National College of Ireland

BSc in Computing

2017/2018

Dylan Kilkenny

x14472008

x14472008@student.ncirl.ie

Cryptocurrency Analysis

Technical Report



**Table of Contents**

[Executive Summary 4](#_Toc499861897)

[1 Introduction 6](#_Toc499861898)

[1.1 Background 6](#_Toc499861899)

[1.2 Aims 7](#_Toc499861900)

[1.3 Technologies 8](#_Toc499861901)

[1.4 Structure 10](#_Toc499861902)

[2 System 11](#_Toc499861903)

[2.1 Requirements 11](#_Toc499861904)

[2.1.1 Functional requirements 13](#_Toc499861905)

[2.1.1.1 Use Case Diagram 14](#_Toc499861906)

[2.1.1.2 Requirement 1: Gather Data 14](#_Toc499861907)

[Description & Priority 14](#_Toc499861908)

[Use Case 14](#_Toc499861909)

[2.1.1.3 Requirement 2: Scrub and Standardise data 16](#_Toc499861910)

[Description & Priority 16](#_Toc499861911)

[Use Case 16](#_Toc499861912)

[2.1.1.4 Requirement 3: Sentiment Analysis 18](#_Toc499861913)

[Description & Priority 18](#_Toc499861914)

[Use Case 18](#_Toc499861915)

[2.1.1.5 Requirement 4: Data Analysis 20](#_Toc499861916)

[Description & Priority 20](#_Toc499861917)

[Use Case 20](#_Toc499861918)

[2.1.1.6 Requirement 5: Visualise Results 22](#_Toc499861919)

[Description & Priority 22](#_Toc499861920)

[Use Case 22](#_Toc499861921)

[2.1.1.7 Requirement 6: Registration 24](#_Toc499861922)

[Description & Priority 24](#_Toc499861923)

[Use Case 24](#_Toc499861924)

[2.1.1.8 Requirement 7: Login 26](#_Toc499861925)

[Description & Priority 26](#_Toc499861926)

[Use Case 26](#_Toc499861927)

[2.1.1.9 Requirement 8: View top N cryptocurrencies 28](#_Toc499861928)

[Description & Priority 28](#_Toc499861929)

[Use Case 28](#_Toc499861930)

[2.1.1.10 Requirement 9: Sentiment Analysis Result 30](#_Toc499861931)

[Description & Priority 30](#_Toc499861932)

[Use Case 30](#_Toc499861933)

[2.1.1.11 Requirement 10: Specific Cryptocurrency data 32](#_Toc499861934)

[Description & Priority 32](#_Toc499861935)

[Use Case 32](#_Toc499861936)

[2.1.1.12 Requirement 11: Analysis Results 34](#_Toc499861937)

[Description & Priority 34](#_Toc499861938)

[Use Case 34](#_Toc499861939)

[2.1.1.13 Requirement 12: Download Dataset 36](#_Toc499861940)

[Description & Priority 36](#_Toc499861941)

[Use Case 36](#_Toc499861942)

[2.1.2 Non-Functional Requirements 37](#_Toc499861943)

[2.1.2.1 Availability requirement 37](#_Toc499861944)

[2.1.2.2 Recover requirement 37](#_Toc499861945)

[2.1.2.3 Reliability requirement 38](#_Toc499861946)

[2.1.2.4 Environmental requirement 38](#_Toc499861947)

[2.1.2.5 Extensibility requirement 38](#_Toc499861948)

[2.1.3 Data requirements 38](#_Toc499861949)

[2.2 Design and Architecture 41](#_Toc499861950)

[2.3 GUI 42](#_Toc499861951)

[2.3.1 Login 42](#_Toc499861952)

[2.3.2 Register 43](#_Toc499861953)

[2.3.3 Statistical Analysis Result 44](#_Toc499861954)

[2.4 Implementation 44](#_Toc499861955)

[2.5 Testing 44](#_Toc499861956)

[2.6 Customer testing 44](#_Toc499861957)

[2.7 Evaluation 45](#_Toc499861958)

[3 Conclusions 45](#_Toc499861959)

[4 Further development or research 45](#_Toc499861960)

[5 References 45](#_Toc499861961)

[6 Appendix 46](#_Toc499861962)

[6.1 Project Proposal 46](#_Toc499861963)

[6.1.1 Objectives 46](#_Toc499861964)

[6.2 Project Plan 50](#_Toc499861965)

[6.3 Monthly Journals 51](#_Toc499861966)

[6.3.1 September 51](#_Toc499861967)

[My Achievements 51](#_Toc499861968)

[My Reflection 52](#_Toc499861969)

[Supervisor Meetings 52](#_Toc499861970)

[6.3.2 October 52](#_Toc499861971)

[My Achievements 52](#_Toc499861972)

[My Reflection 53](#_Toc499861973)

[Intended Changes 53](#_Toc499861974)

[Supervisor Meetings 53](#_Toc499861975)

[6.3.3 November 53](#_Toc499861976)

[My Achievements 53](#_Toc499861977)

[My Reflection 53](#_Toc499861978)

[Intended Changes 53](#_Toc499861979)

[Supervisor Meetings 54](#_Toc499861980)

[6.4 Other Material Used 54](#_Toc499861981)

# 

# Executive Summary

The purpose of this project is to perform an analysis on the comments by users on cryptocurrency discussion platforms and the prices of the mentioned cryptocurrencies. Although the blockchain has been around close to 10 years, cryptocurrencies are still a relatively young market. Bitcoin which has the first mover advantage has always been a dominating force within the sphere, but within the last year the market has seen a huge increase in interest. One year ago, the entire market capitalization of all cryptocurrencies was $13 billion and bitcoin held a dominance of 88% with a market cap of $11.5 billion. Compare that to today and all cryptocurrencies have a market cap of $310 billion with bitcoin only having a 54% dominance. The entire sphere has seen growth of more than 2284% in less than a year and bitcoin is beginning to lose a lot of its market share to newer more promising currencies. Unlike traditional stocks, cryptocurrencies have no tangible value attached and prices fluctuate with perceived value. Cryptocurrency trading relies on speculation more so than fundamental analysis and as a result most of the price speculation takes place online. Due to the abundance of data regarding speculation around prices on various cryptocurrency discussion platforms, I hope to find a correlation between the sentiment of the posts and the prices of the mentioned cryptocurrencies. This project will follow the KDD methodology.

# Introduction

## Background

I first heard about bitcoin in 2012 and made my first purchase in 2013. At the time, I had very little understanding of how the underlying technology that powered bitcoin, the blockchain, worked and as far as I was aware bitcoin was the only cryptocurrency. Apart from checking the bitcoin price occasionally I had not looked into the cryptocurrency space much further for a few years.

At the beginning of 2017 I came across an article detailing a new currency called ethereum that promised to dethrone bitcoin. Unlike bitcoin, ethereum isn't just another digital currency, ethereum is a decentralised platform that runs applications on top of the blockchain, opening up the blockchain space to an array of new use cases. One of the most important features of the ethereum blockchain is the ability to create tokens. These tokens are treated the same as digital currency and can be used in the decentralised applications they are associated with as well as traded for other digital currencies. With the discovery of ethereum I began to look into the cryptocurrency space a little further and to my pleasant surprise there was more than just bitcoin and ethereum. Before I knew it, I was a frequent visitor of various cryptocurrency subreddits and forums.

