# Assessment 3: WebCrawler and NLP System

**Type:** Written document and Jupyter Notebook

**Weight:** 50%

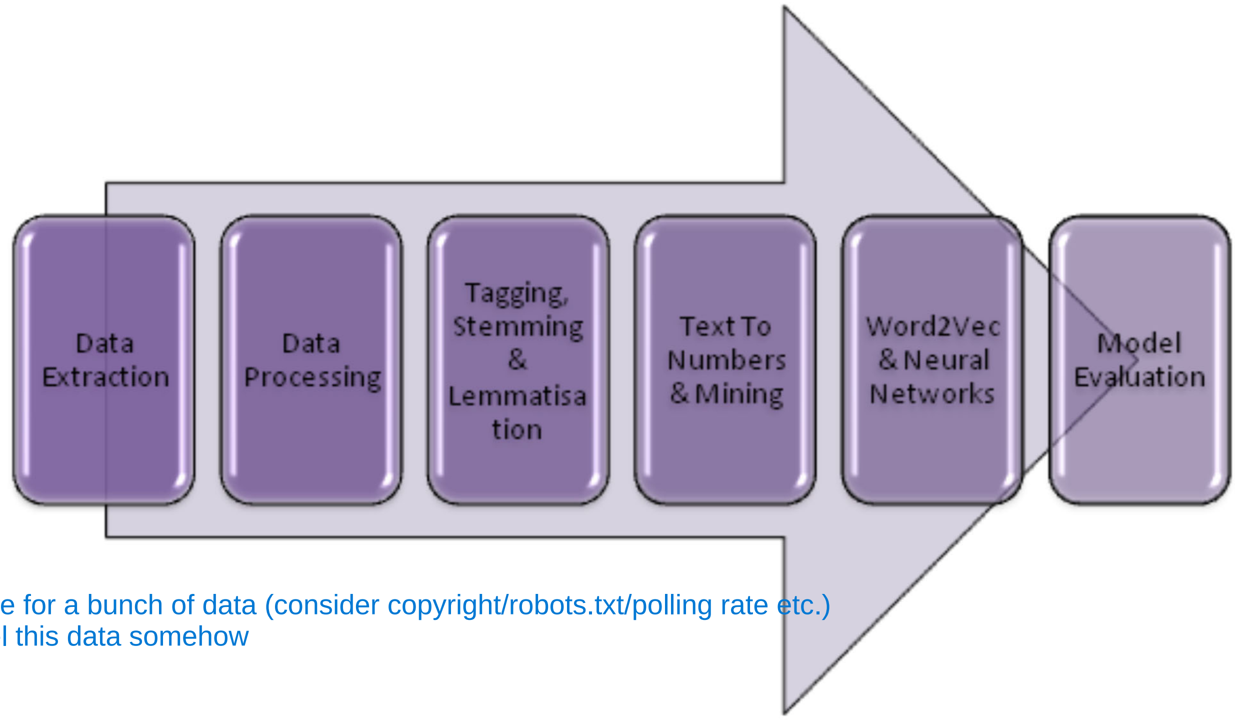
**Length:** Up to 3000 words written document, excluding code, references, and output

## Overview

This assignment involves building a prototype NLP solution using web scraping and machine learning. The initial part of the NLP solution is gathering data using a web scraper. The web scraper collects information from relevant websites and supplements that website data with metadata from additional knowledge databases (if needed). Once the data for the NLP solution is gathered, the data need to be processed, cleaned, and normalised.

A part of modern text normalisation is using machine learning are word embeddings. Word embeddings are a type of word representation that allows words with similar meaning to have a similar representation. Word embeddings are a distributed representation for text that is perhaps one of the key breakthroughs for the impressive performance of machine learning methods on challenging natural language processing problems.

To assist a development team integrating your WebCrawler and machine learning task, you will need to publish your documentation and code in a Git-repository.



- scrape for a bunch of data (consider copyright/robots.txt/polling rate etc.)
- model this data somehow

## Learning outcomes

- Apply NLP data science skills, knowledge, and techniques to solve problems in data science NLP projects with a focus on web crawler and content extraction from webpages.
- Apply NLP tasks in Python
- Understand how to deploy data science projects into production pipelines

BeautifulSoup

## Deliverables

For this assessment, you are to produce a report detailing all four tasks AND a Jupyter Notebook file with the final version of the Python code used.

