlines, they ultimately contain a proper amount of information to deem: Sentence:event :: word:trope
>
> As Einstein viewed the speed of light to be the foundation which time is relative to, the speed of our daily cycle of news was the baseline which events were relative to. We did not try to keep up with the speed of the modern news cycle. We pushed through with our one day cycle with the rationalization that one day old news would have a slightly less biased average meaning than minute old news. This made election time exciting (because we had to wait a day).

Since the election occurred during the algorithm writing process, we of course applied the algorithm to the presidential candidates’ speeches. We found that Trump had Existence based natural language and Hillary had Expression based language. This was not used to project as to who would win the election; it was more used to predict the nature of news for the winner’s time in office. If Trump one, news would be existence based. If Hillary won, news would be expression based. The speeches were assumed to set precedence for the terms to come. We also assumed that the state of primal semantic lean is cyclical in nature. Now that Trump is President, Is the news existence based? There is room for analysis here now that so many presidential speeches and daily news are on record. In an expression based cycle, we argue that the temporary precedence set is that he says/she says (hearsay) is an acceptable form of argument. In an existence based cycle, the temporary precedence set is that arguing something is fake is acceptable. One could imagine that if Hillary won, the buffer against false news would have been called “hearsay news” instead of the current title “fake news.” Either allows the sufferer of poor publicity to claim plausible deniability and is therefore a necessary part of the news domain/paradigm.

2018-06-20
Jefferson Richards

By stringing together associated phen, we created associative webs. These webs were pseudo industries - treated as industries but not traditional. Webs set scope for determining which companies to use to determine risks. [Created method for finding indices that are associated with high freq phen of associative web]

> We then extracted Risk Factors from 10Ks of top companies in industry. We applied Cortext to the risk factors but ultimately found that risk factors must be read manually to fully grasp them. Risk factors can be categorized into systematic and unsystematic. Systematic risks are well understood. Unsystematic risk were the difference between the total risks and the known systematic risks (in risk library). Incorporated risks into news scraping pipeline to trigger decision algorithms. Ultimately, the goal was to identify, classify, evaluate, and treat risks. And our personal developments allow for notification system that reports to those who are responsible or accountable. Used ISO 31000 as a framework. Reverse engineered to create a database for the risk management process. A major method of determining actual strings to search for in the news was to synonymize risk categories (found in ISO 31000). Then combined phen, primes, timestamp, risk triggers into notation that made the management of events much easier. One of the goals was to create an operational notation for marking up events - say, in the newspaper. Looks like the following:
>
> {{timestamp},{subjects},{concepts},{primes},{risks}}
>
> The mental model was that a layman could markup a newspaper on a daily basis to objectify and organize events. A “pencil-and-paper” method is just as valuable (if not more valuable) than all of the computational methods. Enabling daily event management is a step forward in textual analysis.

To handle all of the news being stored each day, we had to create an application that could access the database. We wrote an Android app and an iPhone app. They utilized search capability and daily article metadata reporting. We also created an application for generating event notations. This was intended to improve reporting efficiency and event processing after the events were uploaded to the database. Organizing event attributes in this manner shifted our mindsets on how to regard events.

2018-06-20
Jefferson Richards

The Capital Asset Pricing Model is a model that helps a person determine the required rate of return for an investment. The equation for the model is as follows:

$$E(R_i) - R_f = \beta_i(E(R_m) - R_f)$$

This equation can be plotted to form a Security Market line - with beta on the horizontal axis and asset returns on the vertical axis.

Neel,

I have the piece of information that I needed to start a research paper. What do you think about publishing a paper with our NLP findings?

I needed a model that depicted our final purpose for pursuing what we have pursuing with Sensutec and all. The model that I believe is the resting place of our research is the Capital Asset Pricing Model. There is underlying implications in the fact that the CAPM led to its founders winning the nobel prize in economics.

$$E(R_i) - R_f = \beta_i(E(R_m) - R_f)$$

I view the risk free rate as the baseline, the Beta as the measure of systematic risk, and the premium as the measure of unsystematic risk. This is exactly what we needed. I think you could explain this model to me in more detail.

I have not exactly worked it all out but my thoughts are that we can introduce a new top-down analysis for determining diversified asset portfolios.