Unlike traditional stocks, cryptocurrencies have no tangible value attached and prices fluctuate with perceived value. Cryptocurrency trading relies on speculation more so than fundamental analysis and as a result most of the price speculation regarding cryptocurrencies takes place online. Around May 2017 the cryptocurrency market seen an explosion in interest. As a result, thousands of new users came flooding into the cryptocurrency forums searching for the next best coin to invest in. Most of the cryptocurrencies or tokens being recommended had great potential for making a real-world impact but at the time had no product, or development was a few years away from completion. Despite a lot of the cryptocurrencies having no product or being half finished they began to increase in price anyway. With the influx of new users came a lot of inexperienced traders putting their money where ever the consensus was. The cryptocurrency market is still quite young so there is a lot of volatility. Regulations are scarce and price manipulation from pump and dumps or traders with a large bankroll is a lot easier than traditional stock markets. Compared to the stock market it does not take a lot of buying or selling power to change the price and with the arrival of novice traders that invest in cryptocurrencies solely because they see them recommended a lot, I believe it may be possible to find a correlation between the sentiment of posts on the various platforms and the prices of the mentioned cryptocurrencies.

## Aims

**Aim 1:** The overall aim of this project is to perform sentiment analysis on the comments of various cryptocurrency discussion platforms. Sentiment analysis is the computational task of determining what opinion a user is expressing in text through NLP. For example, one of the simplest forms of sentiment analysis is to classify words as “positive”, “negative” or “neutral” and then average the values of each word to classify the text. This analysis technique is very useful for gaining an overview of the publics opinion on certain topics. By measuring the sentiment of comments on the most popular cryptocurrency platforms this project hopes to find a correlation between the sentiment of the discussions and the price movements of cryptocurrencies.

**Aim 2:** Although sentiment analysis is great for discovering the overall opinion of a piece of text, it is not perfect and as with any analysis there will be errors in the result. Topic modelling will combat this by adding another layer of classification, allowing us to summarise, understand and organise the dataset. Topic modelling can be described as a method for finding a group words from within a dataset which best represents the information within. By applying topic modelling to the comments from the various platforms the system will gain useful insight into how topics on the platform are evolving over time.

**Aim 3:** The third aim of this project is to analyse the top 50 cryptocurrencies ordered by market capitalisation with the help of technical indicators. Technical indicators derive from technical analysis (TA), which is a trading tool employed to forecast prices by analysing statistics gathered from trading activity, such as volume and price movement. The theory of TA assumes that price movements are not random but follow identifiable trends that repeat over time. Primarily designed for short term price movements, technical indicators can turn trends into actionable trading plans. The relative strength index (RSI) is one of the most popular indicators and is primarily used to identify when an asset is overbought or oversold.

**Aim 4:** The final aim of this project is to develop a web app for users to login and interact with the results. The web app will provide the results of this analysis with an R shiny interface, allowing users to select and manipulate the results to make their own conclusions. The web app will also provide users with trading signals based on technical analysis of the top 50 cryptocurrencies listed on coinmarketcap.com

## Technologies

**Python 3.6**

For this project, I will be using python as my main scripting language. Python is an object oriented, high level language with dynamic typing and binding, making it very useful for rapid development. I will be using python to scrape posts and comments from the various platforms. With regards to twitter.com, coinmarketcap.com and reddit.com I will be getting most, if not all the data, through their API’s. Python has a built-in requests module which makes the process of sending GET requests to the API’s all that easier. Coinmarketcap API is quite simple and with few endpoints, so I will be querying their API directly with the requests library. Twitter and reddit have quite a lot of resources on their API and can often be tricky to navigate.

**BeautifulSoup (Python Third-party Library)**

Bitcointalk.org, notably one of the largest forums for cryptocurrency discussion has no API, to tackle this issue I will write a custom script using a third-party python library called BeautifulSoup. BeautifulSoup, with the help of the requests module, will allow me to extract data from bitcointalk by querying html elements directly.

**Python API wrappers**

To simplify the development approach I will be using two open source python wrappers found on GitHub for the reddit and twitter API’s. For reddit I will be using *PRAW* (Python Reddit API Wrapper), and for twitter I will be using *python-twitter.* Both these wrappers will help speed up the development process.

**R**

R language will be used to perform the data scrubbing, classification and analysis. R offers a wide variety of statistical and graphical techniques and is highly extensible.

**R Shiny**

Shiny is an R package that simplifies the process of building web apps for data visualisation. Shiny builds interactive applications which automatically updates, allowing outputs to change instantly as users modify inputs.

**RStudio**

RStudio is an open source IDE designed for the R language. RStudio will serve as the main development environment for this project and all data manipulation and analysis will take place within the IDE.

**MongoDB**

MongoDB is a free and open source document-oriented database and uses JSON like documents with schemas. MongoDB will serve the purpose of storing the user credentials for the web app.

**NodeJS**

NodeJS is an open source JavaScript run-time environment for executing JavaScript code server-side. The web server for the web application will be built using the node framework expressJS. Express is a flexible node.js web framework and provides a powerful set of features for web and mobile apps.

**CSV**

CSV is a text file which separates the values within by commas. This allows data to be saved in a table structured format. CSV files are smaller in size and are easier to generate and parse. They also require less load and setup than a traditional database. All comments, posts, prices along with any other metadata extracted will be saved in CSV format.

## Structure

The structure of this project is broken up into 3 sections:

* **Requirements:** This section details the functional and non-functional requirements of the system along with the use case diagrams.
* **Design and Architecture:** This section contains a diagram of the system architecture from a high level view and information on the flow of the system.
* **GUI:** This section contains screenshots of the web application mock ups.
  1. ***Definitions, Acronyms and Abbreviations***

|  |  |
| --- | --- |
| Coin | Cryptocurrency or token |
| Token | Cryptocurrencies specific to applications built on top of the blockchain |
| KDD | knowledge discovery in databases |
| R | R programming language |
| CSV | Comma separated values file |
| BeautifulSoup | Python third party library for scraping web pages |
| Bitcointalk | Bitcoin forum |
| Altcoin | Cryptocurrencies other than bitcoin |
| NLP | Natural Language Processing |
| TA | Technical Analysis |
| RSI | Relative Strength Index |

# 

# System

## Requirements

The following requirements will adhere to the KDD methodology. KDD stands for knowledge discovery in databases and refers to process of finding knowledge in data and is essential for a data analysis project. The KDD life cycle is explained below:

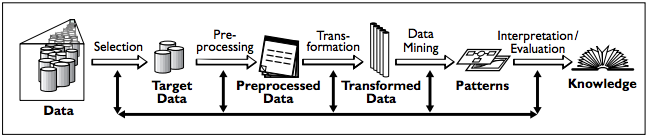


Figure 1: KDD Methodology

**Selection**

This step involves identifying a suitable dataset for the project. This project will be focused on comments and posts from reddit.com, twitter.com and bitcointalk.org. Twitter and reddit API’s offer many endpoints, most of which are not relevant to the scope of this project so this step will also include choosing relevant information.

