**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.** |
| Intro The technologies featured in this study have been developed and studied for since the 1950s leading to rapid advances in the possibilities of Natural Language Processing and the AI field in general. The earliest signs of Web scraping happened around the time of the invention of the World Wide Web. In 1993 “The Wanderer” was written to systematically traverse the Web and create an index since the Web was still in its infancy and did not have any form of the directory *(M.Grey,1996)*.  Natural Language Processing and its underlying algorithms have been in development since 1949 **(***D. Eggers et al,2017).* 50 years later, the school of thought grown into a very powerful analytical skill set that could be implemented into a myriad of situations. Artificial Intelligence Artificial Intelligence has come a long way since it’s conception in 1956 where it was launched by a DARPA-Sponsored conference at Dartmouth College in the USA. Sixty-three years later, the field of Artificial Intelligence has grown into a field greater than the initial question “Can a machine think?” *(Turing, 1950)*. Currently split into three Approaches: The Intelligent Agent Continuum, Logic-Based AI, Non-Logicist AI and finally any AI that does exist within current paradigms. The field experienced explosive growth due to bloom in Machine learning *(T. Mueller, 2006)*. The development of algorithms has split Machine learning into three different learning mechanisms. Some of these learning mechanisms run certain capabilities of Natural Language Processing.  Using Supervised Learning, the computer learns a function’s purpose through the examples and by matching the parameters to the final results. Unsupervised learning allows the machine to find useful knowledge and information when given raw data. It is then left unsupervised so the machine can uncover interesting correlations within the data. This type of AI learning is used in Data mining, were machines search and filter through large amounts of information with the sole incentive of finding something interesting. Google’s PageRank , an early search engine algorithm that ranks web pages falls within the school of thought. A balance of these two schools of thought is Reinforcement Learning were the machine starts the learning activity like the unsupervised learning technique but occasionally it received feedback in the form of rewards or punishments granted that the machine has learned to behave rationally according to the feedback. This type of machine learning exists commonly computer games. Natural Language Processing Current day NLP has many technologies that could be applied to many current problems. The main technology utilised in this study is Text Processing by use of the NLTK python library to aid Information extraction. This technology is used to source meaningful information from large amounts of unstructured text (D. Eggers et al,2017). This could be done through Part of Speech Tagging and token system. Important terms (‘Tags’) are set as parameters in the script. The program will go through the information dump and Count / Highlight all occurrences of the tags. Researchers applied these techniques to filter through police reports and related news article to identify key information such as weapons, vehicles , locations and people with high precision (C.D.Manning,2011). Web Scraping Web scraping is used for many different purposes, from data gathering to web page indexing. As mentioned beforehand , the first very first web scraper was in fact a web crawler based on the Perl programming language. The sole purpose of the World Wide Web Wanderer (WWWW) was to measure the size of the Word Wide web and to generate an index (called Wandex) in 1993 *(M. Grey,1996)*. The primary functionality of the Web crawler was to build the indexes for search engines since in the early days of the Web there were not that many websites and required website administrators to collect the links and enter them manually into a search engine Index.  Beautiful Soup is the one of the most common approaches to web scraping since it does not depend on an API interface (Application programming interface) but instead it parses web page content directly from the HTML container. Designed for Python , the library receives continuous updates and new functionalities every year. Conclusion The key to a successful Natural Language Processing implementation is the outlining of the solution requirements. Establishing the correct requirements translates to the correct pairing of NLP technologies and supporting systems. In the case of this study, the NLTK library with support from Beautiful Soup for data aggregation conform to the requirements. |

|  |
| --- |
| **Significance of the Study.** |
| 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 helps 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.  The primary aim of the research is to discover the efficiency and accuracy of the Part-of-Speech Tagging Analysis and how it can be exploited to solve a typical data problem. |

|  |
| --- |
| **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 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 random web searches. 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 articles/ 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 number 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 the 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 the 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 the 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  the 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. |

|  |
| --- |
| **Anticipated Contributions of the Study.** |
| *Enter details here*  *The results of the study will confirm the accuracy of the Tagging Method. If the accuracy is upheld than the system could be expanded and utlised for data models . In fact the plan for the following dissertation is use this same system as a data aggregator to create live data sets that will be used to teach a data model.* |

|  |
| --- |
| **Dissertation Project Plan.** |
| *The project will span over 3 months with the following chart depicting how research tasks will be split up accordingly. The Timeline is tentative for not all possible setbacks can be taken into consideration at the time of writing.* |

