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

MCAST
Malta College of Arts, Science & Technology

Matthew De Giorgio | Alan Gatt | Institute of Information and Communication Technology

Abstract

The aim of this project was to test the accuracy of the end data after using a web scraping and Natural Language Processing based system with the intention of using the system as an alternative data source and aggregator for feeding a data model with a vast amount of usable data

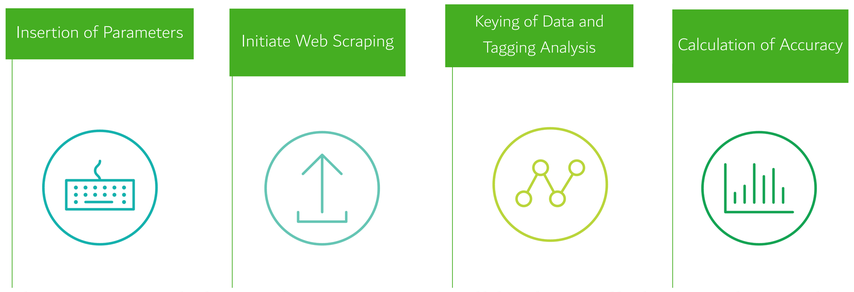
Research Approach

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 a Positivist-based Research Philosophy the will tagging mechanism within the NLTK python library was isolated and tested for accuracy.

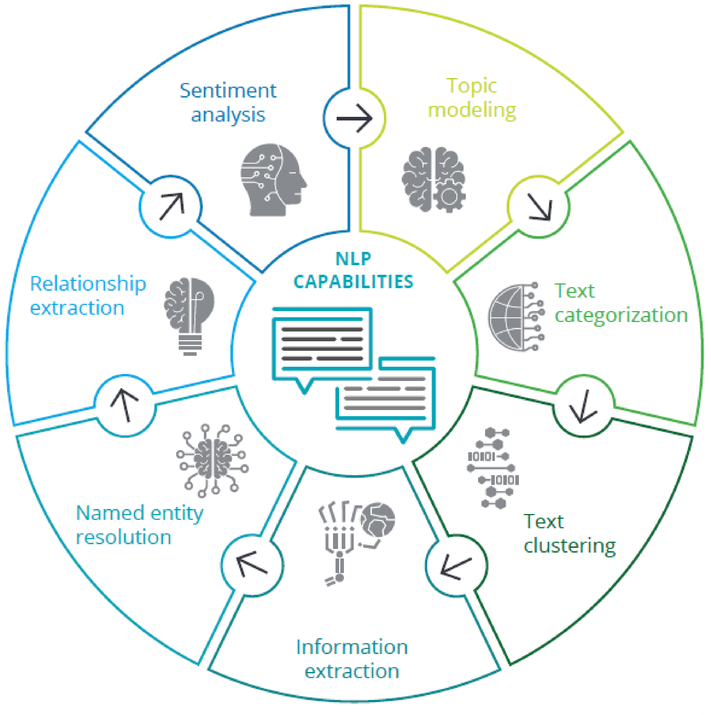
Project Overview

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 the data. Primarily based on two Python Libraries , BeautifulSoup and NLTK , the program sources the information from the designated websites , downloads it and processes it using Part-of-Sentence Tagging.

Methodology



.When executed the python program will look up all the articles and load their content locally. 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 is utilised for Symbolic and Statistical Natural language processing designed for Python. The article content than undergoes Part-of-speech tagging analysis. The accuracy is then calculated by comparing the actual amount of tagged words in the article with the number of tags given by the python program.



Source : Deloitte Insights

Results and Analysis

- *Hypothesis : The program will achieve an accuracy above 90%*
- *Null Hypothesis : The program will not achieve the stated accuracy due to malfunction.*

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 ween by the viewer under normal circumstances.hich is not s

Conclusion

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.

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].