**Pre-processing**

Pre-processing step cleans and scrubs the dataset removing any unwanted punctuation, whitespace and special characters that may interfere with the analysis. This step is important as the dataset this project is using is comments and posts from online platforms they are most certainly going to contain emoji’s and other special characters that may affect the result. As this project is gathering data from three different sources some of the fields such as date may be represented differently, the pre-processing step will also standardise the datasets into a consistent format.

**Transformation**

This step within the KDD involves generating better data from our cleaned dataset. Using transformation methods or dimensionality reduction it will reduce the number of variables under consideration or to find invariant representations for the data. This step will also classify the data into either positive neutral or negative sentiment by utilising sentiment analysis techniques.

**Data Mining**

Data mining is one of the most vital within the project as here is where we will apply algorithms to the dataset to find trends and relevant information. During this step both cryptocurrency historical prices and comments will be analysed looking for similarities or trends between the sentiment of the comments and price movements of the mentioned cryptocurrencies. Machine learning and statistical algorithms will be utilised within this step.

**Interpretation**

The last step of the KDD methodology involves interpreting the results of the mined data and documenting the conclusions. The results will also be displayed using visualisations to give users a better understanding of the findings.

### Functional requirements

The functional requirements of the system that will be required to complete the project:

1. The system will gather data and save to in CSV format.
2. The system will scrub and standardise the data.
3. The system will classify the data using text analysis.
4. The system will analyse the data.
5. The system will visualise the result.
6. The user will be able register
7. The user will be able to login.
8. The user will be able to view full analysis result.
9. The user will be able to view data for a specific cryptocurrency.
10. The user will be able to view data based on sentiment.
11. The user will be able to view top N cryptocurrencies.
12. The user will be able to download the dataset.

### Use Case Diagram

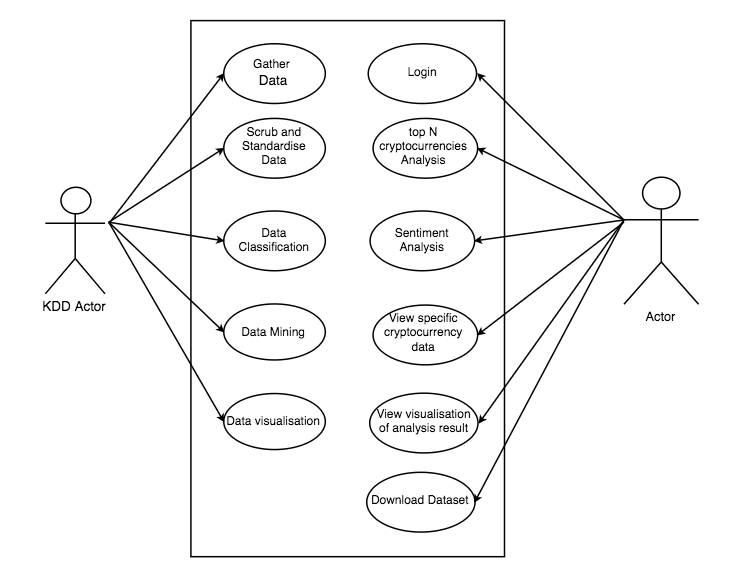


Figure 2: Use Case Diagram

### Requirement 1: Gather Data

#### Description & Priority

This requirement is the most essential for the system. This is the first step and one of the most important as data is required in order for an analysis to be made.

#### Use Case

**Scope**

Identify which information from the platforms is relevant for the purpose of this analysis and gather the relevant data.

**Description**

This use case describes the gathering of data from various cryptocurrency discussion platforms by means of scraping or pulling from API’s.

**Use Case Diagram**



Figure 3: Gather Data

**Flow Description**

**Precondition**

The system has no data

**Activation**

This use case starts when an <Actor> runs the python scripts which will either pull data from the API’s or scrape data from the sites with no API.

**Main flow**

1. The <Actor> identifies relevant data
2. The <Actor> runs the script
3. The script pulls the data
4. The data is saved to a CSV file

**Post condition**

The system imports the file to RStudio

### Requirement 2: Scrub and Standardise data

#### Description & Priority

The data will be scrubbed and standardised into a consistent format by removing punctuation and whitespace. This step is needed to ensure our analysis is as accurate as possible by removing any unwanted data.

#### Use Case

**Scope**

The scope of this use case is to clean the comments by removing any unwanted characters and standardise the data into a common format

**Description**

This use case describes the scrubbing and standardisation of the comments

**Use Case Diagram**

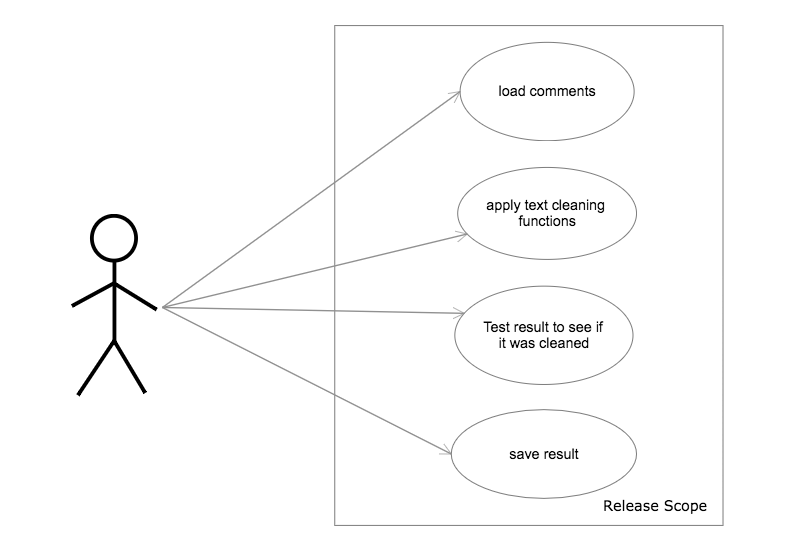


Figure 4: Scrub and Standardise

**Flow Description**

**Precondition**

The comments contain unneeded whitespace and punctuations

**Activation**

This use case starts when an <Actor> imports the dataset to R

**Main flow**

1. The <Actor> loads the dataset into RStudio.
2. The <Actor> applies data scrubbing and standardisation functions.
3. The <Actor> checks the result to see if it was scrubbed correctly.
4. The comments are scrubbed and standardised.
5. The result is saved in CSV format.

**Post condition**

The comments are ready for analysis.

### Requirement 3: Sentiment Analysis

#### Description & Priority

The comments will be classified using sentiment analysis. This requirement is necessary as the focus of the project is on sentiment analysis of the data.

#### Use Case

**Scope**

The scope of this use case is to categorise the comments into being either positive, negative or neutral.

**Description**

This use case describes the process of performing sentiment analysis on the comments.

