Assessment 2: Capstone Planning

Type: Oral presentation.

Weight: 15%

Length: Up to maximum of 10 minutes for oral presentation.

Overview

This assignment involves the initial planning for the capstone assignment. Effective planning of the corpus is a key defining factor leading to a successful NLP system. Strategic planning of the data to be scraped and collated into the corpus, as well as the population sampling design will be highlighted in this oral assessment.

Learning outcomes

- Understand how to deploy data science projects into production pipelines
- Apply new data science skills, knowledge, and techniques to solve problems in data science using natural language processing (NLP).
- Apply data science skills, knowledge, and techniques to solve problems in data science NLP projects with a focus on web scraping
-

Deliverables

For this assessment, you are to give an oral presentation.

Tasks

Oral Presentation:

This oral presentation comprises of four main elements – to presented in the same order:

1. Outlining a potential single issue to be investigated or address using NLP methodologies
2. Identification of potential sources of data - webpages and supplementing data from knowledge sources relevant to the issue
3. Identification of website technologies, such as Cloudflare, that may inhibit web scraping of the identified webpages from (3)
4. Consideration of the following:
    a. potential sample size,
    b. corpus size

- how many documents you are grabbing
- how long each document is (how many words)

Google colab

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/

Video production

For this assessment, you are required to a video and post them using the Panopto software provided in LearnJCU.

Marking criteria. Oral Presentation

Criteria	HD/D Above Expectations (100%)	Pass: Attempted and Requires Revision (50%)	Unsatisfactory / Not Attempted (0-49%)
Issue **40% of section grade**	Identifies and discusses: • The Issue, with brief reference to previous peer review application in a similar issue(s) ✓ • Outline of a proposed NLP machine learning approach to be used with comparative strengths relative to other machine learning approaches. Discussions are specific and targeted towards clearly identified a NLP task. Discussions are supported with credible references sources.	Identifies: • The Issue • Outline of a proposed NLP machine learning approach to be used. Discussions are in a general nature of NLP tasks routine data science related situation.	Partially identifies and/or explains some key issues superficially. The issues is vague or unclear.
Potential Data **30% of section grade**	Identifies and discusses: • Relevant domain(s) on the world wide web. ✓ • Demonstration of assessing any of website technologies, such as Cloudflare, that may inhibit web scraping. ✓ Discussions are supported by visual evidence ✓ and include insights to solve any potential problems. Discussions also include insights gained from previous applications in a similar NLP solution. ✓	Identifies and discusses: • Relevant domains on the world wide web • How an assessment of any of website technologies, such as Cloudflare, that may inhibit web scraping can be performed Discussions are general in nature and identify most criteria.	Partially identifies and/or explains some key issues in a superficial data science related situation
Corpus and sampling **30% of section grade**	Identifies and discusses: • potential sample size, ✓ ✓ • corpus size ✓ • Any potential data biases and text data limitations ✓ Discussions involve integrating sampling and linguistic theory to elicit insight to the application of NLP data wrangling to improve machine learning tasks. Discussions also draw upon relevant theory from a wide range of credible sources. ✓ ✗	Identifies and discusses: • potential sample size, • corpus size Discussions are general in nature and identify most criteria.	Partially identifies and/or explains some key issues in a superficial data science related situation