SFU BIG DATA

CMPT 733 Big Data Science - Capstone Project Idea

Machine Learning Applied to Web Scraping

Description

Web scraping has become a valuable source of information, as we increase our data processing capabilities systems are hungry to ingest information and use those to draw meaningful insights, furthermore, we often encounter data pipeline processes where we have a web scraping tool used to collect data from websites and machine learning algorithms to process that data, however, the web scraping itself is a pattern matching problem that could be tackled using machine learning techniques, the naive approach to extract the information from the websites by matching HTML tags does not scale well for different websites and often requires changes if the website layout change

In many of our research projects at SFU we often require to collect data from news outlets in order to perform further analysis one example is the https://gendergaptracker.informedopinions.org where we use the content of news articles to find the representation gap between male and female sources on media one of the challenges is to scale the number of news outlets and still be able to retrieve accurate information about authors, published date, content of the articles, etc… having a more holistic approach to this problem should be able to contribute to projects like this as well as others, not to mention it could also be highly beneficial to commercial applications on industry

Interesting video about the problem: https://www.youtube.com/watch?v=q0lQAMqQViA

Datasets

Datasets could be mainly scrapped online, however, if the students are not willing or cannot do so I would be able to provide them with a dump of some html pages containing news articles of canadian outlets

Additional Resources

https://newsapi.org/s/google-news-api
https://www.diffbot.com/dev/docs/article/

Contact person/Contributor of the Project Idea

Please stay free to reach me out if you have any questions
Alexandre Lopes - Big Data Analyst - Research Computing Group
alopes@sfu.ca
ASB 10978