**Use Case Diagram**

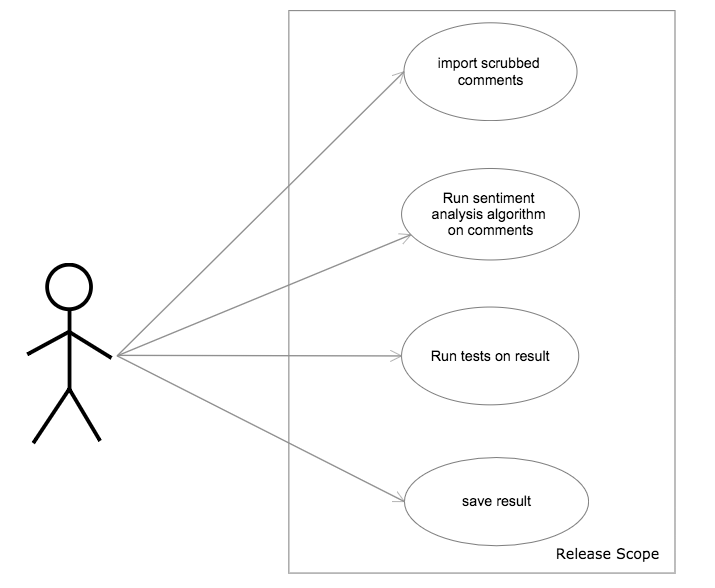


Figure 5: Sentiment Analysis

**Flow Description**

**Precondition**

The comments have been scrubbed of any unwanted data that would interfere in the analysis process

**Activation**

This use case starts when an <Actor> imports the scrubbed data to RStudio

**Main flow**

The <Actor> loads the scrubbed data into R.

The <Actor> runs sentiment analysis algorithm on the comments.

The <Actor> reviews the results.

The <Actor> save the results.

**Post condition**

The system is ready to analyse the comments and prices.

### Requirement 4: Data Analysis

#### Description & Priority

This requirement will make use of statistical algorithms and machine learning to find a correlation between sentiment of users comments and the prices of cryptocurrencies

#### Use Case

**Scope**

The scope of this use case is to do an analysis on the user comments using statistical algorithms in order to create models which will allow us to determine if they have any effect on the price of the cryptocurrency they are discussing

**Description**

This use case describes the process of analysing the comments and prices.

**Use Case Diagram**

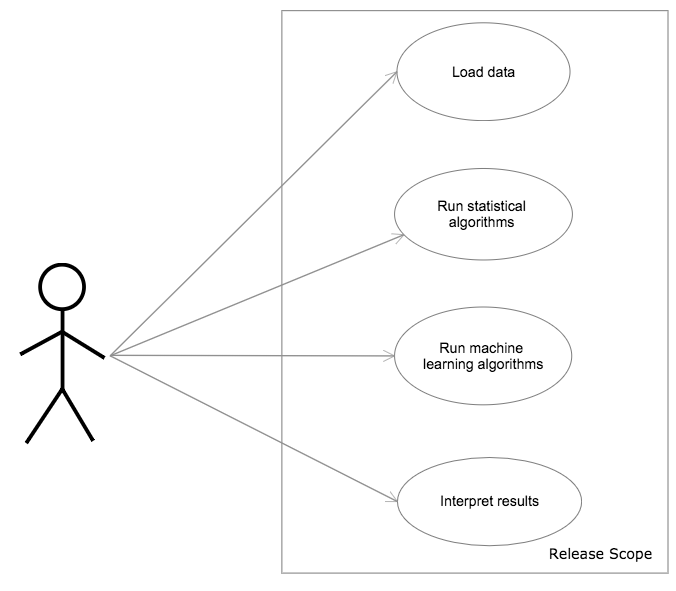


Figure 6: Data Analysis

**Flow Description**

**Precondition**

The comments have been classified using sentiment analysis.

**Activation**

This use case begins when an <Actor> loads the data

**Main flow**

1. The <Actor> loads the data into R.
2. The <Actor> runs statistical analysis algorithms on the comments.
3. The <Actor> runs machine learning algorithms on the comments.
4. The <Actor> interprets the results.

**Post condition**

The system is ready to visualise the results.

### Requirement 5: Visualise Results

#### Description & Priority

This requirement will visualise the results of the statistical analysis to help the user interpret the results. This is the final stage of the project and is crucial for the user to be able to make a conclusion from the results.

#### Use Case

**Scope**

The scope of this use case is to visualise the results of the analysis using various charting techniques within R such as histograms, word clouds, scatter plots, etc.

**Description**

This use case describes how the system will display the results of the analysis.

**Use Case Diagram**

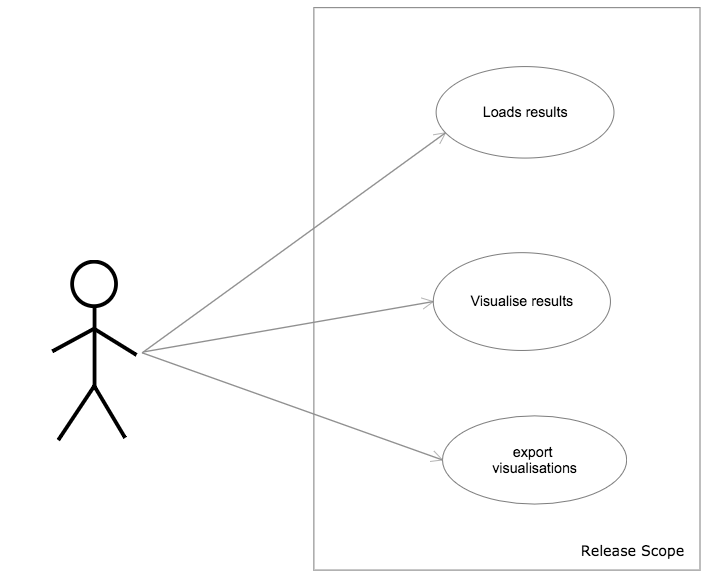
****

Figure 7: Visualise Results

**Flow Description**

**Precondition**

The system has performed statistical analysis on the prices and comments.

**Activation**

This use case begins when an <Actor> loads the results of the analysis

**Main flow**

1. The <Actor> loads results into R.
2. The <Actor> uses R visualisation algorithms to visualise the results.
3. The <Actor> exports the visualisations

**Post condition**

The system enters a wait state

### Requirement 6: Registration

#### Description & Priority

The user will be able to register for the web app to login and view the results. Without registration, the user will not be able to access the dashboard.

#### Use Case

**Scope**

The scope of this use case is to register the user.

**Description**

This use case describes how the user will register with their details and become a member of the system.

**Use Case Diagram**

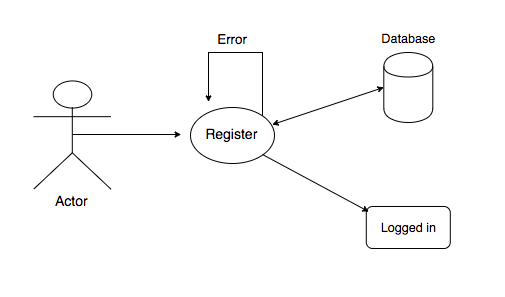
****

Figure 8: Registration

**Flow Description**

**Precondition**

The user is not registered

**Activation**

This use case begins when an <Actor> clicks the register button

**Main flow**

1. The Actor clicks the register button.
2. The Actor enters their details.
3. The Actor is registered and logged in.

**Exceptional flow**

E1: Invalid Details

1. The Actor enters their details.
2. The system returns them invalid and displays a warning.
3. The use case continues at step 2 of the main flow.

**Termination**

The system logs the user in.

