AI Final Report

Hong Kong University of Science and Technology

Hua Neng 20549375
Cheng Yu 20568266
Wang Xuan 20549507
Bi Huarui 20568008
Wang Sunan 20541361

Abstract

In this report, we first introduce the process and output of our research on **Yewno**, an AI company. After that, two reflection on an AI article and an finance one are presented. Then several brief outlooks are put forward for future's study. Lastly, individual's contribution are noted.

1 Progress and learning from group project

1.1 Progress

Yewno is a company that can be regarded as a eveloper of dynamically evolving knowledge graphs that provide inference strength across concepts, events and themes using computational linguistics and deep learning.

And its mission is extracting knowledge from an overwhelming quantity of unstructured and structured data. Using technology helps to overcome the Information Overload problem and to research and to understand the world in a more natural manner.

We first collect the information of the company's backgrond which covers several fields: Finance,Education,Publishing,Government and so on.Because of his vast field of involvement and type of an IT company, we feel that their network information must be very developed, so we decided to take four ways to get information about the company.And the four ways are:

- **Social Media**-Choose a public website about the selected company, such as youtube,and download all comments by Python
- **Recruiment Information**-Crawling and tabulating recruitment information published by target company over time
- **Web Crawler**-Crawling and tabulating news re- garding target company over time, and mapping articles to topics of interest to reveal trends.
- **API**-Get the current business and direction they are trying to focus on. This kind of method is straight forward but may be a little lag, because the code is published when the company is already make the decisions.

1.2 Learning from group project

1.2.1 Social Media

The second approach is to refer to the social platform for relative information. We have studied several websites including YouTube, Facebook and Twitter and it turns out that

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

yewno is relatively active on Twitter. The main content of their tweets include three categories. The first one is the campaigns they have launched, including multilingual, do you know competition and also the corporation with local government.

Also they seem to attempt to attract more attention by the interaction with other accounts with a larger number of followers, such as @LibraryJournal and @PNASNews, with 203k and 105k followers, respectively. But still there shows little changes in their followers amount.

So they are paying attention to recently is that they set up a new branch called Yewno Finance. Basically the two companies do quite similar things, except that the finance one mainly focus on the finance-relative knowledge. This is worth thinking because the narrowed range of their business may lead to loss of customers but at the meantime, since finance industry is so profitable, and most people barely have any knowledge in this field, it might have a very promising future.

1.2.2 Recruiment Information

Human resources are the primary resources of a technology company. Thats why we choose to analyze the recruitment information about Yewno. Whats more, as a start-up, Yewno may frequently post job vacancies because of business expansion. Therefore, we can get updating information. Besides, all recruitment information is public and can be easily found.

We started from the companys website. Yewno posts three jobs on its website: Quantitative Finance Analyst, Full Stack Software Engineer and Alternative Data Research Analyst. They also upload a pdf file as job description about every job. We can get the most detailed information about the key responsibilities and requirements about jobs on the company website. Among the three vacancies, two mention financial data and services.

We also searched Yewno on several job boards, like Linkedin, Glassdoor, Indeed, Monster, Career Builder, Simply Hired, Dice. Yewno just post its jobs on 4 websites: Linkedin, Glassdoor, Indeed and Simply Hired. And then we will analyze searching results on the 4 websites respectively.

- **Linkedin**
  Find two jobs. One of them, DevOps Engineer, is not found on companys website. And this job board provide us one more information about jobs: the publish date. Therefore, we can know the more accurate deduction about the company business progress.
- **Glassdoor**
  Find one job Full Stack Software Engineer, which we have already found in company website. But job description on this website includes salary. Besides, we can get rating from employees about this company and their comments. The rating of Yewno is just 1.6 which is a very low grade. Comments are mainly negative, and employees complain about bad management and high pressure environment about this company.

- **Indeed**
  Jobs posted on this website are completely same as that on company website. But we noticed that the benefits of jobs are little different. The Engineer position, whose work office is at Redwood, enjoys more benefits than research analysts, whose work office is at New York.
- **Simply Hired** Jobs posted on this website are completely same as that on company website. But recruitment information on this website disclose salary of every job. Salary of Alternative Data Research Analyst is $73,000 - $99,000 per year, which of Quantitative Finance Analyst is $110,000 - $150,000 per year, and of Full Stack Software Engineer is $120,000 - $160,000. Yewno pays engineer the highest salary.

Generally, hiring more sales means this company is able to expand more markets and clients, and get higher operating income. But Yewno may not achieve that, based on their hiring process.