## Tasks

This assessment comprises of four tasks

1. Defining of a single issue to be investigated or address using NLP methodologies
2. Sourcing data from webpages and supplementing data from knowledge sources relevant to the issue
3. Data wrangling: Cleaning, normalisation, feature extraction of the sourced data. Normalisation <u>may</u> include applying a word embedding algorithm.
4. Modelling using machine learning and valuation of the model.

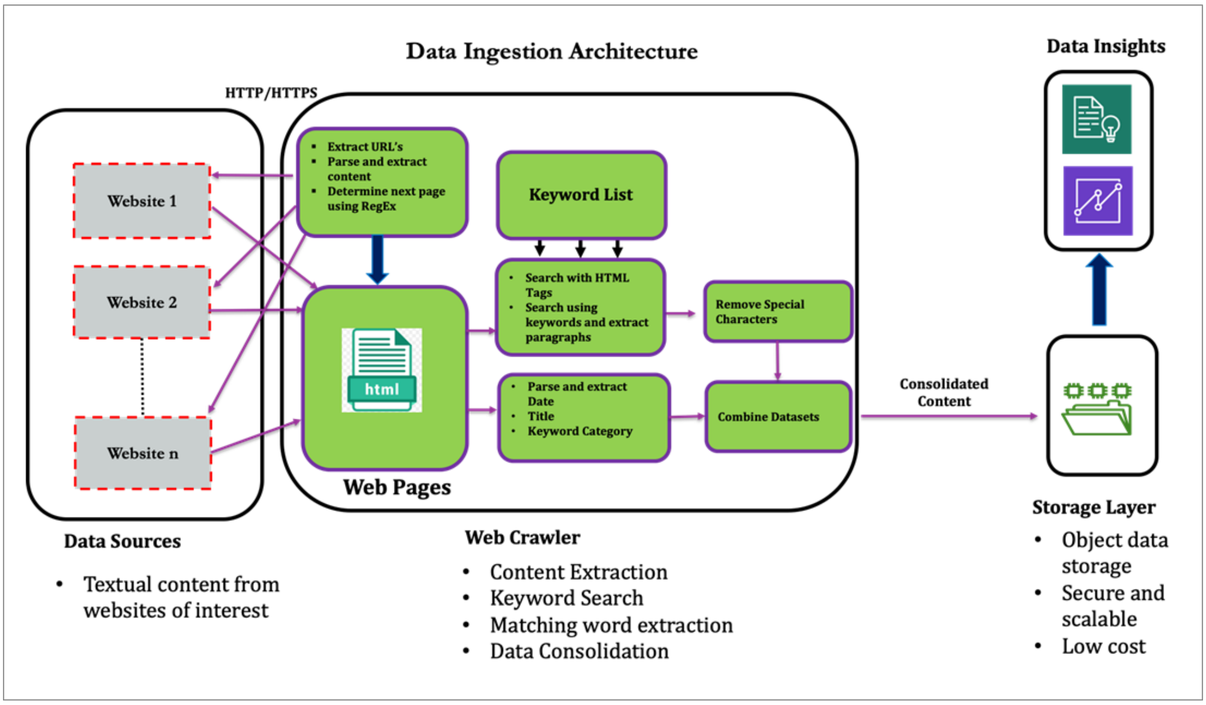


*Figure 1 https://aws.amazon.com/blogs/apn/gathering-market-intelligence-from-the-web-using-cloud-based-ai-and-ml-techniques/*

## Task Descriptions

Task 1. **Overview**: Length: < 200 words (excluding code and references)

a. An overview of the Issue
b. Where the Issue is present on the world wide web
c. How machine learning can be applied to provide a solution to the Issue

Task 2. **WebCrawler**: Length < 500 words (excluding code and references) Detailing

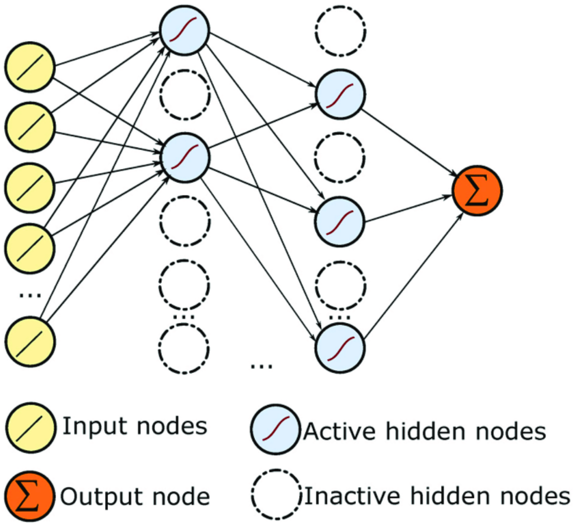
a. Websites consumed.
b. Website/data copyright considerations
c. Methodology of applying the web crawler/scraper
d. Limitations of the WebCrawler and the harvested data.
e. Methodology of storing harvested data

Task 3. **Data Wrangling** Length < 500 words (excluding code and references) Detailing:

a. Cleaning, normalisation, feature extraction of the sourced data. Normalisation may include applying a word embedding algorithm
b. Summary and visualisation of the harvested data. Preliminary EDA is acceptable in this section as well.

