Assessment 3: WebCrawler and NLP System

Type: Written document and video presentation

Due: 11:59 PM AEST 30 May 2021

Weight: 60%

Length: 3500 words +/-10% written document, excluding code, references and output AND 3 minute +/- 30 second video presentation.

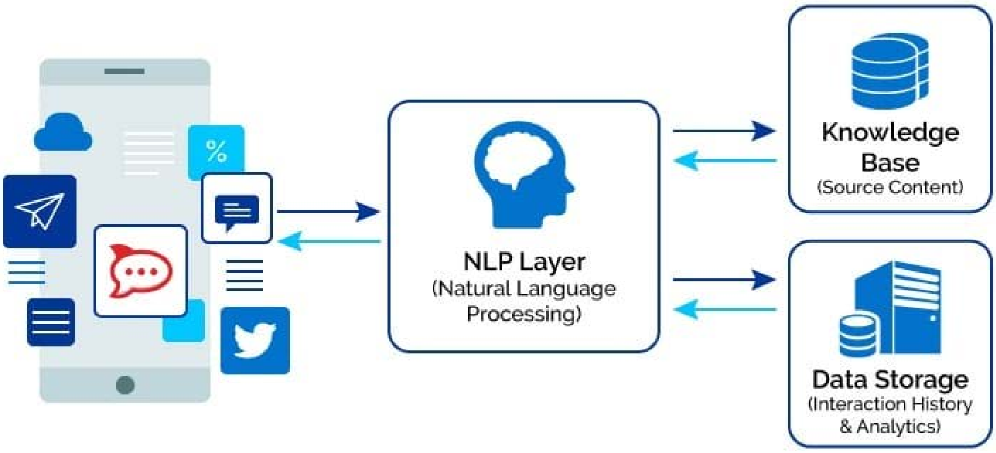
Knowledge Base
(Source Content)
NLP Layer
(Natural Language Processing)
Data Storage
(Interaction History & Analytics)

Figure 1 https://research.aimultiple.com/natural-language-platforms/

Overview

This assignment involves building parts of a prototype NLP solution and investigating the deployment of the NLP solution that could be used by a development team. 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. Once the data for the NLP solution is gathered, the data need to be processed and cleaned for NLP tasks.

The NLP tasks use the harvested data to solve or investigate a wider NLP issue. Here, you are to apply two NLP tasks, such as Name Entity Recognition, Semantic analysis, Sentiment analysis, etc. The output of one of the NLP tasks may even be one of the inputs to the other NLP task. Once the two NLP tasks and web crawler have been written and documented, these project components need to be made accessible to a wider development team.

To assist a development team integrating your WebCrawler and NLP tasks, you will need to publish your documentation and code in a Git-repository. Additionally, you will provide a high-level, end-to-end practical data science solution design video presentation.

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
- Effectively communicate the results of the project as a video

Work-based skills

- The ability to extract key values (e.g. structured data) from HTML
- Image, video, and text recognition (e.g. unstructured data).
- Undertake applied industry research

Background

Many NLP solutions are comprised of multiple NLP tasks jointly working together. The auto complete function in a programming IDE is an example where there is an NLP NER task identifying the names of the variables and functions which feeds into an NLP semantic parser to trap simple errors before compiling/running. Another example is website forms. On many web forms, address matching and prediction NLP tasks jointly run on incomplete user input to suggest addresses. Predictive addressing is a useful tool to ensure accurate user input and mitigates human errors.

You don't need to produce an end-to-end solution in this assignment task but you need to clearly indicate what the targeted problem is, the adoption issues that a wider solution might have and how your project can reasonably contribute towards addressing it using NLP tasks.

