Assessment 3: WebCrawler and NLP System

Type: Written document and Jupyter Notebook

Weight: 50%

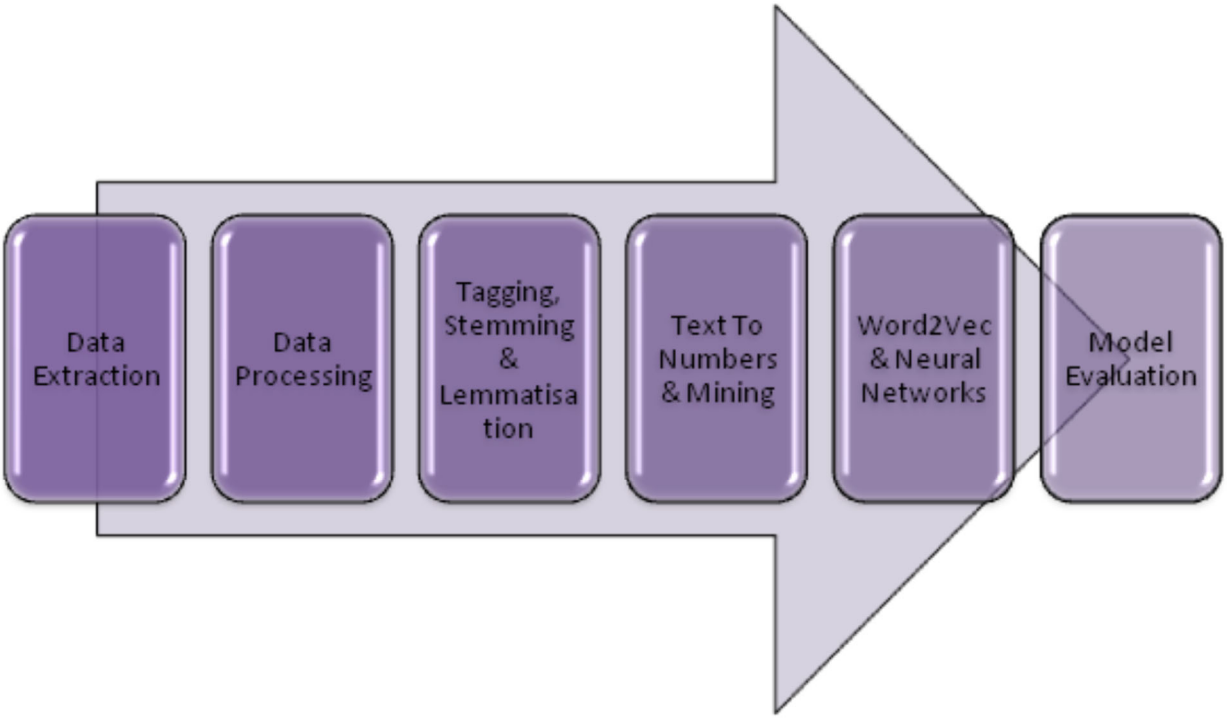
Length: Up to 5000 words written document, excluding code, references, and output

Overview

This assignment involves building a prototype NLP solution using web scraping and deep 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 deep 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 deep learning methods on challenging natural language processing problems.

Using the normalised data from the word embeddings, the NLP issue derived from the web scrape is modelled using a neural network. Forward feed neural networks are common neural networks used in NLP tasks. To assist a development team integrating your WebCrawler and deep learning task, you will need to publish your documentation and code in a Git-repository.



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

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 may include applying a word embedding algorithm.
4. Modelling using a neural network and evaluation of the model.