Task 4. **Machine Learning** Length < 800 words (excluding code and references) Detailing:

a. Specification and justification of the implementation of the ML model
b. Evaluation and visualisation of the machine learning model performance
c. Effect of the data limitations and sampling biases on the machine learning performance



**Word lengths are recommendations and may change relative to your reporting needs.**

## Permitted guidelines for web scraping

1. **Public data only:** Available to anyone on the web where nothing in the data is behind any kind of walled garden, pay or otherwise.
2. **Previously allowed:** Some sites that have tacitly accepted that scraping occurs. For example, some services are openly acknowledged that this occurs (e.g. media intelligence and media monitoring).
3. **Non-copyright-protected content:** The data involved appears to mostly, if not exclusively, be facts and information not protectable under copyright.

**Permitted use of copyright-protected:** If the site has a copyright protection notice, then the material scraped must be within the permissible use. Normally there is a standard notice on a website that will allow to download, display, print and reproduce its material in unaltered form only, provided that appropriate acknowledgment is made for your personal, non-commercial use. Take, for example, James Cook University website copyright and terms of use. James Cook University's copyright states that using a reading list for metadata analysis would be possible as long as an appropriate acknowledgement is made

## NOTES: Size of Corpus

The NLP system is a prototype so the number of documents in the corpus will be limited in size.  However, the size of the corpus will need to be sufficient to demonstrate the issue and to calculate quality metrics.  As an indicative guide, the number of documents in your corpus will depend on the length of the documents.

- **Small** length documents such as social media posts, posts on discussion boards or phone text messages, you can expect to have 500 to 1000 documents in your corpus.
- **Medium** length documents such as online new articles or extracts from reports (or long documents) you can expect to have 100 to 300 documents in your corpus.
- **Long** length document such as complete company reports, you can expect to have 50 to 200 documents in your corpus.

## NOTES: CloudFlare

Website may use technologies that actively prohibit web scraping to protect IP or to mitigate potential website downtime due to denial of service (DOS).  Web scrapers and web crawlers can cause DOS outcomes.  CloudFlare is a very common technology that is used to keep a website operating by preventing headless web browsing scraping, like Selenium and Scrapy.

You can check if a website is protected by CloudFlare at sites like http://www.doesitusecloudflare.com/

## Assessment submission guidelines

Use MS Word or PDF for the written report.

Your submission for Assessment 3 should be uploaded to LearnJCU as two (2) separate files:

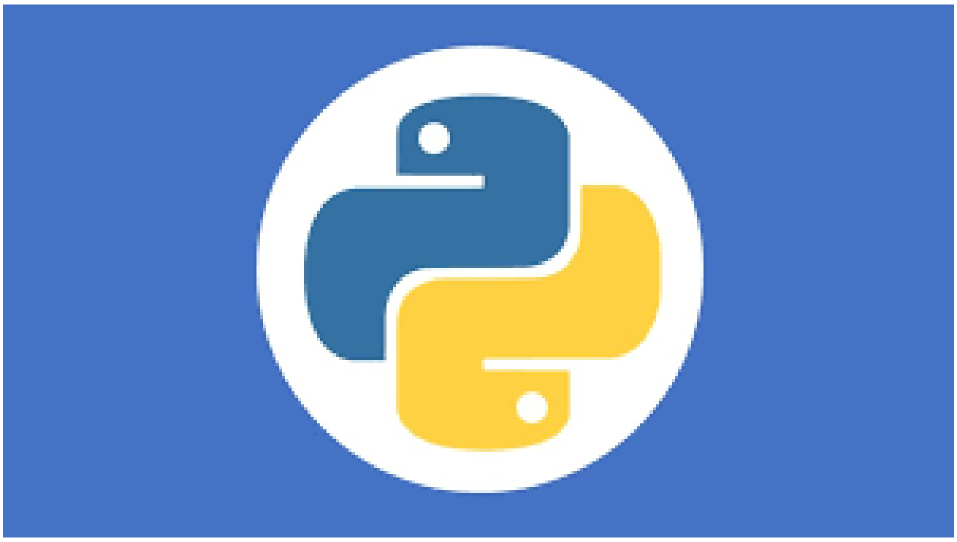
**File 1 – the written report. File 2 the Jupyter Notebook.** Your report meeting following requirements:

- Filename: A3_firstname_lastname.pdf (or *.ipynb)
- 12pt font size with single line spacing (preferred)
- APA referencing style applied (preferred)

You may upload as many times as you want, but only the last submission is graded.

