Using Natural Language Processing to assure Data Quality in Web Scraping Operations

Matthew De Giorgio
MCAST
Institute of Communication and
Information Technology
Poala, Malta
matthew.degiorgio.b31281@mcast.edu.
mt

***Abstract*— 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 a few NLP oriented 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 a vast 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.

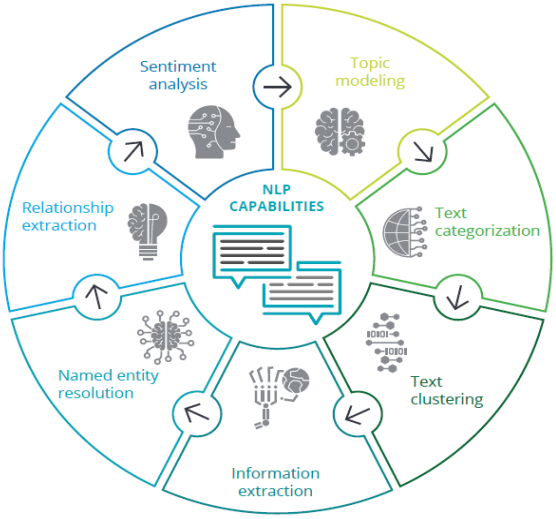
Keywords*— *NLP, Natural Language Processing, Web Scraping, AI, Artificial Intelligence, Python 3

I. INTRODUCTION

This study aims to explore the many benefits one gains when merging two technologies, Web Scraping and Natural Language Processing to provide an alternative source of raw data that could be used for analysis and Insight generation. This concept is being tackled by many institutions around the world including DARPA (America's Defence Advanced Research Project Agency) who create the DEFT (Deep Exploration and Filtering of Text) Program "which utilises natural language processing (NLP), a form of artificial intelligence , to automatically extract relevant information and help analysts derive actionable insights from it"(Deloitte Citation). In this case, Web scraping will be used as the data aggregator. The python library Beautiful Soup is highly configurable, the user could allow it to scrape a whole Domain with particular terms or dial it down to a single webpage. Once collected, the data will undergo a few analytical tests through the use of various python NLP Libraries.

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.

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.



II. Literature Review

A. Introduction

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).

> "The Wanderer was the primary tool for the collection of data to measure the growth of the web. It was the first automated Web agent or "spider". The Wanderer was first functional in spring of 1993 and performed regular traversals of the Web from June 1993 to January 1996. "(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. *Maintaining the Integrity of the Specifications*

B. 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.

C. Natural Language Processing

The first documented use of Natural Language processing occurred in 1949 when IBM was a key component in the making of the "Index Thomisticus", a computer readable compilation of St. Aquina's works. "Artificial Intelligence ", "Pattern Recognition", "Speech Recognition ", "Topic Modelling", "Deep Learning "and "Neural Machine Translation" ; all these terms are milestones in the 50-year long development of the Artificial Intelligence and NLP field.

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).

D. 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.

Due to all the web scraping activity, some countries and website owners have introduced some laws / terms and conditions to restrict the use of web scraping. Currently enforcement is largely present in the United States, with a few court cases occurring around the world *(Eurostat,2017)*.

E. 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.

III. Research Methodology

A. Research Style

The study undertook a qualitative research style since the focus was to establish a proof of concept for a later project that would be on a larger scale. Using this research method, the focus is to establish concepts and test concepts from a small and concise dataset (J. W. Cress,2013). The articles being used by the study are a representation of typical webpages; bodies of text segmented by HTML dividers and classes. The hypothesis being tested on this small dataset can easily be extrapolated to a larger data set in further studies

B. Research Philosophy

Using a positivist-based research philosophy to observe and describe phenomena from an objective viewpoint without interfering the phenomena that are under observation (Levin, 1988). The phenomena will be isolated, and the observations and annotations are easily replicated within other studies. In this case, the tagging mechanism of the NLTK library is being isolated and tested for accuracy. Based on previous research and tests, the prediction that the accuracy of the tagging module is above 90 per cent (C.D Manning, 2011) can be made.