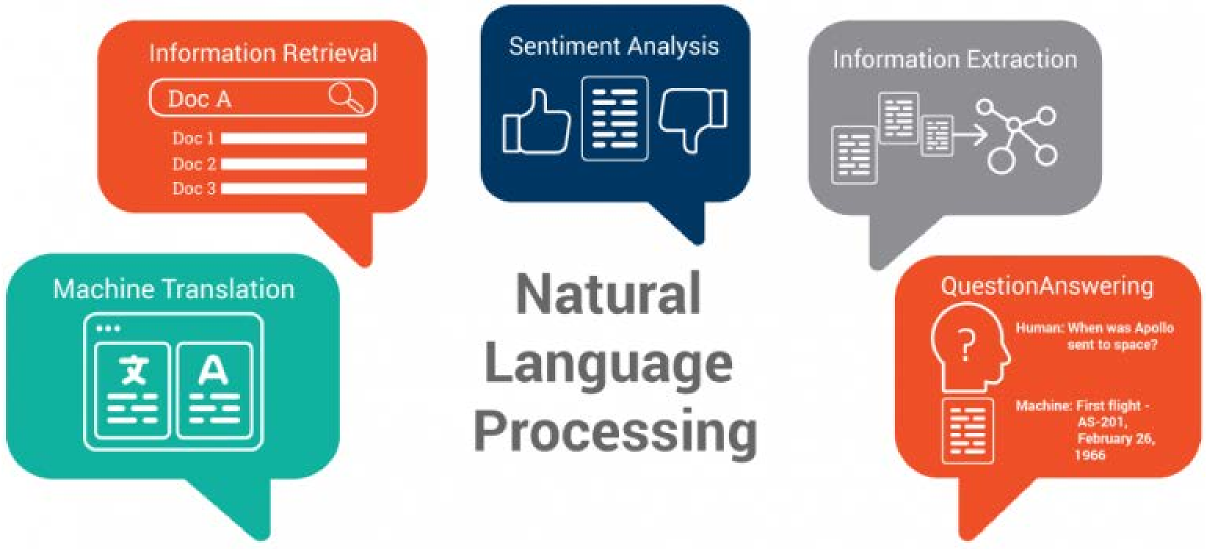

Figure 2 https://www.ontotext.com/blog/top-5-semantic-technology-trends-2017/

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. Developing proto-type NLP tasks that contribute towards addressing the issue
4. Publishing and communicating your contributions to a development team

Deliverables

For this assessment, you are to produce a set of three Markdown documents and a three-minute video presentation which are aligned to the assessment tasks.

Markdown documents | Tasks 1-3

Document 1. **Overview | Task 1**: An overview of the issue to be investigated and how the WebCrawler and NLP tasks align to the issue. Length: < 500 words.