## Important note

The **entire project** must be accomplished using **Python**. Any calculations, visualisations, results and so on produced using software other than Python (e.g. R, Excel, Tableau etc.) is **not** accepted and, therefore, will not be assessed. The code itself must be prepared using **Python either as a script in notebook form or standalone Python files**. Refusal to comply with these requirements will result in your work being considered as **not delivered**.



## Marking criteria. Task 1: Overview 10% of Overall grade

| Criteria | High Distinction / Distinction: Sophisticated/Exceeds Expectations (75-100%) | Credit /Pass: Above/Meets Expectations (50-74%) | Fail: Unsatisfactory / Below Expectations (0-49%) |
|---|---|---|---|
| **Overview**<br>**100% of section grade** | Identifies and discusses:<br>• The Issue<br>• Where the Issue is present on the world wide web, with linkages to how the chosen domains could be expanded<br>• How machine learning can be applied to provide a solution to the Issue with a brief literature review of peer reviewed literature relevant to the chosen NLP machine learning task;<br>Discussions are specific and targeted towards clearly identified a NLP task. Discussions are supported with credible references sources. | Identifies and discusses:<br>• The Issue<br>• Where the Issue is present on the world wide web<br>• How machine learning can be applied to provide a solution to the Issue<br>Discussions are in a general nature of NLP tasks routine data science related situation. | Partially identifies and/or explains some key issues in a superficial data science related situation |

Rubric continued next page

## Marking criteria.  Task 2: WebCrawler 30% of Overall grade

| Criteria | High Distinction / Distinction:  Sophisticated/Exceeds Expectations (75-100%) | Credit /Pass: Above/Meets Expectations (50-74%) | Fail: Unsatisfactory / Below Expectations (0-49%) |
|---|---|---|---|
| **Domains**<br>**25% of section grade** | Identifies and discusses with justifications:<br>• Website URLs to be crawled with consideration of: coverage of the chosen domains on the issue relative to the www; limitations of the consumed domains with linkages to sampling design and ethical considerations<br>• Copyright of the chosen domains and linkages to appropriate legal frameworks<br>• The Natural Language data, meta-data, or other data on each domain and how these data align to the issue<br>Discussions are in a complex data science related situation, drawing upon relevant theory from a wide range of credible sources; eliciting insightful knowledge linking to broader relationships and, bring in originality of perspective | Identifies and discusses:<br>• Website URLs to be crawled<br>• Copyright of the chosen domains<br>• The type of Natural Language data used in the domains.<br>Discussions are general in nature and identify most criteria | Partially identifies and/or explains some key issues in a superficial data science related situation |
| **WebCrawler workflow**<br>**75% of section grade** | Identifies and discusses with justifications:<br>• Technology components used for the web crawler with comparisons to other similar technology components<br>• Complexity of the domains and where the targeted data resides<br>• Methodology and sequencing of the crawler(s), using the complexity, data structures and website access restrictions to optimise the crawler<br>• Data storage<br>Discussions are in a complex data science related situation, drawing upon relevant theory from a wide range of credible sources; eliciting insightful knowledge linking to broader relationships and, bring in originality of perspective | Identifies and discusses:<br>• Technology components used for the web crawler<br>• Where the targeted data resides on the domains<br>• Methodology and sequencing of the crawler(s)<br>• Data storage<br>Discussions are in a routine data science related situation, using code extracts in discussions and demonstrations, drawing upon relevant theory | Partially identifies and/or explains some key issues in a superficial data science related situation |