To analyze the company, especially its human resources, and detect its newest business development in the future, a better method it to build a database about its previous and current employees, as well as its recruitment information.

Use IT technologies like Crawler to automatically grab data and update our database, and AI technology to analyze this data and get some reliable prediction combining with other information about this company.

1.2.3 Web Crawler

With the help of web crawler, we can gather huge number of data from internet, using the data we can do further analysis of the company. We have implemented three web crawler to handle three webpagessearch engine, tian Yan cha, Linkedin.

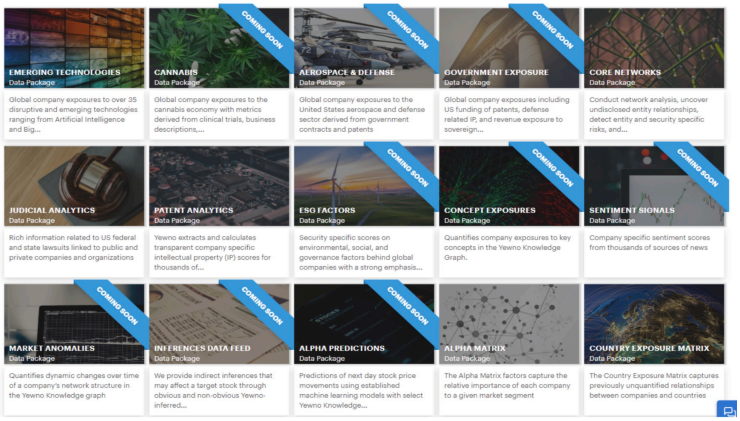
- 1) Search engine is the whole network information gathering center, we can get basicly every url about the company from search engine like google, bing and baidu. We use Scrapy a powerful python crawler skeleton to crawl the search engine data.
  In the results, we can get many urls related to yewno. However, different webpage have different web structure, it is very hard to further crawl the urls in the result from search engine crawler.
- 2) Tianyancha is a website which provides great number of company and people information, the information is professional and comprehensive. But this website is mainly serve for Chinese user.
  We chose another AI company sensetime to do analysis. Use Chrome-webdriver to manually handle the log in and verification problem.
  From the crawler result we can get the brief introduction of the company and the business scope of sensetime, the actual controller of the company, and other important staff composition.
- 3) Linkedin contain more detain information of the company employees and their resumes usually is also showed on Linkedin.
  But Linkedin includes too much personal information, so their anti-craw system is very powerful, to get the data from Linkedin we use cookies pool to pretend human visit and use Redis to record what we have crawled, and change the timestamp to pretend human.
  The working flow is first we search the company yewno in the landing page and get the employees list, then crawl the detail person webpage one by one. In the result we get the title of the person in the company, and the history working experience.
  After analysis the result, we can get the conclusion that over 70 of the employees have once worked in a quantitative financial company or artificial intelligence company before, so the personnel structure of yewno is very professional.

1.2.4 API

Yewno's Intelligent Alpha investment strategies seek to take advantage of non-obvious relationships defined by Yewno's Knowledge Graph structure, generating a unique and persistent source of alpha.

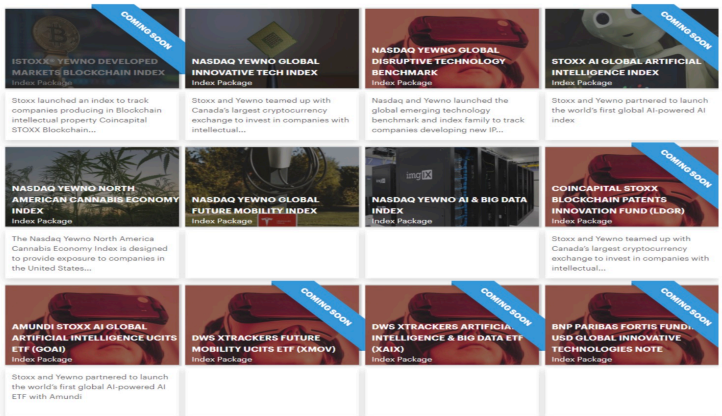
By increasing the speed of understanding, logical trades can be entered before other market participants react to market anomalies. Yewno leverages this Knowledge Arbitrage phenomenon to create multiple Intelligent Alpha Strategies.

Yewno is a leading provider in the index and ETF industry by leveraging its dynamic knowledge graph which aggregates a large volume of structured and unstructured data in order to find companies that are exposed to specific themes.



