Final Project: Proposal

Shawn Anderson, Ka Hang Jacky Lok, Vimohitha

1. List 3 questions that you intend to answer (1 point)

Can we visualize the cryptocurrency ecosystem?
Can we cluster cryptocurrencies, or market participants?
Is it possible to predict the future of the cryptocurrency ecosystem?

2. List all the datasets you intend to use (1 point)

We will implement web scrapers to obtain our data. We will also stream twitter data. We will run a 3 component experiment in data collection. Our goal is to collect and analyze over 500GB of data.

Component 1:
For each 100 coins listed on coinmarketcap.com, we plan to scrape that coin's:

1. GitHub Repo (For statistics: stars/forks/issues/contributors)
2. Home website (For raw text, images, videos)
3. Forum (For raw text, sentiment, opinion, key word extraction, TFIDF, Jaccard Sim)

All of the above can be scraped from coinmarketcap.com for each coin.

Component 2:
A general purpose, topic specific crawler for cryptocurrencies. This component will crawl the internet, maintaining a pagerank-like data structure. It will be topic specific, so only pursue pages which mention one of our 100 coins.

Component 3:
Additionally, we plan on collecting twitter data for each coin to build a social media inference component. For each of the same 100 coins, we will obtain streaming twitter data. For each coin, tweets will be collected if they contain a coin name.

3. Give us a rough idea on how you plan to use the datasets to answer these questions. (2 points)

- Data Collection: Where/how to get data
    - Webscraping
    - coinmarketcap.com
    - github.com
    - twitter.com
    - coin forums and web pages
- Data Exploration: Do you need to conduct EDA in order to understand the data?
    - Yes. Our analysis process will be iteratively explorative. We plan on performing EDA on data as we begin collecting them from sources. We believe our EDA end product will depend on where early EDA leads us (Like Detectives).
- Data Cleaning: Do you need to clean data? How to clean them?
    - We need to structure scraped data such that it can be served via an API to our visualization front-end - which means, we need to clean the data.
    - We clean the data by removing unrelated data , Ex: html tags, symbols
    - We may also remove some ID like data, Ex: names, id
    - We see all of our final data as models with fields, which can be rendered as JSON objects. JSON can be easily passed around the system from scraper, to django, to hive, to spark, to JS for visualization. So each scraper must boil data down to a JSON object.
    - So we need to define a rigid shema for our data in terms of models and fields.
- Data Integration: Do you need to integrate data from multiple sources?
    - Yes we will be integrating data from multiple sources - Ex: multiple webpages, twitter, github, forum data
- Data Analysis: What analysis do you intend to do? (e.g., SQL, Statistics, Deep Learning) How to evaluate your analysis results? (e.g., evaluation metrics, confidence intervals, benchmark)
    - Statistics - Confidence Intervals
    - Natural Language Processing
    - Language Visualization (Word Cloud, Vector/Graph embedding)
    - Clustering
    - Cluster Visualization
    - Machine Learning(Bonus)

- Data Product: What product do you want to build? (e.g., visualizations, an interactive web app, a jupyter notebook)
    - Web based visualizations
- Project Architecture:
    - Data Collection: Python, lxml, requests, scrapy
    - Data Storage: Hive File System (HDFS) / Cassandra
    - Data Interface: Spark / Django Rest API
    - EDA: pandas, matplotlib, seaborn, bokeh
    - Deep Analysis: sparkml(Bonus), pytorch(Bonus))
    - Data Visualization: Javascript, D3, Echarts, EDA

4. Think about that once your project is complete, what impacts it can make. Pick up the greatest one and write it down. (1 point)

We are very inspired by the paper “Above the Clouds: A Berkely Perspective on Cloud Computing”. We wish to do for the cryptocurrency ecosystem, what this paper did for the cloud computing ecosystem. That is, to provide a high level analysis of the major trends that are unfolding in the dynamics of an emerging industry - The Cyptocurrency Ecosystem.

5. Questions we have.

Should we use hive hdfs as our data backend, with spark as our interface, or cassandra as our data backend, with spark/django as our interface.
We want fast queries on 500GB of data, how do we do it?

5. References.

1. [Deep Reinforcement Learning for the Financial Portfolio Management Problem](https://arxiv.org/pdf/1706.10059.pdf)

[implementation](https://github.com/ZhengyaoJiang/PGPortfolio)
[replication](https://github.com/wassname/rl-portfolio-management)

2. [Evolutionary Dynamics of the Cryptocurrency Market](http://rsos.royalsocietypublishing.org/content/4/11/170623)

3. coinmarketcap.com

4. [Analyzing Cryptocurrency Markets Using Python](https://blog.patricktriest.com/analyzing-cryptocurrencies-python/)

5. [Predicting Cryptocurrency Prices with Deep Learning](https://dashee87.github.io/deep%20learning/python/predicting-cryptocurrency-prices-with-deep-learning/)