**Post condition**

The system enters a wait state

### Requirement 7: Login

#### Description & Priority

The user will be able to log into the app with their registered details

#### Use Case

**Scope**

The scope of this use case is to log in the user.

**Description**

This use case describes how the user will log into the web app with their email and password.

**Use Case Diagram**

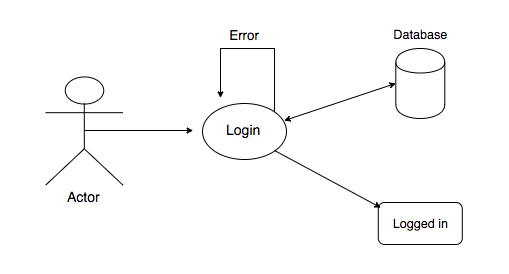
****

Figure 9: Login

**Flow Description**

**Precondition**

The Actor is not logged in

**Activation**

This use case begins when an Actor clicks the log in button

**Main flow**

1. The Actor clicks the log in button.
2. The Actor enters their email and password.
3. The Actor is logged in.

**Exceptional flow**

E1: Invalid credentials

1. The Actor enters their details.
2. The system returns a warning for invalid email or password.
3. The use case continues at step 2 of the main flow.

E2: User not registered

1. The Actor enters their details
2. The system returns an error for email not recognised

**Termination**

The system logs the user in.

**Post condition**

The system enters a wait state

### Requirement 8: View top N cryptocurrencies

#### Description & Priority

The user will be able to view the top N cryptocurrencies based on the analysis results.

#### Use Case

**Scope**

The scope of this use case is for the user to view the top 5, 10, 20 or 50 cryptocurrencies based on the result of the analysis.

**Description**

This use case describes how the will view the top N cryptocurrencies.

**Use Case Diagram**

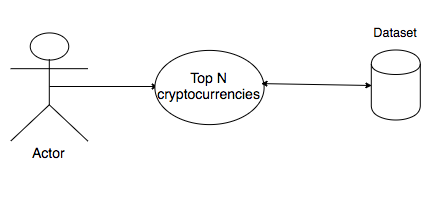
****

Figure 10: View top N cryptocurrencies

**Flow Description**

**Precondition**

The Actor is logged in

**Activation**

This use case begins when an Actor selects top N cryptocurrencies section

**Main flow**

1. The Actor selects top N cryptocurrencies.
2. The Actor selects number of cryptocurrencies to view.
3. The system returns the top N.

**Exceptional flow**

E1: Retrieval error

1. The Actor selects top N cryptocurrencies.
2. The Actor selects number of cryptocurrencies to view.
3. The dataset returns an error.
4. The use case begins at step 2 of the main flow.

**Termination**

The Actor logs out.

**Post condition**

The system enters a wait state

### Requirement 9: Sentiment Analysis Result

#### Description & Priority

The user will be able to view the sentiment associated with each coin from the various cryptocurrency platforms

#### Use Case

**Scope**

The scope of this use case is for the user to view and interact with the result of the sentiment analysis of the comments from the cryptocurrency platforms. The user will be able to view the current sentiment associated with each coin

**Description**

This use case describes how the will interact with the results of the sentiment analysis

**Use Case Diagram**

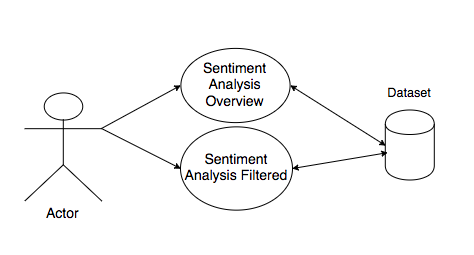
****

Figure 11: Sentiment Analysis Result

**Flow Description**

**Precondition**

The Actor is logged in

**Activation**

This use case begins when an Actor clicks Sentiment analysis section

**Main flow**

1. The Actor clicks sentiment analysis section.
2. The system returns the sentimental analysis overview.
3. The Actor views analysis overview

**Alternate Flow**

A1: Filtered selection

1. The Actor clicks sentiment analysis section.
2. The system returns the sentimental analysis overview
3. The Actor filters by sentiment
4. The system returns results filtered by sentiment
5. The Actor views filtered results.

**Exceptional flow**

E1: Retrieval error

1. The Actor clicks sentiment analysis section.
2. The dataset returns an error.

**Termination**

The Actor logs out or leaves site.

**Post condition**

The system enters a wait state

### Requirement 10: Specific Cryptocurrency data

#### Description & Priority

The user will be able to view all data relating to a specific cryptocurrency, such as price, 24-hour volume, sentiment and any other relatable results from the analysis.

#### Use Case

**Scope**

The scope of this use case is for the user to view and interact with all data relating to a specific cryptocurrency.

**Description**

This use case describes how the user will view data relating to a specific cryptocurrency.

**Use Case Diagram**

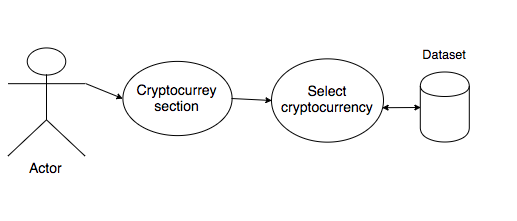
****

Figure 12: Specific Cryptocurrency data

**Flow Description**

**Precondition**

The Actor is logged in

**Activation**

This use case begins when an Actor clicks cryptocurrency section

**Main flow**

1. The Actor clicks cryptocurrency section.
2. The Actor selects a cryptocurrency.
3. The System fetches and returns all data for the cryptocurrency
4. The Actor views the data.

**Exceptional flow**

E1: Retrieval error

1. The Actor clicks cryptocurrency section.
2. The dataset returns an error.
3. The use case continues at step 2 of the main flow.

**Termination**

The Actor logs out or leaves site.

**Post condition**

The system enters a wait state

### Requirement 11: Analysis Results

#### Description & Priority

The user will be able to view and interact with the results of the statistical analysis

#### Use Case

**Scope**

The scope of this use case is to visualise the results of the analysis using various charting techniques within R such as histograms, word clouds, scatter plots, etc.

**Description**

This use case describes how the user will view and interact with the results of the statistical analysis

**Use Case Diagram**

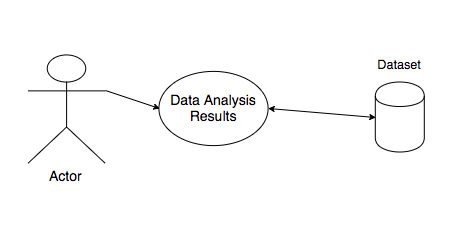
****

Figure 13: Register

**Flow Description**

**Precondition**

The Actor is logged in

**Activation**

This use case begins when the Actor clicks statistical analysis section

**Main flow**

1. The Actor clicks statistical analysis section.
2. The System fetches and returns the results.
3. The Actor views and interacts with the results.

**Exceptional flow**

E1: Retrieval error

1. The Actor clicks statistical analysis section.
2. The dataset returns an error.
3. The use case continues at step 1 of the main flow

**Termination**

The Actor logs out or leaves site.

**Post condition**

The system enters a wait state

### Requirement 12: Download Dataset

#### Description & Priority

The user will be to download the dataset used within the web app.

#### Use Case

**Scope**

The scope of this use case is to allow the user to download the dataset.