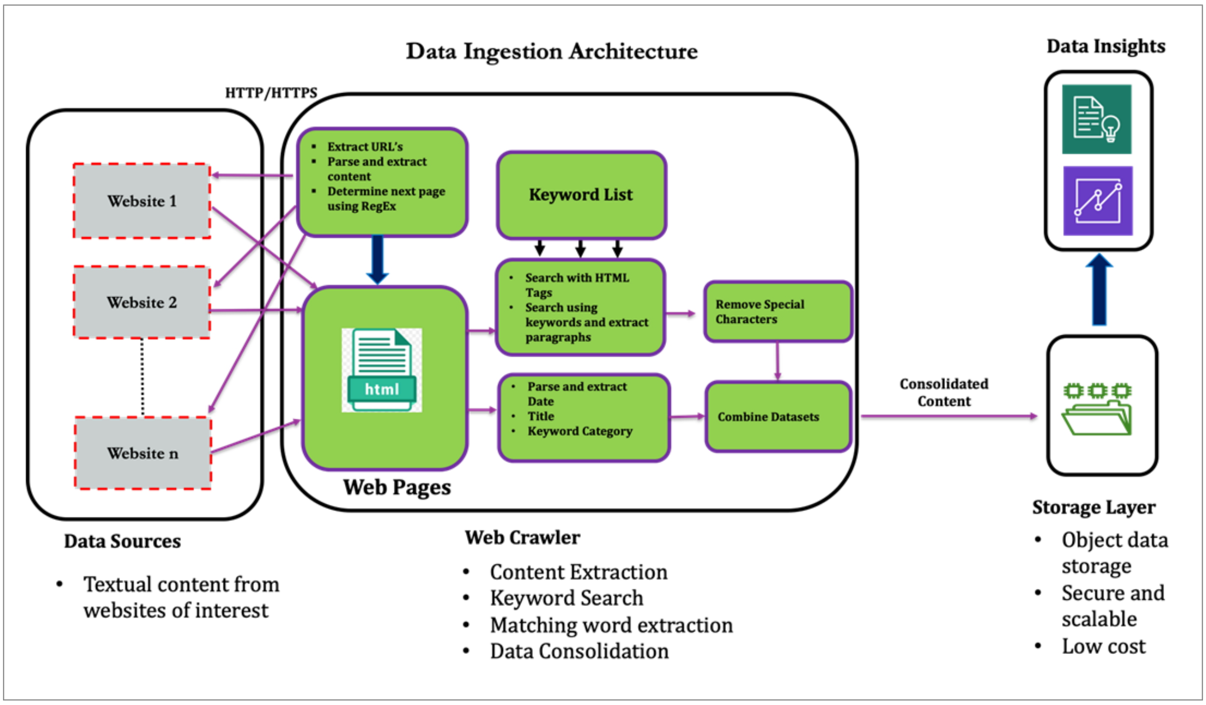

Figure 3 https://aws.amazon.com/blogs/apn/gathering-market-intelligence-from-the-web-using-cloud-based-ai-and-ml-techniques/

Document 2. **WebCrawler | Task 2**: 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. Complexity of the content layout
e. Website/data copyright considerations
f. Metadata supplementation and rational for the supplementation
g. Content extractor to export the important aspects of the data and/or metadata
h. Relevant python coding

i. Demonstration of the application of the WebCrawler (i.e. screen shots)
j. Methodology of processing, cleaning, and storing harvested data for NLP tasking
k. Summary and visualisation of the harvested data. Preliminary EDA is acceptable in this section as well.

The WebCrawler activity must adhere to the Permitted guidelines for web scraping (pg 5).

Document 3. **Proto type NLP Tasks | Task 3:** Length < 1500 words (excluding code and references). For each NLP task, provide:

a. Brief literature review of the NLP task
b. Rational for selection of the NLP task
c. Data pre-processing of inputs and outputs, separate from the WebCrawler harvesting
d. Specification and justification of hyperparameter
e. Preliminary assessment of NLP Task performance
f. Code

Video Presentation | Task 4

The video presentation, of length 3 minutes ± 30 seconds, is a short summary of:

a. Ther online repository your deliverables and code are published, and which framework modification of the code can be made.
b. Examples of code management using your code repository
c. Limitations of the WebCrawler, harvested data and the NLP tasks

The video presentation can be a screen recording or an orally annotated PowerPoint. There is no requirement to include a video of yourself – just your voice.

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

Assessment submission guidelines

If you use MS Word or any other program, save your work as a PDF for submission.

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

File 1, 2 and 3: Markdown documents for Deliverables 1, 2 and 3. Your report meeting following requirements:

- Filename: A3_DocumentNumber_X_firstname_lastname.pdf
- Length: 1000 words (+/-10%)
- 12pt font size with 1.5 spacing
- APA referencing style applied.

File 4: Video presentation. If the video presentation is an orally annotated PowerPoint, then submit the PowerPoint file. If the video is another format, publish the video on YouTube as a private file using the JCU guide to sharing YouTube video (https://www.jcu.edu.au/__data/assets/pdf_file/0005/1132493/How-to-Upload-Edit-and-Share-YouTube-Video.pdf). Submit a pdf document detailing the URL for the assessor to access the YouTube video.

There is no specific requirement on the tools to make your recording – you may use any that you are comfortable with. Some suggestions are PowerPoint (recording audio and visual), OBS (Open Broadcasting Software), Zoom, and the default media recorders for Mac OS or Windows. Videos can be edited prior to uploading to YouTube.

