Individual Project (CS3IP16)

Department of Computer Science
University of Reading

Project Initiation Document

PID Sign-Off

Student No.	**24005432**
Student Name	**Andrew Sotheran**
Email	**andrew.sotheran@student.reading.ac.uk**
Degree programme (BSc CS/BSc IT)	**BSc CS**
Supervisor Name	**Kenneth Boness**
Supervisor Signature	
Date	

SECTION 1 – General Information

Project Identification

1.1	**Project ID** (as in handbook)
	N/A
1.2	**Project Title**
	Cryptocurrency market and value prediction tracking
1.3	**Briefly describe the main purpose of the project in no more than 25 words**
	To provide a means to predict the value of cryptocurrencies that will aid in investor decision making in investment of the market

Student Identification

1.4	**Student Name(s), Course, Email address(s)** e.g. Anne Other, BSc CS, a.other@student.reading.ac.uk
	Andrew William Sotheran BSc CS Andrew.sotheran@student.reading.ac.uk

Supervisor Identification

1.5	**Primary Supervisor Name, Email address** e.g. Prof Anne Other, a.other@reading.ac.uk
1.6	**Secondary Supervisor Name, Email address** Only fill in this section if a secondary supervisor has been assigned to your project

Company Partner (only complete if there is a company involved)

1.7	**Company Name**
	N/A
1.8	**Company Address**
	N/A
1.9	**Name, email and phone number of Company Supervisor or Primary Contact**
	N/A

SECTION 2 – Project Description

2.1 **Summarise the background research for the project in about 400 words. You must include references in this section but don't count them in the word count.**

To create a tool that aims to predict the price of cryptocurrencies that aids in investor decisions. Research will need to be conducted into the following topics that surround data mining, machine learning and artificial neural networks.

This research will consist along the lines of;
Natural Language processing and analysis – To analyse and process fed in data gathered through RSS data feeds and social media feeds, through the underlying tasks of Natural language processing. Content categorisation (search and indexing, duplication detection), Topic discovery and modelling (Obtain meanings and themes within the data and perform analytic techniques), sentiment and semantic analysis (which will identify the mood and opinions within the data), summariser (to summarise a block of text and disregard the rest).

Machine learning algorithms: The three types of machine learning (Supervised, Unsupervised and Reinforced)
The types of common algorithms used, each of these will be researched to identify the most suitable for this project and only one will be used: (Linear Regression, Logistic Regression, Decision Tree, SVM, Naive Bayes, kNN, K-Means, Random Forest, Dimensionality Reduction Algorithms, Gradient Boosting algorithms (GBM, XGBoost, LightGBM, CatBoost).

Artificial Neural Networks: To identify the drawbacks and benefits of using them or other computational models within machine learning. Recurrent Neural networks and 3rd generation Neural Networks.

Data mining: To investigate the different techniques and algorithms used (Same as the ones listed above for machine learning including C4.5, Apriori, EM, PageRanks, AdaBoost and CART) these will be researched and the most appropriate identified.

To investigate techniques: for storing and processing large amount of data, such as Hadoop, Elasticsearch utilities, Graphing and data modelling and visualisation.

To identify appropriate libraries for python or C for each of the topics above to aid in the creation of this project. Libraries such as:
Natural Language Toolkit (NLTK) – python
Pandas - python
Sklearn - python
Numpy – python - scientific computation for working with arrays
Matplotlib - python - data visualisation

Investigate into types of databases. Sql and nosql for a storage medium between receiving data and feeding it into the machine learning algorithm.
Investigate into the use of REST API and other web-service based technologies (GRPC, Elasticsearch)
Investigate into frameworks for the thin client, such as Angular vs React, Nodejs, Leafelt.js, charts.js
Additionally Web scraping may be needed if certain website that don't either have an API or JSON for the data needed.

https://www.sas.com/en_gb/insights/analytics/what-is-natural-language-processing-nlp.html
https://blog.algorithmia.com/introduction-natural-language-processing-nlp/
https://gerardnico.com/data_mining/algorithm
https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/
https://www.kdnuggets.com/2015/05/top-10-data-mining-algorithms-explained.html
https://www.datasciencecentral.com/profiles/blogs/artificial-neural-network-ann-in-machine-learning
http://scikit-learn.org/stable/index.html
https://grpc.io/docs/