Thanks to AI technology, Yewno provides a highly scalable and cost-effective way for index and ETF providers to create alpha looking for thematic index within the ever-growing emerging megatrends. Yewno can be filtered out under specific topics. The company to build the index, so that the correlation between the internal companies built by the index is more in line with this specific theme, the strategy is more effective and more financial intuition when constructing strategies for specific topics.

Yewnos Intelligent Data Feeds provide customers in the financial services industry with access to unique insights directly from Yewnos Knowledge Graph. From quantitative managers to fundamental PMs to research analysts, Yewno offers a diverse set of data feeds that measure systemic risk, country risk, sentiment, company intellectual property and more. Navigate the product grid below for more information and contact sales if you would like to trial a data set.



Yewno has a wealth of new data that is rarely seen in traditional quantitative platforms, such as the judicial analytic. These scores can be used to analyze the historical legal exposure of companies, courts, and judges on core topics such as data privacy and antitrust. Investors can build portfolios based on legal risk exposures.

The first update is to improve the ease of use of api, making it easier to do screening, indicating that the company is relying on optimization for ordinary users, and it is easier to write strategies.

The second update is to increase the breadth of api usage so that the api can run for a longer period of time. When users want more stable and long-term data to test, the user really wants to check whether the policy is valid, indicating that the company is actually providing users.

Traditional quantitative platforms focus on volume and fundamental data analysis, but with artificial intelligence technology, yewno can process text and other data, making it possible to analyze events and emotional data, which makes data more versatile, and other traditions.

2 Reflection on articles

2.1 Finance Article

Title: Volatility Analysis of Bitcoin Price Time Series
Author: Luká Pichl, Taisei Kaizoji

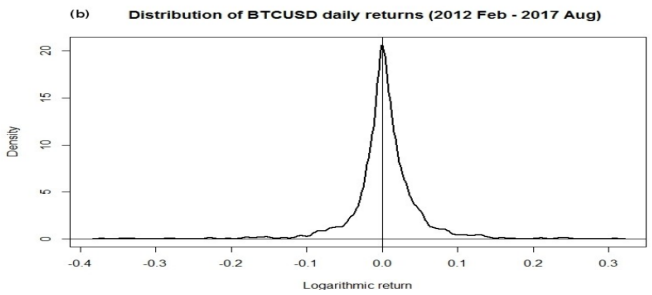
International Christian University, Osawa 3-10-2, Mitaka, Tokyo 181-8585 Japan
Received: 10 September 2017 , Accepted: 19 November 2017 , Published: 13 December 2017
Source: Quantitative Finance and Economics, 2017, 1(4): 474-485. doi: 10.3934/QFE.2017.4.474

In recent years, Bitcoin is increasingly accepted in real economy as a means of payment, and definitely become a popular but risky investment instruments. Actually, at the last semester, the project of one of our classes is to design trading strategy for cryptocurrencies including Bitcoin. Thats why we choose this essay. In this essay, authors focus on the price of Bitcoin in terms of standard currencies and their volatility from 2012 Feb to 2017 Aug, and to get results in the following 5 parts: Bitcoin price and return distribution, distribution of trading volume, Bitcoin arbitrage opportunities involving different currencies, estimation of realized volatility and prediction of log-returns using neutral network.

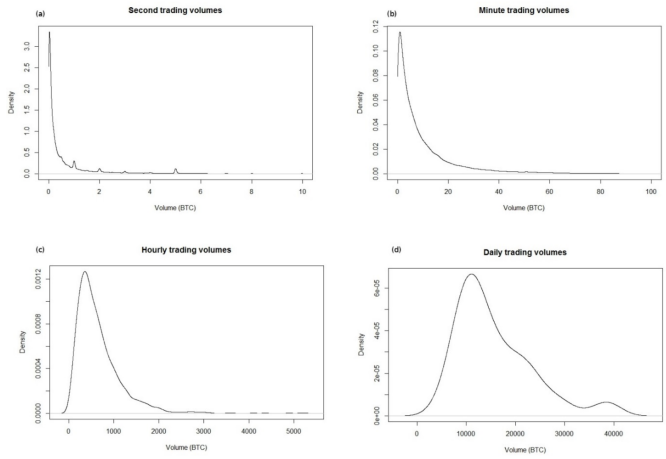
For the first part, the Bitcoin prices in terms of a standard currency CRS, i.e. the BTCCRS time series, are denoted as Bi, with the sampling frequencies of 1 second, 1 minute, 5 minutes, 1 hour and 1 day. The logarithmic return is defined as:

$$R_i = log(\frac{B_i}{B_{i-1}})$$

Bi stands for the price of Bitcoin at time step i. The distribution of the daily return for BTCUSD times series illustrate the approximate symmetry of the R-distribution.