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: MA 3831 Assignment Document 1: Overview

Total marks 10	Exemplary (100% marks)	Good (75% marks)	Competent (50% marks)	Limited (25% marks)	Needs work (0% marks)
Proposed High-level Solution design (10 marks)	Clearly articulates a high-level solution design to explain the targeted outcome using an NLP-based data science pipeline solution approach.	Exhibits aspects of columns to the left and right	Partially articulates a high-level solution design to explain the targeted outcome using an NLP-based data science pipeline approach, but the explanation of the approaches and system components is not clear in places.	Exhibits aspects of columns to the left and right	Unclear, does not give an overview of the solution design

Rubric continued next page

Marking criteria: MA 3831 Assignment Document 2: WebCrawler

Total marks 25	**Exemplary (100% marks)**	**Good (75% marks)**	**Competent (50% marks)**	**Limited (25% marks)**	**Needs work (0% marks)**
Website selection (5 marks)	Clear justification for selecting a website from where the student will extract web content to prepare a text corpus. The justification addresses both content coverage and complexity of the content layout along with the copyright issue succinctly.	Exhibits aspects of columns to the left and right.	Partial or non-pragmatic justification for selecting a website from which to extract web content to prepare a text corpus. The justification is unclear in addressing both content coverage and complexity of the content layout along with copyright issue.	Exhibits aspects of columns to the left and right.	Very weak justification for selecting a website from which to extract web content to prepare a text corpus. The justification lacks addressing of both content coverage and complexity of the content layout along with copyright issue.
Description of the web crawler workflow (5 marks)	Clearly articulates to the reader an overview of the crawler being developed and explains different components of the web crawler with a clear justification of using any particular Python package or framework to build the crawler.	Exhibits aspects of columns to the left and right.	Partially articulates to the reader an overview of the crawler being developed, but the explanation of different components of the developed web crawler is not clear. The justification of using any particular Python package or framework to build the crawler is not addressed clearly or only addressed partially.	Exhibits aspects of columns to the left and right.	Fails to provide a clear overview of the crawler being developed and the explanation of different components of the developed web crawler is limited or poorly addressed. The justification of using any particular Python package or framework to build the crawler is not addressed at all or addressed poorly.
Data extraction, collection method and description of corpus (10 marks)	Provides clear and easy-to-follow background of the data extracts. Clearly explains the suitability of data as a text corpus and method by which the target data is structured within HTML tree or document object model (DOM). Where applicable, clearly explains: • The process where unstructured data is structured via HTML/DOM • The process where target data (from multi-page or multi-source) is extracted and saved into files by the built crawler • The data schema and format for final output	Exhibits aspects of columns to the left and right.	Provides a partial and limited background of the data extracts. Partial or limited explanation of the suitability of data as a text corpus and method of which the target data is structured within HTML tree or DOM. Where applicable, provides a partial explanation and some example snippets of: • The process where unstructured data is structured via HTML/DOM • The process where target data (from multi-page or multi-source) is extracted and saved into files by the built crawler • The data schema and format for final output	Exhibits aspects of columns to the left and right.	Little to no information and/or explanation of the background of the data extracts. Unsuitable/no explanation of the suitability of data as a text corpus and method of which the target data is structured within HTML tree or DOM. Where applicable, provides little to no explanation or example snippets of: • The process where unstructured data is structured via HTML/DOM • The process where target data (from multi-page or multi-source) is extracted and saved into files by the built crawler • The data schema and format for final output
Working crawler code with screenshots (5 marks)	Provides a working crawler code with instructions to run on the reader's machine and reader is able to run it. Evidence of a successful run is shown using screenshot of your 'web crawler in action'.	Exhibits aspects of columns to the left and right.	Provides a crawler code without instructions to run on the reader's machine and/or the reader is unable to run the program. Provides limited evidence of successful run of the web crawler through the attachments of screenshots of the crawler application.	Exhibits aspects of columns to the left and right.	Provides no code and/or the reader is unable to run any code snippet. Provides no screenshots of the crawler application.