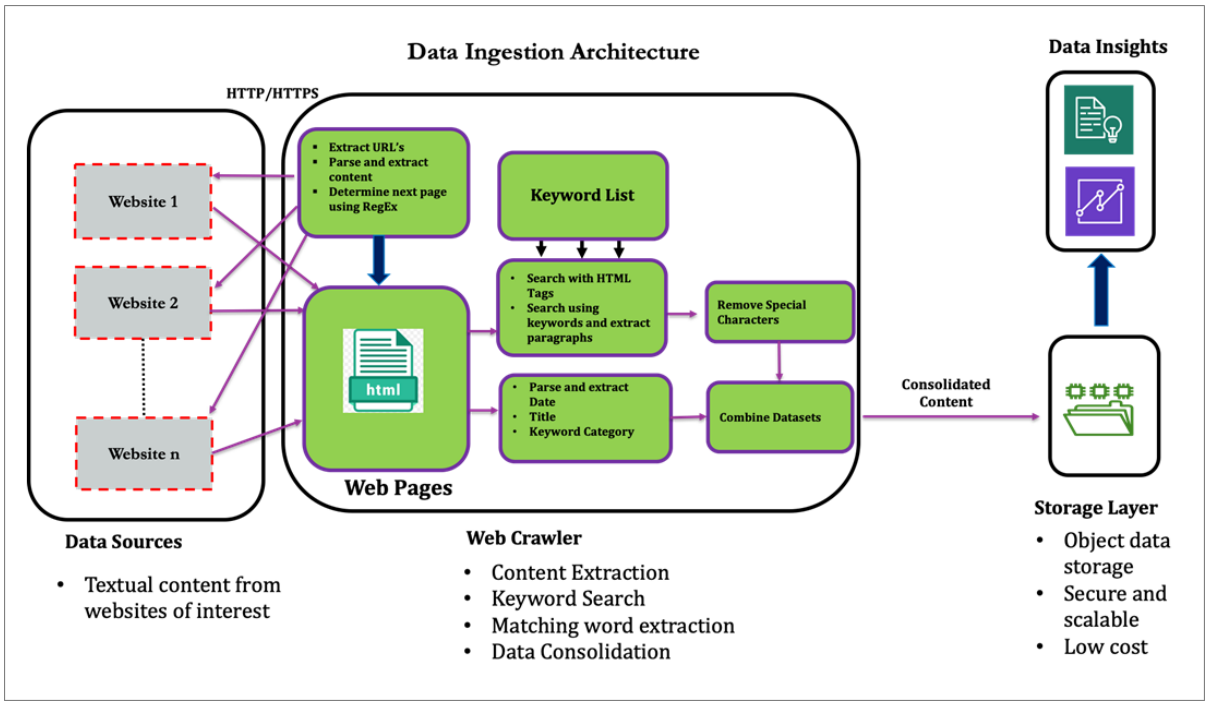


Figure 1 https://aws.amazon.com/blogs/apn/gathering-market-intelligence-from-the-web-using-cloud-based-ai-and-ml-techniques/

Task Descriptions

Task 1. **Overview**: Length: < 500 words (excluding code and references)

a. An overview of the Issue
b. Where the Issue is present on the world wide web
c. How deep learning can be applied to provide a solution to the Issue
d. Brief literature review of peer reviewed literature relevant to the chosen NLP deep learning task

Task 2. **WebCrawler**: Length < 1500 words (excluding code and references)

Detailing

a. Websites to be consumed
b. A rationale for extracting the web content
c. Content coverage of the data extracted
d. Methodology of applying the web crawler
e. Website/data copyright considerations
f. Metadata supplementation and rational for the supplementation
g. Limitations of the WebCrawler and the harvested data.
h. Methodology of storing harvested data

Task 3. **Data Wrangling** Length < 1500 words (excluding code and references)

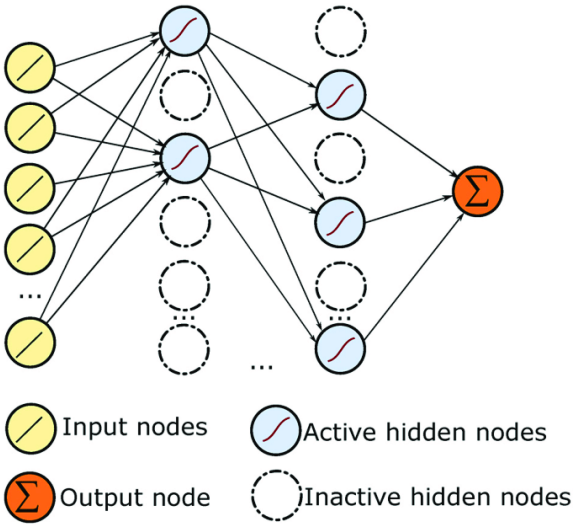
Detailing:

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. **Deep Learning** Length < 1500 words (excluding code and references)

Detailing:

a. Specification and justification of any hyperparameters
b. Detailed evaluation and visualisation of the neural network performance
c. Effect of the data limitations and sampling biases on the deep 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:

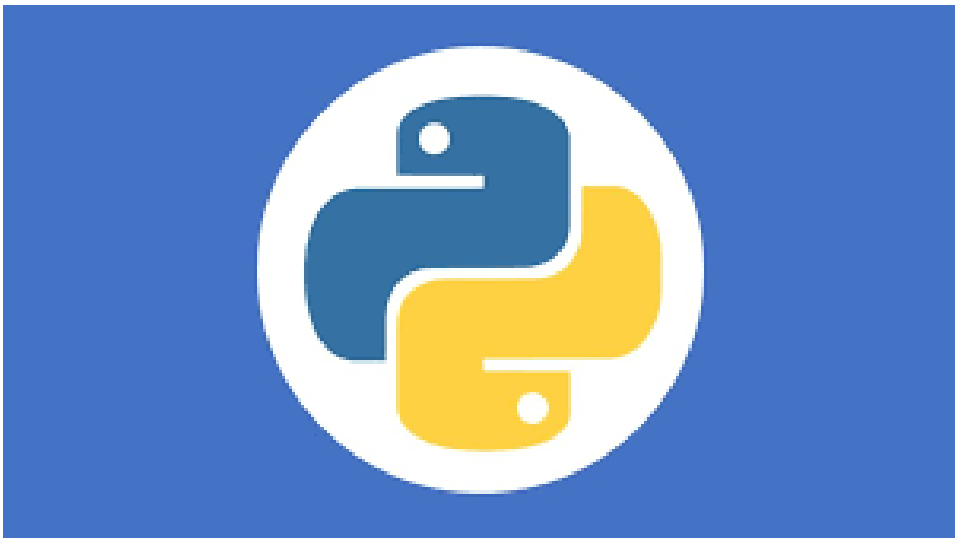
File 1 – the written report. File 2 the Jupyter Notebook. Your report meeting following requirements:

- 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 **100% of section grade**	Identifies and discusses: • The Issue • Where the Issue is present on the world wide web, with linkages to how the chosen domains could be expanded • How deep learning can be applied to provide a solution to the Issue • Brief literature review of peer reviewed literature relevant to the chosen NLP deep learning task; with linkages between the word embedding, data sources and/or NLP task(s) Discussions elicit insightful knowledge linking to broader relationships and, bring in originality of perspective	Identifies and discusses: • The Issue • Where the Issue is present on the world wide web • How deep learning can be applied to provide a solution to the Issue • Brief literature review of peer reviewed literature relevant to the chosen NLP deep learning task Discussions are in a routine data science related situation.	Partially identifies and/or explains some key issues in a superficial data science related situation

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 **25% of section grade**	Identifies and discusses with justifications: • Website URLs to be crawled • Coverage of the chosen domains on the issue • Limitations of the WebCrawler domains with linkages to sampling design and ethical considerations • The Natural Language data, meta-data, or other data on each domain and how these data align to the issue • Copyright of the chosen domains and linkages to appropriate legal frameworks 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: • Website URLs to be crawled • Coverage of the chosen domains on the issue • Limitations of the WebCrawler domains • Copyright of the chosen domains • Metadata supplementation and rational for the supplementation Discussions are in a routine data science related situation, drawing upon relevant theory	Partially identifies and/or explains some key issues in a superficial data science related situation
WebCrawler workflow **75% of section grade**	Identifies and discusses with justifications: • Technology components used for the web crawler • Complexity of the domains and where the targeted data resides • Methodology and sequencing of the crawler(s), using the complexity, data structures and website access restrictions to optimise the crawler • Data storage 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: • Technology components used for the web crawler • Complexity of the domains and where the targeted data resides • Methodology and sequencing of the crawler(s) • Data storage 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

Marking criteria. Task 3: Data Wrangling. 15% 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 **50% of section grade**	Identifies and discusses with justifications: • Cleaning and normalisation of the corpus • Feature extraction appropriate to the intended NPL task • Hyperparameters of the feature extraction task 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: • Cleaning and normalisation of the corpus • Feature extraction appropriate to the intended NPL task • Hyperparameters of the feature extraction task 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
Data Summarisation **50% of section grade**	Identifies and discusses: • Summary of the generated corpus • Visualisation of the corpus • Descriptive statistics of the corpus Discussion of the corpus are inclusive of population sampling considerations and population strata. 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: • Summary of the generated corpus • Visualisation of the corpus • Descriptive statistics of the corpus 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

Marking criteria: Task 4: Deep 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%)
Deep learning Structure **50% of section grade**	Identifies and discusses with justifications: • Structure of deep learning layers • Activation functions • Loss function • Hyperparameters of deep learning algorithm • Computation environment 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: • Structure of deep learning layers • Activation functions • Loss function • Hyperparameters of deep learning algorithm • Computation environment Discussions are in a routine data science related situation, drawing upon relevant theory	• Partially identifies and/or explains some key issues in a superficial data science related situation
Evaluation **50% of section grade**	Identifies and discusses: • Detailed evaluation of the deep learning performance • Visualisation of the model performance • Detailed effects of the data limitations and sampling biases on the deep learning model performance 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: • Preliminary evaluation of the neural network performance • Visualisation of the model performance • Some effects of the data limitations and sampling biases on the deep learning model performance 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

Marking criteria: Reporting and Coding 15% 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 **33% of section grade**	• Sequencing of sections logical and coherent. No out of sequence material or discussions. • Output results, code, figures appear in the sections where initially discussed • Grammar and spelling errors are rare • Internal cross referencing always used • External referencing style appropriate	• Sequencing of sections logical and coherent. Some out of sequencing of content. • Output results, code, figures appear in the sections where initially discussed • Grammar and spelling contain some errors • Internal cross referencing sometimes used • External referencing style appropriate	• Sequencing of sections routinely illogical and/or incoherent, frequent out of sequencing of content. • Output results, code, figures routinely do not appear in the sections where initially discussed • Grammar and spelling contain frequent errors • Internal cross referencing rarely/not used • External referencing style inappropriate
Code **67% of section grade**	Demonstrates • Python code extracts use appropriate PEP8 and PEP 256 code format • Robust, fault tolerant coding practices • Efficient coding practices • Use of relevant python packages • Code contains comments and markdown explanations for all key sequences. • Coding of variables names used as part of the commenting process • Code management tools, such as GIT, used.	Demonstrates • Python code extracts use appropriate PEP8 and PEP 256 code format • Use of relevant python packages • Code contains comments for all key sequences.	Demonstrates • Python code extracts use rarely follow PEP8 and PEP 256 code format • Code rarely contains comments • Key code sequencing elements rarely identified