# Title: Using Natural Language Processing to ascertain relative Web Scraping

# 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 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.

**Keywords** : NLP , Natural Language Processing , Web Scraping , AI , Aritificial Intelligence, Python 3

# Literature Review

## Intro

## Subtopic 1 -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.

*Note :*

*The different Approaches to AI;*

* *The Intelligent agent continuum,*
* *Logic-based AI,*
* *Non – logicist AI ,*
* *AI beyond the clash of Paradigms.*

<https://plato.stanford.edu/entries/artificial-intelligence/#HistAI> (AI Date Reference)

## Subtopic 1 - Web Scraping

## Subtopic 2 – Natural Language Processing

# Research Methodology

## Research Style 🡪 Quantitative

## Research Philosophy

Philosophy of Artificial Intelligence

## Sample Method

## Data Collection Method

Python based Web scraping through the use of the beautiful soup python 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 in a full implementation by utilising Gigabit class ethernet and Network Attached Storage arrays.

## Conclusion

# Evaluation

Charts : Sentiment Analysis of Sentences in articles,

Data Quality of articles , the amount of key words found in each lead.

Compare single vs multiple tagging runs

Explain Sentiment Analysis of the Textblob Library.

# Conclusion

# References

<https://plato.stanford.edu/entries/artificial-intelligence/#HistAI> (AI Date Reference)

<https://towardsdatascience.com/https-medium-com-hiren787-patel-web-scraping-applications-a6f370d316f4> (Web Scraping)

<https://www.datacamp.com/community/tutorials/web-scraping-python-nlp> (Web scraping)

<https://emeraldinsight.com/doi/full/10.1108/JEIM-01-2015-0005> (NLP)

<https://github.com/ZackAkil/web-scraping-and-nlp-CLDSPN/blob/master/%5BCOMPLETED%5D%20Web%20Scraping%20and%20NLP.ipynb>

<https://textblob.readthedocs.io/en/dev/api_reference.html#textblob.blob.TextBlob.sentiment>

The main objective of this study is to analyse the Data Quality of the resultant data file and check whether it’s within standards to be used as an alternative data source.