Rubric continued next page

Marking criteria: MA 3831 Assignment Document 3: NLP Tasks

Criteria task 1 (Total: 40 marks)	Exemplary (100% marks)	Good (75% marks)	Satisfactory (50% marks)	Limited (25% marks)	Poor (12.5% marks)	Incomplete (0% marks)
Proposes strategies or partial solutions (5 marks)	The solution provided is correct and solves the problem effectively.	Exhibits aspects of columns to the left and right.	The solution provided is mostly correct and partially solves the problem.	Exhibits aspects of columns to the left and right.	The solution provided is incorrect and does not solve the problem.	No solution or Python notebook or output provided.
Formal quantitative assessment of results and output (10 marks)	The write up of results and output solves the problem and is consistent with the explanation of the algorithm used.	Exhibits aspects of columns to the left and right.	The write up of results and output mostly solves the problem and is consistent with the explanation of the algorithm used.	Exhibits aspects of columns to the left and right.	The write up of results and output does not solve the problem and is inconsistent with the explanation of the algorithm used.	No solution or Python notebook or output provided.
Effective and correct use of text mining package in the solution (20 marks)	The text mining elements used are appropriate and correctly used.	Exhibits aspects of columns to the left and right.	The text mining elements used are mostly appropriate and correctly used.	Exhibits aspects of columns to the left and right.	The text mining elements used are mostly inappropriate or incorrectly used.	No solution or Python notebook or output provided.
Analysis (5 marks)	The analysis was thorough, accurate and succinct. The results clearly follow from the data collection and the methods. Description of model or experimental runs on the dataset is well explained	Exhibits aspects of columns to the left and right	The results are explained correctly, clearly and in sufficient detail most of the time. There exists a connection of some type between the results and the data collection and methods. The description of model or experimental runs on the dataset is not well explained.	Exhibits aspects of columns to the left and right	The results are not explained correctly, clearly and in sufficient detail. The connection of some type between the results and the data collection and methods is missing and not explained at all.	No solution or Python notebook or output provided

Rubric continued next page

Marking criteria: MA 3831 Assignment Video Presentation

Total marks 15	Exemplary (100% marks)	Good (75% marks)	Competent (50% marks)	Limited (25% marks)	Needs work (0% marks)
Publishing (5)	The publishing discussion and justifications are thorough, accurate and succinct. The solution is shared via a public link to your code repository with a few commits (at least 5 commits) as an evidence of code management.	Exhibits aspects of columns to the left and right	The publishing discussion is briefly described with limited justification. The solution is shared via a public link to your code repository with a single or initial commit.	Exhibits aspects of columns to the left and right	The publishing discussion is not evident and there are no justifications made. No repo link provided.
Limitations Discussion (10)	The original objectives and/or problem is restated and contrasted against the obtained achievements. Limitations of the WebCrawler harvested data and NLP tasks are highly appropriate and concise. The discussion summarises and draws a clear, effective conclusion of the investigation and enhances the impact of the report. It may also highlight other unavoidable limitations of the investigation.	Exhibits aspects of columns to the left and right	The discussion is clearly stated and connections to limitations of the WebCrawler, harvested data and NLP tasks are described.	Exhibits aspects of columns to the left and right	The discussion may not be clear and/or the connections to the work reported and limitations are incorrect or unclear.

Marking criteria: MA 3831 Reporting

Total marks 10	Exemplary (100% marks)	Good (75% marks)	Competent (50% marks)	Limited (25% marks)	Needs work (0% marks)
Presentation (5 marks)	Presents to audience with a high level of clarity using academic language and highly effective sequencing and explanatory techniques Visual aids align with the message; are concise and accurate, and falls within the allocated time allowance	Exhibits aspects of columns to the left and right	Presents in a generally clear manner using academic language and sequencing and explanatory techniques Visual aids align with the message; and falls within the allocated time allowance	Exhibits aspects of columns to the left and right	Has difficulty conveying meaning to audience due to inappropriate language and lack of sequencing and explanatory techniques Visual aids are illogical, and/or falls outside within the allocated time allowance
Written Report (5 marks)	Demonstrate a sophisticated ability to write a formal written document where: all elements are clearly, correctly, and concisely presented; Grammar and spelling contains no or rare errors, Justifications and discussions are all appropriately referenced.	Exhibits aspects of columns to the left and right	Meets expectations writing a formal written document where: all elements are correctly presented; Grammar and spelling contains some/few errors. Justifications and discussions are sometimes appropriately referenced.	Exhibits aspects of columns to the left and right	Below expectations writing a formal written document where: some elements are missing; Grammar and spelling errors are frequent; Justifications and discussions are not appropriately referenced.