**INITIAL RESEARCH PROPOSAL FORM**

(also referred to as ‘Statement of Intent Form’)

***To be submitted by the researcher to the Institute Research Sub-Committee (IRC)***

|  |  |
| --- | --- |
| **Research Title:**  Using Natural Language Processing to assure Data Quality in Web Scraping Operations. | |
| **Institute name**  Institute of Information and Communication Technology, | |
| **Course / Programme:**  Bachelor of Science (Honours) in Business Analytics | |
| **Level and year of study**  MFQFL 6, Second Year | |
| **Main area of study being proposed:**  This study explores the possibility of coupling Natural Language processing with web scraping techniques to create a data stream with a high standard of data quality.  Data analysis requires a vast amount of data to produce accurate insights and return value. Datasets can be had in a myriad variety of methods with various amounts of resource requirements.  Using the Beautiful Soup and Text Blob Python Libraries, Web Scraping and Natural Language Processing capabilities have become highly accessible to programmers of all levels. This project explores the possibility of using web scraping and Natural Language Processing as an alternative data source and aggregator for feeding a data model with vasts amount of usable data. The main objective is to analyse the Data Quality and Integrity of the resultant data files and ensure they are within standards to ensure the accuracy of all analysis and insights won't be hindered by a poor data foundation. | |
| **Name of Researcher:**  Matthew De Giorgio | **Researcher’s I.D. Number:**  0047796M |
| **Signature of Researcher** | **Date of submission of Form**  5th April 2019 |
| **Name of Tutor (or Recommended Tutor):**  Mr. Alan Gatt | |

|  |
| --- |
| **Personal Motivation for the Choice of Research Theme.** |
| Coming from a family background of pharmaceutical importation and then starting to work in a leading technology consulting company, I felt that this research area was a perfect fit. Amidst daily operations, our director of logistics would encounter many situations were quick decisions had to be made. If a product shipment got delayed, we would have to decide whether we leave it alone and incur the late delivery fine or we purchase another shipment from a different supplier that would guarantee the product would arrive on time at a much higher cost. By the time the Net Profit calculations for each situation were completed, the time window for a positive outcome would have closed.  Having experienced this situation from a personal level, I was intrigued to explore the possible benefits that this technology could bring if it was to be implemented in the local market space. Seeing that large foreign companies have similar systems in place to make such situations as easy as possible, why not try to make a similar system and apply to a small company?  With this context, the purpose of this research is to create a prototype and assess whether it’s feasible to be implemented into the local companies to reduce the strain on resources and management whilst increasing profitability in the highly competitive local market and perhaps bring some new insight and value to the family’s business. |

|  |
| --- |
| **Outline of Key Literature and Theoretical Framework or Propositions.** |
| *Enter details here* |

|  |
| --- |
| **Significance of the Study.** |
| The main focus of the study and the prototype is to predict the result of a shipment with a certain degree of accuracy using past and live data. A dashboard will be created to show the key stats of a shipment, the Net profit if it were to arrive on time and the net profit if it were to arrive late. Other Suppliers that also sell the product will be listed including an estimated purchase price, the last possible date to order for the shipment to arrive on time and the Net Profit if that supplier was to be used.  Another focus is to derive the main factors that are involved during transportation which could be used to predict the outcome of the shipment (the expected arrival vs the actual arrival time).  All data sets used will be based on European Air Traffic since that the operating zone of many local companies. If the data sets were to be expanded to include Air Traffic within the United States, future research could be conducted to answer the following questions:   1. Would the algorithm work with Transocean air-freight? 2. Could the algorithm evolve to predict the most efficient purchasing strategy if a big enough database of suppliers was linked to it?   The prototype may bring value to any business that exports and imports to sustain business deals and tenders that require goods to arrive on particular days. |

|  |
| --- |
| **Hypotheses and/or Research Question/s** |
| 1. *Can Shipments be predicted accurately using past data?* 2. *Would further broadening of the data sources increase the accuracy of algorithm?* 3. *Would the algorithm be accurate enough to take in changing parameters such as weather changes or changes in air traffic flow due to airport problems?* |