2.2	**Summarise the project objectives and outputs in about 400 words.** These objectives and outputs should appear as tasks, milestones and deliverables in your project plan. In general, an objective is something you can do and an output is something you produce – one leads to the other.
	To produce a thin web client that provides a dashboard that provides tangible and useful information to users such as; Their current price (Updated every 5 minutes), exchange rates, network hashrates, historical price data. It will also display statistics about sentiment analysis conducted on social media about the currency, graphical predictions on what the price may be, in a given time, and will also compare this to other currencies for aid in investment. To produce significant research into the topics in and around data mining, machine learning and Artificial Neural network and the underlying tasks and algorithms used, the efficiency, drawbacks and advantages of each to identify the most suitable for the use in this project. To produce a system that analyses a data set obtained through social media feeds and posts on news sites regarding crypto currencies. It should perform sentiment analysis using Natural Language processing and analysis techniques to identify features and identifies the type of sentiment in the data and categorises it for machine learning. To utilise machine learning techniques and algorithms to produce a system that learns from historical data to predict to an extent the possible future price of a given currency. To compare this with the use of an Artificial Neural Network and to analyse the drawbacks of both.
2.3	**Initial project specification - list key features and functions of your finished project.** Remember that a specification should not usually propose the solution. For example, your project may require open source datasets so add that to the specification but don't state how that data-link will be achieved – that comes later.
	The finished project should provide a thin client single page application. This will provide a means to users the ability to view various statistics on crypto currencies on a dashboard that incorporates text analysis through natural language analysis, and will utilise various machine learning and data mining techniques to provide price predictions to the users. The nature and level of this will depend on the research conducted into the areas of data mining, machine learning, natural language processing and artificial neural networks, along with the algorithms used. The data set will be created from scratch for this project as it will require the gathering of data from numerous sources and performing text analysis on them to for the data needed. Data sets for the characteristic and data for the currencies can be obtained from pre-existing data sets such as: https://www.kaggle.com/sudalairajkumar/cryptocurrencypricehistory https://www.kaggle.com/jessevent/all-crypto-currencies Web scraping may be included if certain news/social media websites do not provide an API or RSS feed for the analysis engine to perform text analysis on Additionally, there will be a server between the analysis/prediction engine and the thin client that will maintain a database, either SQL or NoSQL, that will hold statistics about the currencies and data about the price predictions about the currencies. It will not hold any of the data used in the analysis engine, as this database will only hold data available to the end users.

2.4	**Describe the social, legal and ethical issues that apply to your project. Does your project require ethical approval? (If your project requires a questionnaire/interview for conducting research and/or collecting data, you will need to apply for an ethical approval)**
	The project will not be handling any user related data, therefore it does not need ethical approval.
2.5	**Identify and lists the items you expect to need to purchase for your project. Specify the cost (include VAT and shipping if known) of each item as well as the supplier.** e.g. item 1 name, supplier, cost
	None Needed
2.6	**State whether you need access to specific resources within the department or the University e.g. special devices and workshop**
	Possibly a server to host the database and analysis engine on to perform the computation necessary, and a server to host the thin client.

SECTION 3 – Project Plan

3.1	**Project Plan** Split your project work into sections/categories/phases and add tasks for each of these sections. It is likely that the high-level objectives you identified in section 2.2 become sections here. The outputs from section 2.2 should appear in the Outputs column here. Remember to include tasks for your project presentation, project demos, producing your poster, and writing up your report.

Task No.	Task description	Effort (weeks)	Outputs
1	**Background Research**		
1.1	Investigate into RPC frameworks and REST APIs	0.3	To identify the type of API/RPC framework that would be most suitable
1.2	Research into Natural Language processing and analysis techniques	0.5	To get an understanding of how NLP works and how it could be used
1.3	Research into the use of machine learning – types and algorithms	0.5	To grasp how ML paradigms work and how this project will use it
1.4	Research into the application of Neural Networks – drawbacks and advantages of using them	0.3	To identify whether there will be a need for a neural network or ML paradigms can be used instead
1.5	Research techniques for storing and processing large amount of data, such as Hadoop, spark or Elasticsearch utilities.	1	To understand the uses, application and whether the use of these are more viable solution than standard ML practices
1.6	Identify appropriate libraries for data modelling and visualisation, NLP and Machine Learning	1	To identify what libraries will aid in the construction of this project
1.7	Investigate into frameworks for the front-end thin clients	0.3	To identify what frameworks the thin client should be used with, along with drawbacks and advantages
1.8	Research web scraping techniques	0.3	To understand the application of these techniques and learn how to apply them
2	**Analysis and design**		
2.1	Resolve issues discovered by background research	0.2	…
2.2	Identify limitations discovered from research and what is not feasible	0.1	…
2.3	UML Diagrams/ XUML	0.2	
2.4	Wire frames for frontend	0.1	
2.5	Data Flow	0.1	
2.6	User Flow	0.1	
3	**Develop prototype**		
3.1	Develop thin client	2	
3.2	Develop analysis Engine	4	
3.3	Develop Prediction Engine	3	
3.4	Develop Unit tests	2	
4	**Testing, evaluation/validation**		
4.1	Unit testing	1	
4.2	Acceptance Testing	0.8	
4.3	User testing	0.8	
5	**Assessments**		
5.1	write-up project report	2	Project Report
5.2	produce poster	0.5	Poster
5.3	Log book	0.5	