1. We created a universal algorithm for parsing and categorizing concepts and subjects (has potential for 26 languages)
2. Highest frequency concepts and subjects in global financial news can be assumed to be indicators of the most important economic policies, people, companies, government spending,... etc (First axiom)
3. Looking at the subjects and concepts of articles that contain the most frequent concepts and subjects (web) can clustered person, places, concepts, and things to formulate "industries."
4. Extracting and finding trends in risk factors from the industry leaders will help to determine systematic and unsystematic risk (unsystematic can be diversified). Can even apply textual analysis and ISO 31000 to this (risk management standard)
5. Neel adds in some novel approaches at valuation of the firms that have been isolated in this process.

2018-06-20
Jefferson Richards

The major value add will need to be better described. The major difference between this approach and traditional approaches is that:

1. Economic policy and trends in "human nature" of a nation are no longer the foundation of economic policy of that nation. We are now a global economy (tie in 2008 Financial crisis and that all economies are interdependent)
    1. Our approach: Trends in global financial news are better indicators of economic policy and pivotal companies/people because high frequency of subjects and concepts in AP and Reuters correlated with significant economic influence (hash out and prove. The idea here is to justify using frequency of subjects and concepts (direct string frequencies) because we will later argue (in the sentiment portion) that using frequency of strings is a very naive approach at textual analysis)
2. Obviously our universal algorithm for extracting phen is awesome by itself
3. We have (actually new branding for our associative web phase) a flexible and scalable method of classifying industries (associative web of subjects and concepts). This is an alternative to the traditional industry classifications.
4. I have confidence that Neel has novel approaches at valuation and fundamental analysis once the assets have been determined (financial statements, etc.).

This is just a quick outline and I am missing a lot. The mathematics I developed behind primes as well as the neural networks I generated for cleaning data after it has been processed are also value adds to the paper.

What do you think? Let's just get published. It will be no problem. I already have all of the documentation.

I know that your primary response is that we need to prove that the approach generates cash positive portfolios. I don't know what to do with this. I want to write the paper and prove it as we go along because I personally need to bring everything together at the same time. I am confident that compiling all of the research and proof of our findings will help with finalizing the process. Basically, we should generate portfolios at the end of writing this paper. Again, I am losing track of the steps we took prior to now and think that I need to go back over those steps before we codify the exact methodology for generating portfolios.

From,
Jefferson

CAPM Assumptions:

All investors:

1. Aim to maximize economic utilities (Asset quantities are given and fixed).
2. Are rational and risk-averse.
3. Are broadly diversified across a range of investments.
4. Are price takers, i.e., they cannot influence prices.
5. Can lend and borrow unlimited amounts under the risk free rate of interest.
6. Trade without transaction or taxation costs.
7. Deal with securities that are all highly divisible into small parcels (All assets are perfectly divisible and liquid).
8. Have homogeneous expectations.
9. Assume all information is available at the same time to all investors.

2018-06-20
Jefferson Richards

Problems with Model:

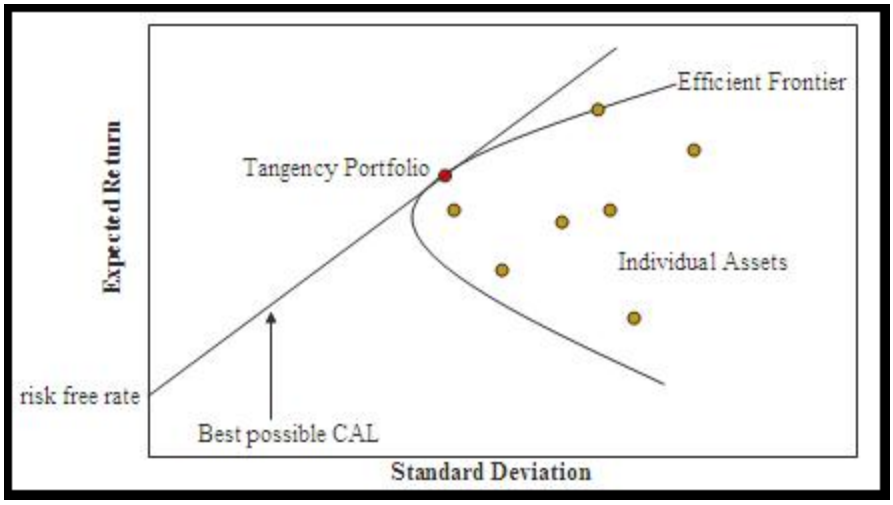
In their 2004 review, Fama and French argue that "the failure of the CAPM in empirical tests implies that most applications of the model are invalid".[3]

