MA5851 Assessments – A3 debrief

- Assessment 3 (50%) – WebCrawler and NLP System
  - Deliverables
  - Tasks
  - Marking Rubric

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."

- Document 1/Part 1 - Overview
- Task 1: An overview of the issue to be investigated [broad picture, objective, business case]
  - how the WebCrawler align to the issue – data source and context
  - NLP tasks align to the issue – why it can address the needs
- Length: < 500 words

Deliverables

- Document 2/Part 2: WebCrawler [technicality, implementation, corpus EDA]
- Task 2 Detailing
  - Websites to be consumed
  - A rationale for extracting the web content
  - Content coverage of the data extracted
  - Complexity of the content layout
  - Website/data copyright considerations
  - Metadata supplementation and rational for the supplementation
  - Content extractor to export the important aspects of the data and/or metadata
  - Relevant python coding
  - Demonstration of the application of the WebCrawler (i.e. screen shots)
  - Methodology of processing, cleaning, and storing harvested data for NLP tasking
  - Summary and visualisation of the harvested data. Preliminary EDA is acceptable in this section as well.
- Length: < 1500 words (excluding code and references)

Deliverables

- Document 3/Part 3: Prototype (solution from the dataset) [PoC, v0.1 of Big Picture]
- Task 3: For each NLP task, provide
  - Brief literature review of the NLP task
  - Rational for selection of the NLP task
  - Data pre-processing of inputs and outputs, separate from the WebCrawler harvesting
  - Specification and justification of hyperparameter
  - Preliminary assessment of NLP Task performance
  - Code
- Length: < 1500 words (excluding code and references)

Example of NLP tasks:

- Sentiment Analysis
- Summarisation
- Information Extraction (templated approach)
- Relationship Extraction
- Entity Recognition
- Text Classification
- Topic Modeling
- Question Answering
- Recommender

Example of NLP techniques:

- Tokenization
- Stemming/Lemmatization
- Shallow parsing (POS, NP, VP)
- Syntax parsing
- Word Embedding
- Similarity
- Annotation

Deliverables

- Part 4: Video Presentation
- Task 4:The video presentation,
  - length 3 minutes ± 30 seconds
  - The online repository your deliverables and code are published, and which framework modification of the code can be made. Examples of code management using your code repository [eg, local repo with SourceTree → GitHub]
  - 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.

Finally

JAMES COOK UNIVERSITY AUSTRALIA

- A document of (500+1500+1500) 3500 words +/-10% (excluding code, references and output)
- 3 minute +/- 30 second video presentation (objective, showing repo, data, crawler in action, comments)

Marking Rubric

JAMES COOK UNIVERSITY AUSTRALIA

Overview:

- Proposed High-level Solution design (10)

WebCrawler:

- Website selection (5 marks)
- Description of the web crawler workflow (5 marks)
- Data extraction, collection method and description of corpus (10 marks)
- Working crawler code with screenshots (5 marks)

NLP Tasks:

- Proposes strategies or partial solutions (5 marks)
- Formal quantitative assessment of results and output (10 marks)
- Effective and correct use of text mining package in the solution (20 marks)
- Analysis (5 marks)

Marking Rubric

Video Presentation:

- Publishing (5 marks)
- Limitations Discussion (10)

Reporting:

- Presentation (5 marks)
- Written Report (5 marks)