TOTAL	**Sum of total effort in weeks**	**21.9**	

SECTION 4 - Time Plan for the proposed Project work

For each task identified in 3.1, please *shade* the weeks when you'll be working on that task. You should also mark target milestones, outputs and key decision points. To shade a cell in MS Word, move the mouse to the top left of cell until the curser becomes an arrow pointing up, left click to select the cell and then right click and select 'borders and shading'. Under the shading tab pick an appropriate grey colour and click ok.

	START DATE: 10/2018 **<enter the project start date here>**												
	Project Weeks												
Project stage	0-3	3-6	6-9	9-12	12-15	15-18	18-21	21-24	24-27	27-30	30-33	33-36	36-39
1 Background Research													
Investigate into RPC frameworks and REST APIs													
Research into Natural Language processing													
Research into the use of machine learning –													
Research into the application of Neural													
Research techniques for storing and													
Identify appropriate libraries for data													
Investigate into frameworks for the front-													
Research web scraping techniques													
2 Analysis/Design													
Resolve issues discovered by background													
Identify limitations discovered from													
UML Diagrams/ XUML													
Wire frames for frontend													
Data Flow													
User Flow													

3 Develop prototype.													
Develop thin client													
Develop analysis Engine													
Develop Prediction Engine													
Develop Unit tests													
4 Testing, evaluation/validation													
Unit testing													
Acceptance Testing													
User testing													
5 Assessments													
write-up project report													
produce poster													
Log book													

RISK ASSESSMENT FORM

Assessment Reference No.		**Area or activity assessed:**	
Assessment date			
Persons who may be affected by the activity (i.e. are at risk)	**Andrew Sotheran**		

SECTION 1: Identify Hazards - *Consider the activity or work area and identify if any of the hazards listed below are significant (tick the boxes that apply).*

1.	Fall of person (from work at height)		**6.**	Lighting levels		**11.**	Use of portable tools / equipment		**16.**	Vehicles / driving at work		**21.**	Hazardous fumes, chemicals, dust		**26.**	Occupational stress	
2.	Fall of objects		**7.**	Heating & ventilation		**12.**	Fixed machinery or lifting equipment		**17.**	Outdoor work / extreme weather		**22.**	Hazardous biological agent		**27.**	Violence to staff / verbal assault	
3.	Slips, Trips & Housekeeping	X	**8.**	Layout , storage, space, obstructions		**13.**	Pressure vessels		**18.**	Fieldtrips / field work		**23.**	Confined space / asphyxiation risk		**28.**	Work with animals	
4.	Manual handling operations		**9.**	Welfare facilities		**14.**	Noise or Vibration		**19.**	Radiation sources		**24.**	Condition of Buildings & glazing		**29.**	Lone working / work out of hours	
5.	Display screen equipment	X	**10.**	Electrical Equipment	X	**15.**	Fire hazards & flammable material		**20.**	Work with lasers		**25.**	Food preparation		**30.**	Other(s) - specify	X

SECTION 2: Risk Controls - *For each hazard identified in Section 1, complete Section 2.*

Hazard No.	Hazard Description	Existing controls to reduce risk	**Risk Level** (tick one)			Further action needed to reduce risks *(provide timescales and initials of person responsible)*
			High	Med	Low	
3	Tripping over wires	Cable management is at a minimum, none are currently properly cable managed and kept out of way			x	Sufficient cable management needed, cables tied together and moved out of way of feet
5	Eye strain from looking at a monitor	Current screen contrast and brightness is acceptable		x		To have periodic breaks from the screen
Name of Assessor(s)			**SIGNED**			
Review date						

Health and Safety Risk Assessments – continuation sheet

Assessment Reference No	
Continuation sheet number:	

SECTION 2 continued: Risk Controls

Hazard No.	Hazard Description	Existing controls to reduce risk	**Risk Level** (tick one)			Further action needed to reduce risks *(provide timescales and initials of person responsible for action)*
			High	Med	Low	
Name of Assessor(s)			**SIGNED**			
Review date						