|  |
| --- |
| **Target Participants and Research Methods for Data Collection and Analysis** |
| 1. Natural language processing would be used on RSS feeds and on web scraping data (using Beautiful Soup for Python) to create a dataset of events and the time delay they put on air freight. 2. A Microsoft SQL database would be built to house the data. Suppliers, products, air freight providers will also be included in the database. 3. Analysis would be conducted on the dataset to find patterns, and key statistics (fastest way from point A to Point B, cost per kg for transport) 4. An aspect of probability would have to be included to factor in the chances of delays during particular times of the year, particular locations (Ex: Snow in the northern Regions of Europe) 5. The results be visualized using Tableau or PowerBi to compare the generated models and graphs. 6. Measuring of the results and statistical tests to validate the algorithm and to show why such results are being given. 7. The results would be compared to actual destination arrival times. |

|  |
| --- |
| **Ethical Considerations.**  **Refer to *guidance points below. You are also additionally required to read MCAST Document 074 ‘Research Ethics Policy and Procedure’ that is available on the College website via link*** [***http://www.mcast.edu.mt/MainMenu/Full-TimeCourses/Rules,PoliciesandRegulations.aspx***](http://www.mcast.edu.mt/MainMenu/Full-TimeCourses/Rules,PoliciesandRegulations.aspx)   1. *Research shall be conducted in such a manner so as to avoid any psychological and physical harm to humans and animals and financial damage to organizations* 2. *Only the supervisor and examiners will have access to any data gathered.* 3. *Participants will remain free to withdraw from the study at any time without having to provide any reason. In the case of withdrawal, all the records and information collection will be deleted.* 4. *The participant, who is the sole proprietor of the data provided, is granting that such data would be processed for this study purposes only.* 5. *The data collection process will be a transparent process.* 6. *All transcriptions and/or electronic recordings reflecting the data collected, once exhausted, are to be deleted* 7. *Confidentiality, anonymity and data protection procedures are to be ethically abided by.* 8. *The researcher would provide a soft copy of the study to the participant, if required.* |
| *Enter details here regarding possibility of issues regarding confidential personal data:*  *How will you ensure that:*   * *No personal data or confidential data is divulged.* * *Participants’ identities are not divulged (ie kept anonymous)* |
| *Enter details here regarding possibility of physical harm:*   * *How will you ensure that no person or animal gets hurt during the implementation of the research?* * *What Personal Protective Equipment (PPE) will you be needing/using?* |
| *Enter details here regarding possibility of moral harm:*   * *What steps will you take to avoid unduly offending or disturbing the well-being of the participants?* * *How will you avoid any possible psychological, spiritual or cultural offence to participants?* * *How will you ensure that the interests of minors / vulnerable / disabled persons are safeguarded where necessary?* |
| *Enter details here regarding possibility of business harm:*  *How will you ensure that:*   * *Participants do not suffer any competitive disadvantage as an outcome of the research?* * *Confidential business ideas and data are protected and not divulged?* |

|  |
| --- |
| **Anticipated Contributions of the Study.** |
| *Enter details here* |

|  |
| --- |
| **Dissertation Project Plan.** |
| *Enter details here* |

**List of Key References:**

***This section is to be filled in by the representative of the Institute Research Sub-Committee prior to forwarding of this Form to the ‘MCAST Research Ethics Committee’ for final ethics approval:***

|  |  |  |
| --- | --- | --- |
| ***Nature of ethical consideration*** | ***Outcome (\*)*** | ***Comments*** |
| *Consideration of possibility of issues regarding confidential personal data:* |  |  |
| *Consideration of possibility of physical harm* |  |  |
| *Consideration of possibility of moral harm* |  |  |
| *Consideration of possibility of business harm* |  |  |

***(\*) Legend to record outcome by Institute Research Sub Committee:***

***A***  *– Ethical considerations have been* ***addressed appropriately*** *by Researcher;*

***B*** *– No (****Nil****) relevant ethical considerations are applicable under purpose of study as described by Researcher.*

***C*** *– Ethical consideration have* ***not been addressed appropriately*** *by Researcher;*

***D*** *– Applicable ethical consideration have* ***not been considered*** *by Researcher.*

|  |  |
| --- | --- |
| **Details of Representative to the ‘Institute Research Sub-Committee.** | |
| Name | Signature |
| Designation | Date |