**Use Case Diagram**

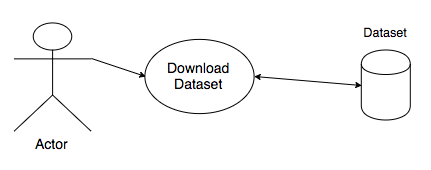
****

Figure 14: Register

**Flow Description**

**Precondition**

The Actor is logged in

**Activation**

This use case begins when the Actor clicks download dataset button

**Main flow**

1. The Actor clicks download dataset button.
2. The System fetches the dataset.
3. The dataset begins downloading.

**Termination**

The Actor logs out or leaves site.

**Post condition**

The system enters a wait state

### Non-Functional Requirements

### Availability requirement

The system and all its components will be made publicly available through GitHub. The users will be able to access the app through a web browser on mobile, tablet and pc.

### Recover requirement

All code and datasets will be backed up locally as well as with cloud based storage providers such as Google Drive and GitHub.

### Reliability requirement

The system will meet all functional requirements. The scripts within the system will run as long as the machine has necessary dependencies installed. The web app will be running at all times and in the event of a server crash the user will have to log back in again.

### Environmental requirement

The system will run on mac, windows and Linux. The machine will be required to have python 3 and RStudio installed. The python and R scripts will come with some dependencies which are required to be installed also. The web app is accessible through the web browser and can viewed via any standard internet enabled device.

### Extensibility requirement

The system will be extensible by both increasing the size of the data applying more complex algorithms for finding patterns and indicators within the comments. With the use of clustering the system could discover hidden semantic structure within the comments.

### Data requirements

To perform the analysis a sufficient dataset is required. Comments will be gathered from the most popular cryptocurrency discussion platforms and prices will be taken from cryptocurrency price aggregator. I chose 3 sources for comments and posts as different platforms cater to different users with different opinions and levels of expertise.

**Reddit.com**

Reddit.com is one of the largest sites for newcomer and experienced crypto enthusiasts. The site is built around subreddits or communities devoted to specific topics where users are free to share links and post questions or topics for discussion. A unique feature of reddit is the ability to create a subreddit for any topic you can think of. This has allowed the cryptocurrency community to expand into communities and sub communities devoted to niche areas. A few examples of such would be the bitcoin subreddit [/r/btc](http://reddit.com/r/btc), which is mainly for bitcoin news and trading discussion, [/r/BitcoinBeginners](http://www.reddit.com/r/BitcoinBeginners/) which is aimed at newcomers to bitcoin who have questions or need advice, and also [/r/CryptoCurrency](http://www.reddit.com/r/CryptoCurrency/) which is intended for open discussion on all cryptocurrencies and not just bitcoin. The variety of user base on reddit will allow this project to gather posts and comments from users ranging from first time buyers to blockchain enthusiasts. The most popular cryptocurrency communities on reddit will be chosen and the data will be pulled using reddit’s API.

**Bitcointalk.org**

Bitcointalk.org is the oldest and most well-known forum for discussion of blockchain technologies. It’s mainly focused around technical details and development issues surrounding bitcoin but offers an array of other boards for trading discussion, alternative currencies discussion and a marketplace where users can buy and sell goods and services for bitcoin. This year alone bitcointalk has been receiving an average of 70,000 posts per month. A lot of speculation regarding new and upcoming cryptocurrency begins on bitcointalk and the altcoin announcement board is the most popular of all by a factor of 4. Unlike users of reddit, bitcointalk users tend to have a better understanding of cryptocurrencies and the underlying technologies. As a result the discussions regarding new and upcoming cryptocurrencies delve a lot deeper than just price speculation and tend to revolve more so around fundamentals. Bitcointalk does not offer an API so this project will utilise third party python library BeautifulSoup to scrape posts directly from bitcointalk.org.

**Twitter.com**

Although twitter is not solely designed for cryptocurrency discussion it offers a great medium of exchange where news spreads quickly. The users base is much larger than reddit and bitcointalk and every level of user from first time buyer to blockchain developers post their opinions there. Twitter offers an API to developers and this project will mainly be focusing on tweets discussing the cryptocurrencies from coinmarketcap.com’s list of top 100 cryptocurrencies.

**Coinmarketcap.com**

Coinmarketcap.com is one of the most popular sources for cryptocurrency prices and statistics and keeps track of cryptocurrency data such as markets, 24 hour volume, market cap and historical data. Coinmarketcap.com offers a public api and this project will utilise it to pull current and historical price data for the top 100 cryptocurrencies.

* + 1. **User requirements**

From a client’s perspective, the system must display the results of the analysis in a concise and easy to read manner. All visualisations must be clearly labelled along with all relevant information regarding methods and algorithms used to compute the results. The web app requires users to have access to a web browser on an internet connected device. Users are also required to register.

* + 1. **Usability requirements**

The web application should have an easy to navigate user interface and no specific level of expertise will be required to use any of its features. The web app should also display clear and concise errors to the user.

## Design and Architecture

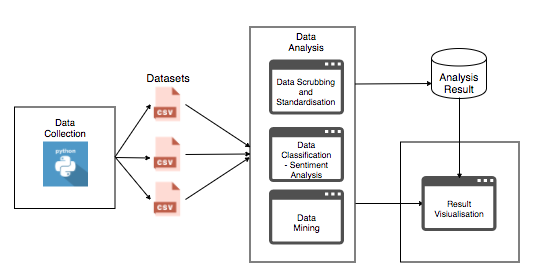
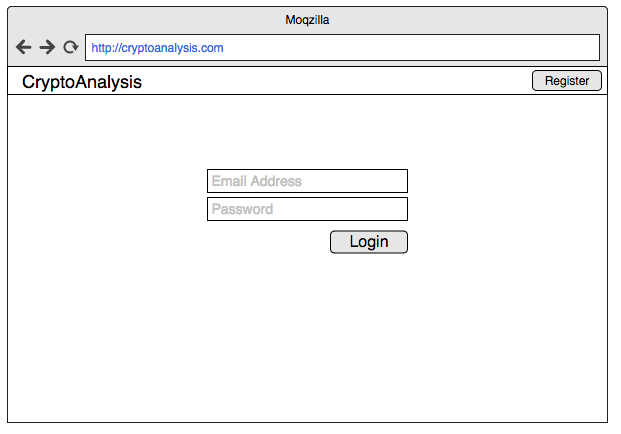


Figure 8: System Architecture

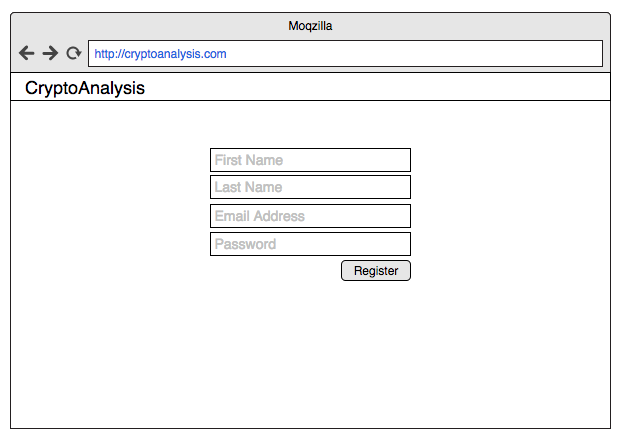
The above architecture diagram shows the system from a high-level view. The system will first gather the dataset from the various platforms and save them to CSV file. The datasets will then be imported to RStudio where they will be scrubbed and standardised, classified according to sentiment and analysed for patterns or trends. The results of the analysis and the dataset will be uploaded to the web server and the results will be visualised using an R shiny interface.