At the second part, in order to produce distribution of Bitcoin trading volumes, which is usually unavailable from the standard high frequency data sources, authors use the application interface (API) of Kraken exchange market, collecting the last 5,000 transactions every minute and transforming them onto regular high frequency grids of 1 sec, 1min, 1 hour and 1 day. And draw figures of distribution of transaction volume in different frequencies.



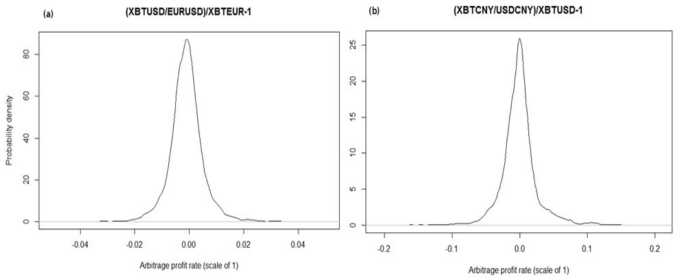
At third part, assuming Bitcoins prices in two currencies, i.e. BTCFC1 and BTCFC2. A hypothetical arbitrage transaction can be defined by buying 1 BTC in currency FC1 (expense BTCFC1), selling it in currency FC2 (revenue BTCFC2), then transforming the received cash back to currency FC1 (foreign exchange rate FC2FC1=1/FC1FC2). The profit rate (relative to the Bitcoin price BTCFC1) is then:

$$\sigma_{1,2} = \frac{-BTCFC1 + \frac{BTCFC2}{FC1FC2}}{BTCFC1} = \frac{BTCFC2}{BTCFC1}/FC1FC2 - 1$$

The direction of the transaction, FC1-FC2 is important, and it holds:

$$\sigma_{1,2} + \sigma_{2,1} \leq 0$$

The data sets used for the evaluation of arbitrage opportunities are XBTEUR, XBTUSD, XBTCNY and the foreign exchange rates EURUSD and USDCNY, all on 1 hour scale, from 2013/2/8 to 2017/4/7, and also draw distribution figures for them.



The fat tail on the right-hand-side of the distribution in figure for BTCEUR-BTCUSD shows substantial arbitrage windows. Care should be taken because:1.Transactions across markets are excluded from the arbitrage opportunity windows. 2. For most Bitcoin holders, who are not financial institutions specialized in technical trading, the most profitable strategy may simply be to hold the Bitcoin over time rather than to engage in trading. 3. There is no transaction fees considered.

The fourth part is to estimate Realized Volatility based on HARRVJ model with adjusted parameters:

$$\sqrt{RV_{i+1}} = \beta_0 + \beta_1\sqrt{RV_i} + \beta_2\sqrt{RV_{i-5}} + \beta_3\sqrt{RV_{i-10}} + \beta_4\sqrt{J_i} + \beta_5\sqrt{J_{i-5}} + \beta_6\sqrt{J_{i-10}}$$

And the fifth part, authors adopt a feed-forward neural network with two fully interconnected hidden layers, input layer of size 10, and output layer of size 1. When used as a statistical regression technique on scaled data of daily logarithmic volume, the method is capable of capturing the shape of the logarithmic return density distribution.

All works in this essay is an empirical investigation into the properties of Bitcoin markets. What is instructive for us are following:

1)It shows a workable method, which is using API on Kraken platform, to get data with the time stamp available for all recorded transactions is resolved to 0.1 millisecond.

2)Provide theoretical and empirical bounds on Bitcoin arbitrage opportunities using different standard currency pairs.

3)Ddemonstrate that a feed-forward neural network architecture is capable of learning the statistical distribution of the logarithmic return.

2.2 AI Article

Title:World Models
Author:David Ha, Jürgen Schmidhuber

Subjects: Machine Learning (cs.LG); Machine Learning (stat.ML)
DOI: 10.5281/zenodo.1207631
Cite as: arXiv:1803.10122 [cs.LG]

In this paper, Ha and Schmidhuber developed an enervative neural network model of popular reinforcement learning environments. It is a world model that can be quickly trained in an unsupervised manner and expressed in terms of the space and time of the learning environment. By inputting features extracted from the world model, a very simple strategy can be trained to solve specific tasks. According to the paper, the agent has succeeded in navigating the race track in the Car Racing task and avoiding the fireballs shot by monsters in the VizDom experiment, which used to be extremely difficult tasks if using previous methods. An interactive version of this paper is available at https://worldmodels.github.io.

