WEB SCRAPING

What is Web Scraping?

- Web Scraping is the process of extracting data from websites. It involves fetching the web page's HTML code and then parsing it to extract the desired information. This is done programmatically using scripts or programming languages.
- We can use web scraping if we want to do,
  - Data collection
  - Price monitoring
  - Market research
  - Social media analysis
  - Sentiment analysis
- We can use R, Pyhton, Java, JavaScript to do web scraping.
- In R, we got some packages to use it,
  - **rvest:** An R package that is used for web scraping. This package is part of tidyverse.
  - **httr:** An R package that provides tools for working with URLs and HTTP.

Literature Review

- **Web Scraping for Data Analytics: A BeautifulSoup Implementation:** https://ieeexplore.ieee.org/document/10145369
  - A web scraper is a tool used for crawling databases and extracting data.
  - In this project, they developed a web scraper that uses the BeautifulSoup4 library in **Python**.
  - They tested their web scraper on the **Amazon** website. It sends a request to Amazon servers to search for a product based on its name. Then it displays the ten cheapest products and the ten products with the highest ratings and graphs the price frequencies as well as the number of reviews per rate.
- **SENTIMENT ANALYSIS ON SHOPEE E-COMMERCE USING THE NAÏVE BAYES CLASSIFIER ALGORITHM:** https://iocscience.org/ejournal/index.php/mantik/article/view/2397/2012

- Using the **RStudio** application and the **Naive Bayes Classifier Algorithm**, this study aims to find or analyze both positive and negative reviews of the e-commerce application through user reviews on **Google Play**.
- Nave Bayes Classifier Method is used to compare training data and test data after text has undergone text processing to produce positive and negative sentiment classes based on the number of word frequencies.

- **Investigating the indoor environmental quality of different workplaces through web-scraping and text-mining of Glassdoor reviews:** https://www.tandfonline.com/doi/abs/10.1080/09613218.2021.1908879
- This study presents an innovative analysis of occupants' feedback about the IEQ(indoor environmental quality) of **different workplaces** based on web-scraping and text-mining of **online job reviews**. A total of 1,158,706 job reviews posted on **Glassdoor** about 257 large organizations (with more than 10,000 employees) are scraped and analyzed.
- (In this article, they didn't write that what programming they used.)

- **Netizens' behavior towards a blockchain-based esports framework: a TPB and machine learning integrated approach:** https://www.emerald.com/insight/content/doi/10.1108/IJSMS-06-2021-0130/full/html

- This study explores user attitudes toward adopting a blockchain-based framework in the **e-sports industry**. The framework suggests a reward system for stakeholders based on their invested attention, coupled with the inherent protocols of **blockchain technology**.
- They used **RStudio**, package **RedditExtractoR** for scraping and analysis of the discussion referring to the keyword "Verasity" on the **Reddit** website.

- **Social Media's Influence on Mental Health:** http://jbox.gmu.edu/bitstream/handle/1920/12184/Courtney-SocialMedia.pdf?sequence=1&isAllowed=y

- Their goal in this research is to identify the diagnoses and symptoms of mental illness prominent in high social media use. Although anxiety and depression were of frequent discussion, we will not limit our findings to those two and plan to explore additional potential diagnoses.
- They used **RStudio** and social media platforms such as **Twitter**, **Reddit**, **Facebook**, and **Instagram**.

- **Mining netizen's opinion on cryptocurrency: sentiment analysis of Twitter data:** https://www.emerald.com/insight/content/doi/10.1108/SEF-06-2021-0237/full/html

- **Sentiment Analysis of Opinions on the Use of Devices in Students Using the Support Vector Machine (SVM) Method:** https://journal.unpak.ac.id/index.php/komputasi/article/view/6558

- The method used in the classification of opinion is Support Vector Machine (SVM).
- The data used in this study amounted to 1354 taken in 2019 using web scraping techniques on the Twitter site which are then pre-processed so that it can be processed into the program and classified into 3 classes of sentiments, namely negative, neutral and positive sentiments.

- **Sentiment Analysis of the Body-Shaming Beauty Vlog Comments:** https://eudl.eu/pdf/10.4108/eai.12-10-2019.2296530

- - In this article, they used **Naïve Bayes algorithm** for classifying **body-shaming** comments.
- - They scraped data from **YouTube**.

- **Research Note: Scraping Financial Data from the Web Using the R Language:** https://publications.aaahq.org/jeta/article-abstract/15/1/169/9257/Research-Note-Scraping-Financial-Data-from-the-Web