**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.** |
| I decided to investigate this subject from a data analyst’s point of view. Data Analytics is one of the main components of my workflow and investigating alternative ways to improve the workflow’s efficiency is part of the improvement process. When conducting data analysis, the more data points one has, the broader and more detailed the ‘picture’ becomes. This allows for increased accuracy in insights and help train a data model to a greater degree. Establishing a new possible data aggregator would allow for comparative analysis (C. Pickvance,2005) and strengthen predictive analytics through stronger data models that could lead to better predictions. |

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

|  |
| --- |
| **Significance of the Study.** |
|  |

|  |
| --- |
| **Hypotheses and/or Research Question/s** |
| *H1 : The tag analysis will pass 90% accuracy based on previous tests*  *H0 : The tag analysis will score below 90% .*  *In relation to H0 , why should it score below 90% ? This could point to a flaw in the program or an anomaly.* |

|  |
| --- |
| **Target Participants and Research Methods for Data Collection and Analysis** |
| The research will undertake a a qualitative approach since the focus of the project is to establish a proof of concept for the dissertation project. A positive-based research philosophy allows for the following methodology by observing phenomena from an objective viewpoint. The phenomena are isolated and easily replicated.  A non-probability sampling method will be used, specifically using Accidental and Snowball Sampling Methods. The articles will be chosen by a few web searches centered around the subject of Air-Traffic delays and accidents. Once again, this subject was chosen since the following dissertation project will use predictive analytics to refine Business decisions concerning Logistics. The analytics will be powered by a data model that requires training.   1. The researcher will select ten article / pages to be scraped by a python script running BeautifulSoup. 2. Using an NLP Library such as NLTK or Textblob , the article content will be keyed 3. Tagging Analysis   The Accuracy will be calculated by counting the actual amount of tagged words in the articles / pages by hand and then comparing that number to the amount of tagged words given by the NLP script. |

|  |
| --- |
| **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:*  Information collected in the proposed research study will be considered as information that the participants have  “disclosed in a relationship of trust” and so no information will be divulged without their permission. Anonymity  means that the research cannot link individual responses with the participants’ identities. All the information  collected during the study will be kept in a multi-factor authentication secured cloud storage facility. All information  processed during the study serves only for the purpose of the study itself and therefore once the research is  complete, all transcripts and any information reflecting the data collected are to be destroyed. All the information  collected during the study will be kept in a biometrically secured cloud storage facility.  Attached to this SOI is a model Consent form that will be issued before any research activities are to be initiated. |
| *Enter details here regarding possibility of physical harm:*  The methodology pertained by this study does require the use of animals for the scope of the study, neither will  there be any physical harm to any interview attendees or respondents of the questionnaire. Furthermore, no  Personal Protective Equipment (PPE) will be needed as the questionnaires and interviews will be dispensed and  carried out in a safe office environment to eliminate the possibility of physical harm. |
| *Enter details here regarding possibility of moral harm:*  Maintenance of the basic ethical principles is to be upheld to highest of standards during the data collection process  including doing good, protecting the autonomy, wellbeing, safety and dignity of all participants. The researcher will  maintain objectivity and will persist to avoid any possible psychological, spiritual or cultural misunderstandings with  participants. Participants always have the right to refuse to answer any questions even though the questionnaire  and interview are designed to not involve any professional and emotional risks. |
| *Enter details here regarding possibility of business harm:*  All research findings will be processed for the purpose of the research only and nothing more and will remain  confidential and therefore, participants should not encounter in any way any competitive disadvantage as an  outcome of the research, nor will there be any form of harm to any businesses locally and abroad.  All Data will be held and view only by the Researcher and no one else. As previously described, Secure online storage  will be heavily utilised to reduce the chance of any data loss or breach.  Anonymity and data protection procedures are always to be upheld and so |

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

|  |
| --- |
| **Dissertation Project Plan.** |
|  |

**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 |