C. Sample Method

Non-probability sampling, specifically Accidental and Snowball sampling methods were utilised to select the articles that will undergo the Part-of-Speech (PoS) Tagging test. A few random searches centred provided enough articles to saturate the tests. This article subject was chosen due to the whole purpose of this test. This study will be used as a steppingstone for a dissertation focusing on using data models to calculate the risk associated with transportation methods used by businesses. Due to the nature of the tests, the sampling method utilised would not affect performance unless the researcher aimed to gather a large data set. Tests that involve NLP tagging and tagged tokens can be run on any type of written content as long as the content is Text-based and in English.

D. Data Collection Method

The article content is aggregated through Python-based Web scraping using the Beautiful Soup library. Apart from a main windows desktop pc, the python script was also tested on a Raspberry PI microcomputer to evaluate whether having a microcomputer running twenty-hours, seven days a week, scraping data was feasible with today's hardware. During testing, the CPU operating temperature hovered between 45 and 50 degrees Celsius. The only bottleneck was the read write speeds of the onboard memory, but this could be rectified for a full implementation by utilising Gigabit-class ethernet and Network Attached Storage arrays.

For testing purposes, the program will be running on a windows machine and not on the raspberry pi just due to the convenience factor. When executed the python program will look up all the articles and load their content into a variable. Next, the text is keyed, separated word by word for future processing by using the Natural Language Tool Kit (NLTK). NLTK is a suite of libraries that are utilised for Symbolic and Statistical Natural language specifically designed for Python.

E. Conclusion

The data collection process has shown that test scaling would not be as taxing as one would find in usual circumstances. The fact that so many well written and well-tested libraries exist shows the sheer modularity and multi-purpose capabilities of the Python programming language. The Beautiful Soup Library has a highly extensive skill set that could be executed on a variety of machines from a small Raspberry Pi running on a low-end mobile chip to a Desktop Workstation Class PC. Smart usage of resources can be further enhanced by using the multi-threading technique in the python script. A well thought out web scraping script could go through terabytes of information in a single day, enough to teach a data model to increase its accuracy efficiently.

With the article content captured and keyed within a few variables, the tagging comparisons may commence.

IV. Evaluation

Word	*Program*	*Manual counting*	*Percentage Found*
amazon	24	24	100%
approval	2	2	100%
prime	9	6	150%
fly	2	2	100%

The Test proved successful. The above dataset shows both possible cases when dealing with web scraping. The program managed to find all user-visible text with just a single anomaly, due to the nature of the scraping, the program found the chosen tag in the web-page code which is not seen by the viewer under normal circumstances.

V. Conclusion

A. Project Outcome

Hypothesis confirmed, within testing bounds, the program managed to find all words visible to the viewer. In one case it found more due to other page elements not viewable by the viewer. There one will need further tune the application to just scrape the visible information. The situation changes from the site, depending on their web design. Extrapolating the dataset could have led to further anomaly discovery.

B. Recommendations for Future Work

The aim of this project was to supplement a reliable data stream for a data model designed for the sole purpose of predicting risk involved in Logistical business decisions. Natural Language processing and Data Aggregation tools such as Web Scraping brought forward the possibility of creating data sets that are constantly being updated as soon as new information enters the World Wide Web.

To Conclude, a drawback to this endeavour was the lack of time. The initial objective was to go through a plethora of

NLP processes and test whether the accuracy of POS Tagging has diminishing returns.

For future projects, the researcher would pursue larger data sets to improve the certainty of the results and observations throughout the project. Natural Language processing is a very powerful collection of tools that automate workflows involving anything to do with raw text. Whilst researching for the literature review, a few interesting research papers exploring the possibility of using NLP to help shift businesses from hard copy to soft copy and providing relational databases with rich semantics and NLP exhibit the capabilities of NLP (K. Hamaz, F.Benchikha, 2017). Furthermore, professional services firms are looking into using Artificial Intelligence like NLP to increase transaction rate within government bodies that contain large amounts of unstructured data (D. Eggers et al,2017).

REFERENCES

[1] 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].

[2] 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].

[3] 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].

[4] 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].

[5] 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].

[6] 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].

[7] 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].

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

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

[10] Murphy, K. (2013). Machine learning. Cambridge, Mass.: MIT Press.

[11] 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].

[12] 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].

[13] 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].

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

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