## GUI

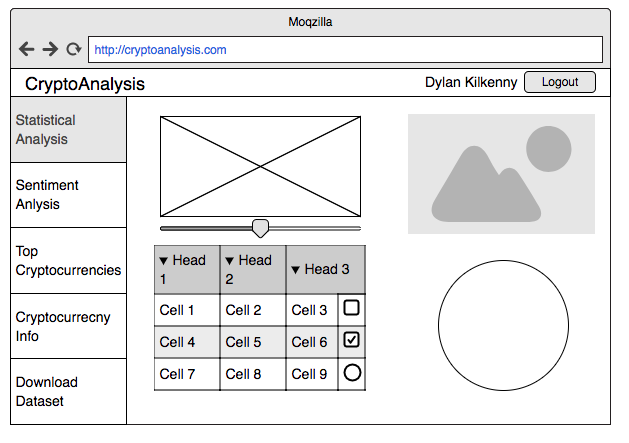
## Login



## Register



## Statistical Analysis Result



## Implementation

Describe the main algorithms/classes/functions used in the code. Consider to show and explain interesting code snippets where appropriate.

## Testing

Describe any testing tools, test plans and test specifications used in the project

## Customer testing

Provide evidence for and results of customer testing. This may include ratings or quotes from the customer.

## Evaluation

How was the system evaluated and what are the results? In many cases this will include usage data and user feedback. It may also include performance evaluations, scalability, correctness, etc. depending on the focus of the project.

Quantative results may be reported in tables or figures. Note that tables have their caption above the table and need to be cross referenced in the text (see **Error! Reference source not found.**). In many cases, tables are better to read if you skip the vertical lines.

Figures have their caption below the figure as shown in **Error! Reference source not found.**. Make sure that if you use colour, the figure is still readable when printed in black & white, e.g., by using additional symbols, patterns, etc.

# Conclusions

Describe the advantages/disadvantages, opportunities and limits of the project.

# Further development or research

With more resources, where could the results of this project lead to?

# References

* [Coinmarketcap.com](http://coinmarketcap.com/). (2017). Historical Snapshots Index | CoinMarketCap. [online] Available at: <https://coinmarketcap.com/historical/> [Accessed 30 Nov. 2017].
* [Kdnuggets.com](http://kdnuggets.com/). (2017).Text Mining 101: Topic Modelling. [online] Available at: <https://www.kdnuggets.com/2016/07/text-mining-101-topic-modeling.html> [Accessed 30 Nov. 2017].
* Staff, I. (2003).Â Technical Analysis. [online] Investopedia. Available at: <https://www.investopedia.com/terms/t/technicalanalysis.asp> [Accessed 30 Nov. 2017].
* Robinson, J. (2017).Â *Text Mining with R*. [online] [Tidytextmining.com](http://tidytextmining.com/). Available at: <http://tidytextmining.com/sentiment.html> [Accessed 1 Dec. 2017].
* Robinson, J. (2017).Â *Text Mining with R*. [online] [Tidytextmining.com](http://tidytextmining.com/). Available at: <http://tidytextmining.com/topicmodeling.html> [Accessed 1 Dec. 2017].

# Appendix

## Project Proposal

# Objectives

**Objective 1:** Gather Datasets - The platforms this project will focus on are reddit.com, twitter.com and bitcointalk.org. Both reddit and twitter offer API’s with an abundance of data, but for the purpose of this project only a subset of relevant information will be required. Bitcointalk.org does not offer an API so this project will make use web scraping technologies to extract the data from the web pages. Historical prices for the top 100 cryptocurrencies will be pulled from coinmarketcap.com’s public API.

**Objective 2:** Scrub and Standardise Dataset - As three different platforms will be used the datasets will slightly differ in format. In order to perform the analysis without any hiccups it is good practice to scrub the dataset of any unwanted characters such as whitespace, punctuations, indentations. The three datasets will also be standardised into a uniform format.

**Objective 3:** Data Classification - By applying sentiment analysis algorithms to the comments they will be classified by either positive, neutral or negative sentiment.

**Objective 4:** Data Analysis - After scrubbing and classifying the dataset statistical analysis will be performed on both prices and comments with the aim of discovering trends or correlations between the two.

**Objective 5:** Visualise Data - Create a visualisation of the research results and display them using tableau for interpretation by the end user.

**Objective 6:** Web Application – After the analysis has been performed the results and dataset should be accessible via a web application. Users will be required to register to view the data and will be able to interact with and interpret the results of the analysis.

* + 1. **Background**

I first heard about bitcoin in 2012 and made my first purchase in 2013. At the time I had very little understanding of how the underlying technology that powered bitcoin, the blockchain, actually worked and as far as I was aware bitcoin was the only cryptocurrency. Apart from checking the bitcoin price occasionally I had not considered the cryptocurrency space much further for a few years.

At the beginning of 2017 I came across an article detailing a new currency called ethereum that promised to dethrone bitcoin. Unlike bitcoin, ethereum isn't just another digital currency, ethereum is a decentralised platform that runs applications on top of the blockchain, opening the blockchain space to an array of new use cases. One of the most important features of the ethereum blockchain is the ability to create tokens. These tokens are treated the same as digital currency and can be used in the decentralised applications they are associated with as well as traded for other digital currencies. With the discovery of ethereum I began to explore the cryptocurrency space a little further and to my pleasant surprise there was more than just bitcoin and ethereum. Before I knew it, I was a frequent visitor of various cryptocurrency subreddits and forums.

Unlike traditional stocks, cryptocurrencies have no tangible value attached and prices fluctuate with perceived value. Cryptocurrency trading relies on speculation more so than fundamental analysis and as a result most of the price speculation regarding cryptocurrencies takes place online. Around May 2017 the cryptocurrency market seen an explosion in interest. As a result, thousands of new users came flooding into the cryptocurrency forums searching for the next best coin to invest in. Most of the cryptocurrencies or tokens being recommended had great potential for making a real-world impact but at the time had no product, or development was a few years away from completion. Despite a lot of the cryptocurrencies having no product or being half finished they began to increase in price anyway. With the influx of new users came a lot of inexperienced traders putting their money where ever the consensus was. The cryptocurrency market is still quite young so there is a lot of volatility. Regulations are scarce and price manipulation from pump and dumps or traders with a large bankroll is a lot easier than traditional stock markets. Compared to the stock market it does not take a lot of buying or selling power to change the price and with the arrival of novice traders that invest in cryptocurrencies solely because they see them recommended a lot, I believe it may be possible to find a correlation between the sentiment of posts on the various platforms and the prices of the mentioned cryptocurrencies.