**List of Key References:**

Bringsjord, Selmer, Govindarajulu and Sundar, N. (2018). *Artificial Intelligence (Stanford Encyclopedia of Philosophy)*. [online] Plato.stanford.edu. Available at: https://plato.stanford.edu/entries/artificial-intelligence/#HistAI [Accessed 1 Mar. 2019].

D. Eggers, W., Malik, N. and Graciee, M. (2019). *Using AI to unleash the power of unstructured government data*. [online] Deloitte Insights. Available at: https://www2.deloitte.com/insights/us/en/focus/cognitive-technologies/natural-language-processing-examples-in-government-data.html [Accessed 24 May 2019].

D.Manning, C. (2011). *Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?*. [ebook] Stanford: Department of Linguistics , Stanford University. Available at: https://nlp.stanford.edu/pubs/CICLing2011-manning-tagging.pdf [Accessed 20 Apr. 2019].

Gray, M. (1996). *Internet Growth and Statistics: Credits and Background*. [online] Mit.edu. Available at: http://www.mit.edu/~mkgray/net/background.html [Accessed 3 Jun. 2019].

Hamaz, K. and Benchikha, F. (2017). A novel method for providing relational databases with rich semantics and natural language processing. *Journal of Enterprise Information Management*, [online] 30(3), pp.503-525. Available at: https://emeraldinsight.com/doi/full/10.1108/JEIM-01-2015-0005 [Accessed 2 Jun. 2019].

Iriberri, A. and Leroy, G. (2007). Natural Language Processing and e-Government: Extracting Reusable Crime Report Information. *2007 IEEE International Conference on Information Reuse and Integration*. [online] Available at: https://ieeexplore.ieee.org/document/4296624 [Accessed 11 May 2019].

Loira, S. (2018). *API Reference — TextBlob 0.15.2 documentation*. [online] Textblob.readthedocs.io. Available at: https://textblob.readthedocs.io/en/dev/api\_reference.html#textblob.blob.TextBlob.sentiment [Accessed 2 Jun. 2019].

M. Turing, A. (1950). *OMPUTING MACHINERY AND INTELLIGENCE, Mind,*. 59th ed. Oxford: Oxford University Press, pp.433-460.

Manning, C. and Schütze, H. (1999). *Foundations of Statistical Natural Language Proccesing*. Cambridge (Massachusetts): MIT Press.

Murphy, K. (2013). *Machine learning*. Cambridge, Mass.: MIT Press.

Patel, H. (2018). *How Web Scraping is Transforming the World with its Applications*. [online] Towards Data Science. Available at: https://towardsdatascience.com/https-medium-com-hiren787-patel-web-scraping-applications-a6f370d316f4 [Accessed 1 Jun. 2019].

Pennington, J., Socher, R. and D.Manning, C. (2014). *GloVe: Global Vectors for Word Representation*. [online] Aclweb.org. Available at: https://www.aclweb.org/anthology/D14-1162 [Accessed 4 Jun. 2019].

Raina, R., Madhavan, A. and Y. Ng, A. (2009). *Large-scale Deep Unsupervised Learning using Graphics Processors*. [ebook] Stanford: Stanford University ,Standford. Available at: http://robotics.stanford.edu/~ang/papers/icml09-LargeScaleUnsupervisedDeepLearningGPU.pdf [Accessed 3 Jun. 2019].

Russell, S. and Norvig, P. (2009). *Artificial intelligence*. Upper Saddle River, N.J.: Prentice Hall.

T. Mueller., E. (2006). *Commonsense Reasoning*. 2nd ed. Morgan Kaufmann Publishers.

***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 |