There are three main components in this model: A variational auto encoder (VAE), a recurrent neural network (RNN) and a controller.

A VAE is responsible for obtaining visual information. For example, it is able to compress an image (as its input, in RGB format) into a vector, which follows a normal distribution, with 32 dimensions.

A RNN is a memory component responsible for forecasting. Based on the previous input and previous reaction, a RNN will be able to forecast what the next image will most likely look like.

A controller is responsible for giving response. Connecting the output from VAE and the hidden state of RNN, it can select the best reaction through simple neural network.

Through the combination of those three components, the model has the following achievements: Firstly, this is the first known solution to the "racing car" reinforcement learning environment. Secondly, the study shows that it is possible to train an agent to perform tasks totally depending on its own simulated space.

This model has a profound meaning in application. For example, when running computationally intensive game engines, we used to train an agent in the actual environment, and that might lead to a waste of heavy compute resources. However now we can achieve the same effect with as many times as needed inside its simulated environment.

After the publication of the paper, it was widely discussed in the AI community as a beautiful work on using neural networks in reinforcement learning and training agents in their own hallucinated worlds. Still, there might be some improvements:

Firstly, by means of changing small RNN into models with higher capacity or connecting with an external memory module, the agent might be able to explore more complicated worlds.

Secondly, instead of using time step by time step approach, as in the paper, experimenting with more general approaches might obtain a more hierarchical planning.

3 Synthesis and suggestion for further study

3.1 Improvement for corporate analysis

- API

  The reason why we choose Yewno is that the company is the only one that provide API. However, when we can get access to its database, we find that it is not a comprehensive one and the current data are not free of change. But we can use the Python packages to do some analysis. For example. analyzing whether the two stocks are competitive by passing in two stock code parameters and the start and end dates. The result is 99.83 probability for competitive relationship. Next, we can do some research on its popularity rate among the public (pay attention to the surveyed sample, the common people, or people who are used in Python or even people who are in interested in venture capital and high-tech startups).

- Recruitment information

  When our team uses the recruitment information, we just download the information from some well-known websites and the companys official website. All types of jobs, including some office staffs, haven be included without filtration. some filtration is supposed to be done in order to select the needed positions related to the core technology. If we simply use the common information, it is likely that we will reach the conclusion that the company is expanding due to the increasing employees, however, unrelated staffs may contribute to the increase. Hence, if we can separate core employees from the rest, a more precise decision can be made.

- Data from social media

  We do collect some data related to Yewno, but the amount of data is limited, which means that there is little information for us to use. Our previous goal is aimed to collect data from Youtube, Twitter and Facebook and use these data to analyze the publics attitude to Yewno. When we search some keywords regarding to the company, such as its CEO and the company name, few of texts can be seen. Therefore, the data availability limits our further analysis.

- Data from ordinary website

  It is an ordinary method, simply but un accurate. Because it is not uncommon to buy some comments on the internet, the news we can gain may be the information the company has process. If we use such data, our valuation may not be overvalued. Therefore, the biggest disadvantage of this kind of data source is biased and the biggest advantage is easy-access. We have to do a tradeoff and do more data cleaning.

- From corporate analysis method

  The corporate analysis method we use is very simple and subjective. Getting public data and then analyze by our human brain, and some imprecision can occur.

  Among data, we use some indirect information and try to induce the situation of the company. The information may be biased, and our subjective judgement is possibly involved in the analysis procedure, leading to less precise result.

  Hence, in our further study, we can try to do more objective analysis, such as inputting the data we collected and get do machine learning, and then get the results purely generated from the computers.

3.2 Further study on Artificial intellectual

It is a very practical project that we can put what we learn in our course into practice.AI and machine learning are buzz word in nowadays, people can get knowledge about AI through class and reading. But after all, it's on paper. This course gives us the opportunity to apply knowledge to our future

work while learning the basics of AI. This approach gives us a deep understanding of the meaning of AI, making it easier for us to apply AI in our future lives and work.

4 Individual contribution

Wang Sunan:
Collecting comments from social media,summarizing the AI article,finishing the social media part of the report

Wang Xuan:
Collecting and downloading recruitment information and summarizing the Finance article,finishing the recruiment part of the report

Cheng Yu:
Collecting API data and finishing the API part of the report,writting the synthesis and suggestion for further study in the report

Bi Huarui:
Using web crawler and finishing the Web crawler part of the report

Hua Neng
Collecting and completing other informaion of the recruiment information,finishing the progress part of the report,using Latex to integrate the report