* + 1. **Technical Approach**

Previous projects I have completed have been focused around software development more so than data analysis. As this is the first project of its kind for me I will begin by researching various techniques within data analysis such as dataset cleansing and text analysis tools. For my dataset I will be gathering user posts and I have chosen 3 sources; bitcointalk.org, reddit.com, twitter.com. I chose these sites as they are the main hubs for online cryptocurrency discussion. Reddit and twitter offer an API which is very helpful. I will not need every piece of information they have to offer on a post and to avoid overabundance of unnecessary data I will only gather relevant information. Bitcointalk does not offer an API unfortunately, but this can be circumvented with the use of Python’s third-party library Selenium which will allow me to directly query html on Bitcointalk and extract necessary data. To make a comparison between the sentiment of posts and cryptocurrency prices I will need historical and current price data. Luckily this data is easily accessible from many public API’s. I chose coinmarketcap.com for the historical and current prices as they are reliable and have a proven track record within the cryptocurrency sphere. Their API is public, requiring no key, and provides helpful data other than price, such as volume and price percent changes over set periods of time. Python will be my language of choice for gathering the dataset and I will gather all relevant posts on the various platforms from the last 12 months. Before performing an analysis on the dataset, it will first need to be cleaned. I will begin by merging my three data sets into one, making sure they are all represented in a consistent format, then identify any missing data and standardising the dataset. To cleanse the data even further I will remove any duplicate data as well as removing any punctuations and unnecessary whitespace. Having scrubbed the dataset I will begin looking for patterns and performing text analysis on the posts using R. More specifically, applying sentiment analysis to the individual comments using RSentiment. Finally, I will interpret the results and create appropriate visualizations to represent it.

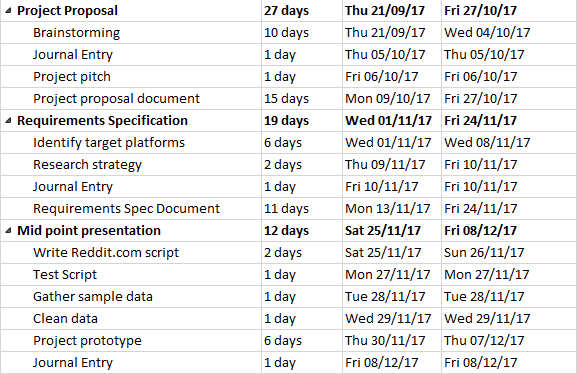
* + 1. **Technical Details**

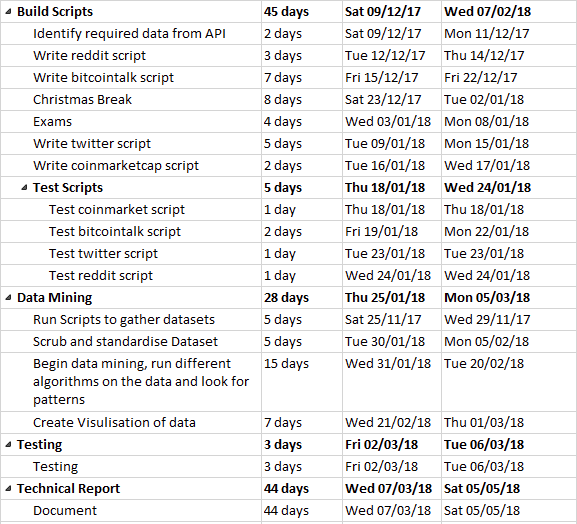
For the purpose of this project I will be using python as my main scripting language. Python is an object oriented, high level language with dynamic typing and binding, making it very useful for rapid development. I will be using python to scrape posts and comments from the various platforms. With regards to twitter.com, coinmarketcap.com and reddit.com I will be getting most, if not all the data, through their API’s. Python has a built-in requests module which makes the process of sending GET requests to the API’s all that easier. Coinmarketcap API is quite simple and with few endpoints, so I will be querying their API directly with the requests library. Twitter and reddit have quite a lot of resources on their API and can often be tricky to navigate. To simplify the approach I will be using two open source python wrappers found on GitHub for the respective API’s. For reddit I will be using *PRAW* (Python Reddit API Wrapper), and for twitter I will be using *python-twitter.* Both these wrappers will help speed up the development process. Bitcointalk.org, notably one of the largest forums for cryptocurrency discussion has no API, to tackle this issue I will write a custom script using a third-party python library called BeautifulSoup. BeautifulSoup, with the help of the requests module, will allow me to extract data from bitcointalk by querying html elements directly. All comments and posts along with any other metadata extracted will be saved in comma separated format (CSV). For my analysis I will be using R to manipulate and visualise the dataset. R includes a built in library called RSentiment which will allow me to perform sentiment analysis on the posts I have gathered.

* + 1. **Evaluation**

I will be running various unit tests on all scripts and API connections to verify the quality of the code. Datasets will be checked for credibility as they are being scraped and pulled. Results of the analysis will be displayed using visualisation techniques to allow the end users to interpret the results.

## Project Plan





## Monthly Journals

## September

Student name: Dylan Kilkenny

Programme (e.g., BSc in Computing):

Month: September

# My Achievements

This Month we began back at college. We were prepped for our pitch in Eamon’s class and he gave us some very helpful advice regarding our projects and our pitch. Luckily I have been thinking about my project idea since I met with Simon Caton last May while on work placement. I decided to go with a project related to my stream, data analytics, rather than build something. Thankfully it was approved.

# My Reflection

My project revolves around an analysis of forum comments and news article related to cryptocurrency’s, with the hope of finding a correlation between the sentiment of these comments/articles and the price movements of motioned currency’s. I got some positive feedback within the pitch and it was suggested I place my focus on digital currency’s that aren’t bitcoin as they aren’t studied as much. Overall I am happy with my choice of project and look forward to beginning.

# Supervisor Meetings

No meeting held .

## October

Student name: Dylan Kilkenny

Programme : BSHC in Computing

Month: October

# My Achievements

This month I began to think about my project a bit more and what is needed to perform the analysis. I decided to focus my attention on three platforms instead of just one as it will give a lot more variety to the dataset. I choose bitcointalk.org, reddit.com and twitter.com as they are the 3 biggest platforms for cryptocurrency discussion. I had a lot of project work this month as well as multiple CA’s so I was beginning to feel the pressure of fourth year already. The project proposal for software project was due at the end of the month also so the last two weeks consisted of researching the strategy needed to implement my project idea.

# My Reflection

I feel as though I need to manage my time better as towards the end of the month the assignments started to pile up and I felt as though I was rushing to complete them.

# Intended Changes

Next month I will be focusing more on my software project. I will gather a sample dataset from one of the platforms and prepare a prototype for December.

# Supervisor Meetings

I had a meeting with Adriana Chis on 26th October where we discussed some details about my project. Adriana also explained the format of our meetings will be group meetings and they will be held twice weekly.

## November

Student name: Dylan Kilkenny

Programme : BSHC in Computing

Month: November

# My Achievements

This month I began writing the python scripts to collect the dataset. We have a prototype due for the first week of November and I will be gathering a test dataset from reddit to work on. I also have begun researching sentiment analysis techniques as this is the core focus of my project.

# My Reflection

This month the pace is starting to pick up and I am really beginning to feel the 4th year work load. My time management needs some improvement.

# Intended Changes

Work on time management.

# Supervisor Meetings

Absent from meeting

## Other Material Used