Rubric continued next page

## Marking criteria. Task 3: Data Wrangling. 20% of Overall grade

| Criteria | High Distinction / Distinction: Sophisticated/Exceeds Expectations (75-100%) | Credit /Pass: Above/Meets Expectations (50-74%) | Fail: Unsatisfactory / Below Expectations (0-49%) |
|---|---|---|---|
| **Data Wrangling**<br>**50% of section grade** | Identifies and discusses with justifications:<br>• Corpus data wrangling methods that begin to feature engineer towards the intended NLP task<br>• Feature extraction appropriate to the intended NPL task<br>• Hyperparameters of the feature extraction task<br>• Generation of an appropriate training and test sets with reference to any sample distributions, biases and or data limitations<br>•<br>Discussions are in a complex data science related situation, drawing upon relevant theory from a wide range of credible sources; eliciting insightful knowledge linking to broader relationships and, bring in originality of perspective | Identifies and discusses:<br>• Cleaning and normalisation of the corpus<br>• Feature extraction appropriate to the intended NPL task<br>Discussions are in a routine data science related situation, using code extracts in discussions and demonstrations, drawing upon relevant theory<br>• | Partially identifies and/or explains some key issues in a superficial data science related situation |
| **Data Summarisation**<br>**50% of section grade** | Identifies and discusses:<br>• Visualisation and interpretation of sample distribution<br>• Visualisation and interpretation of corpus<br>• Descriptive statistics of both the sample and the corpus<br>• Corpus limitations<br>• Sampling biases<br>Discussion of the corpus are inclusive of population sampling considerations and population strata.<br>Discussions, visualisations and tabulations contain linkages to sampling design and limitations/design features of the web crawler. Discussions elicit insightful knowledge linking to broader relationships and, bring in originality of perspective | Identifies and discusses:<br>• Summary of the generated corpus<br>• Visualisation of the corpus<br>• Descriptive statistics of the corpus<br>Discussions are in a routine data science related situation, using code extracts in discussions and demonstrations, drawing upon relevant theory | Partially identifies and/or explains some key issues in a superficial data science related situation |

Rubric continued next page

## Marking criteria: Task 4: Machine Learning. 30% of Overall grade

| Criteria | High Distinction / Distinction: Sophisticated/Exceeds Expectations (75-100%) | Credit /Pass: Above/Meets Expectations (50-74%) | Fail: Unsatisfactory / Below Expectations (0-49%) |
|---|---|---|---|
| **Machine learning Structure**<br>**50% of section grade** | Identifies and discusses with justifications:<br>• Structure of the machine learning<br>• Hyperparameters of the machine learning algorithm<br>• Computation environment<br>Discussions are in a complex data science related situation, drawing upon relevant theory from a wide range of credible sources; eliciting insightful knowledge linking to broader relationships and, bring in originality of perspective | Identifies and discusses:<br>• Structure of the machine learning<br>• Hyperparameters of machine learning algorithm<br>• Computation environment<br>Discussions are in a routine data science related situation, drawing upon relevant theory | •<br>Partially identifies and/or explains some key issues in a superficial data science related situation |
| **Evaluation**<br>**50% of section grade** | Identifies and discusses:<br>• Detailed evaluation of the machine learning performance<br>• Visualisation of the model performance<br>• Detailed effects of the data limitations and sampling biases on the machine learning model performance<br>Discussions are in a complex data science related situation, highlights potential downstream effects related to data distribution, missing data, or data biases. Discussions elicit insightful knowledge linking to broader relationships and, bring in originality of perspective | Identifies and discusses:<br>• Preliminary evaluation of the machine learning performance<br>• Visualisation of the model performance<br>• Some effects of the data limitations and sampling biases on the machine learning model performance<br>Discussions are in a routine data science related situation, using code extracts in discussions and demonstrations, drawing upon relevant theory | Partially identifies and/or explains some key issues in a superficial data science related situation |

Rubric continued next page

## Marking criteria: Reporting and Coding 10% of Overall grade

| Criteria | High Distinction / Distinction: Sophisticated/Exceeds Expectations (75-100%) | Credit /Pass: Above/Meets Expectations (50-74%) | Fail: Unsatisfactory / Below Expectations (0-49%) |
|---|---|---|---|
| **Report**<br>**33% of section grade** | • Sequencing of sections logical and coherent. No out of sequence material or discussions.<br>• Output results, code, figures appear in the sections where initially discussed<br>• Grammar and spelling errors are rare<br>• Internal cross referencing always used<br>• External referencing style appropriate | • Sequencing of sections logical and coherent. Some out of sequencing of content.<br>• Output results, code, figures appear in the sections where initially discussed<br>• Grammar and spelling contain some errors<br>• Internal cross referencing sometimes used<br>• External referencing style appropriate | • Sequencing of sections routinely illogical and/or incoherent, frequent out of sequencing of content.<br>• Output results, code, figures routinely do not appear in the sections where initially discussed<br>• Grammar and spelling contain frequent errors<br>• Internal cross referencing rarely/not used<br>• External referencing style inappropriate |