- The traditional CAPM using historical data as the inputs to solve for a future return of asset i. However, the history may not be sufficient to use for predicting the future and modern CAPM approaches have used betas that rely on future risk estimates.[8]
- Most practitioners and academics agree that risk is of a varying nature (non-constant). A critique of the traditional CAPM is that the risk measured used remains constant(non-varying beta). Recent research has empirically tested time-varying betas to improve the forecast accuracy of the CAPM.[9]
- The model assumes that the variance of returns is an adequate measurement of risk. This would be implied by the assumption that returns are normally distributed, or indeed are distributed in any two-parameter way, but for general return distributions other risk measures (like coherent risk measures) will reflect the active and potential shareholders' preferences more adequately. Indeed, risk in financial investments is not variance in itself, rather it is the probability of losing: it is asymmetric in nature. Barclays Wealth have published some research on asset allocation with non-normal returns which shows that investors with very low risk tolerances should hold more cash than CAPM suggests.[10]
- The model assumes that all active and potential shareholders have access to the same information and agree about the risk and expected return of all assets (homogeneous expectations assumption).[*citation needed*]
- The model assumes that the probability beliefs of active and potential shareholders match the true distribution of returns. A different possibility is that active and potential shareholders' expectations are biased, causing market prices to be informationally inefficient. This possibility is studied in the field of behavioral finance, which uses psychological assumptions to provide alternatives to the CAPM such as the overconfidence-based asset pricing model of Kent Daniel, David Hirshleifer, and Avanidhar Subrahmanyam(2001).[11]
- The model does not appear to adequately explain the variation in stock returns. Empirical studies show that low beta stocks may offer higher returns than the model would predict. Some data to this effect was presented as early as a 1969 conference in Buffalo, New York in a paper by Fischer Black, Michael Jensen, and Myron Scholes. Either that fact is itself rational (which saves the efficient-market hypothesis but makes CAPM wrong), or it is irrational (which saves CAPM, but makes the EMH wrong – indeed, this possibility makes volatility arbitrage a strategy for reliably beating the market).[12][13]
- The model assumes that given a certain expected return, active and potential shareholders will prefer lower risk (lower variance) to higher risk and conversely given a certain level of risk will prefer higher returns to lower ones. It does not allow for active

and potential shareholders who will accept lower returns for higher risk. Casino gamblers pay to take on more risk, and it is possible that some stock traders will pay for risk as well.[*citation needed*]

- The model assumes that there are no taxes or transaction costs, although this assumption may be relaxed with more complicated versions of the model.[14]
- The market portfolio consists of all assets in all markets, where each asset is weighted by its market capitalization. This assumes no preference between markets and assets for individual active and potential shareholders, and that active and potential shareholders choose assets solely as a function of their risk-return profile. It also assumes that all assets are infinitely divisible as to the amount which may be held or transacted.[*citation needed*]
- The market portfolio should in theory include all types of assets that are held by anyone as an investment (including works of art, real estate, human capital...) In practice, such a market portfolio is unobservable and people usually substitute a stock index as a proxy for the true market portfolio. Unfortunately, it has been shown that this substitution is not innocuous and can lead to false inferences as to the validity of the CAPM, and it has been said that due to the inobservability of the true market portfolio, the CAPM might not be empirically testable. This was presented in greater depth in a paper by Richard Roll in 1977, and is generally referred to as Roll's critique.[15]
- The model assumes economic agents optimise over a short-term horizon, and in fact investors with longer-term outlooks would optimally choose long-term inflation-linked bonds instead of short-term rates as this would be more risk-free asset to such an agent.[16][17]
- The model assumes just two dates, so that there is no opportunity to consume and rebalance portfolios repeatedly over time. The basic insights of the model are extended and generalized in the intertemporal CAPM (ICAPM) of Robert Merton,[18] and the consumption CAPM (CCAPM) of Douglas Breeden and Mark Rubinstein.[19]
- CAPM assumes that all active and potential shareholders will consider all of their assets and optimize one portfolio. This is in sharp contradiction with portfolios that are held by individual shareholders: humans tend to have fragmented portfolios or, rather, multiple portfolios: for each goal one portfolio — see behavioral portfolio theory[20] and Maslowian portfolio theory.[21]
- Empirical tests show market anomalies like the size and value effect that cannot be explained by the CAPM.[22] For details see the Fama–French three-factor model.[23]

Efficiency Frontier:



Behavior analysis is a naive approach as compared to laws of physics