Emplay Assignment Report

Name : Shanmukeswara Reddy Medapati **Date :** 02-04-2024
Assignment : Text Summarization from Web Scraping

Table Of Content

A. Introduction :

Web scraping is a data extraction technique for retrieving information from websites and web pages. In order to extract structured data for use in a variety of applications, websites' content is fetched and parsed programmatically. Web scraping is a useful technique for data collection, research, website monitoring, and job automation.

Web scraping includes obtaining data—such as text, photos, videos, links, and more—from the Internet. Depending on the accessibility of the website and your particular needs, this data may be accessible to the general public or information that requires logging in. Web scraping is frequently carried out by automated scripts or programs that are created using programming languages like Python, JavaScript, or Ruby. These scripts submit HTTP queries to web servers, process the results, and then simulate human browsing behavior.

HTML (Hypertext Markup Language) is frequently used to organize web pages. Online scraping technologies analyze the HTML code to find and extract certain parts such as headers, paragraphs, tables, and links in order to extract data from online pages.

Typical usage scenarios include:

1. Analysis of the competition and market research.
2. E-commerce price comparison and monitoring.
3. Aggregation of news and material.

4. Examination of social media data.
5. Monitoring of employment and housing postings.
6. Scholarly investigation and data gathering.

Web scraping may be facilitated by a number of libraries and tools, including Python's BeautifulSoup and Scrapy.

B. Literature Survey :

Web Scraping

The preferred method for online mining and data extraction is web scraping. With web scraping, you can focus more on finding solutions to issues like what data your organization can use and how to utilize that data rather than getting bored with queries like how to collect the desired data. This post will provide you a straightforward introduction to web scraping methods, resources, and advice. I really hope that you can use these suggestions to guide your business decisions.

In simple terms, an automated process that involves gathering information from various websites on the internet. It is also known as data extraction, content scraping, data scraping, web crawling, data mining, content mining, information gathering, and data collection.

Big Data and automation are no longer novel ideas in today's corporate environment. They are used by people to increase their productivity and effectiveness. Huge data is huge in terms of quantity. Automation is the process of carrying out tasks automatically. Web scraping excels at obtaining large amounts of data quickly and with little manual labor. Web scraping is the solution for huge data acquisition. A lot of precise input data will make you happy if you want to train a machine learning model. Your model will learn valuable lessons from this data, which will help you create a more intelligent algorithm. Web scraping comes in handy at that point to efficiently collect your data from several websites and convert it to a machine-readable format for use in automated systems. This is the overall point of web scraping in today’s world.

Text Summarization

A crucial natural language processing (NLP) approach called text summarizing is used to extract the most important details from large textual materials and make them

more comprehensible and accessible. This procedure is condensing the original material into a manageable length while keeping the crucial ideas and information. The two primary methods of text summary are extractive and abstractive summarization. In extractive summarization, the system recognises and chooses phrases or sections straight from the original text, frequently using statistical or machine learning methods to rank the most pertinent information. Contrarily, abstractive summarizing creates summaries by rewriting and paraphrasing the original text. This method offers greater freedom but has special difficulties in preserving correctness and coherence. This technology has a variety of uses, from automating the summaries of news articles to helping scholars sort through lengthy scholarly papers, eventually easing the process of consuming information and making decisions. Text summarization is still an essential technique for processing and understanding massive volumes of textual data, even as the volume of digital material keeps increasing.

Text summarization techniques, there are two methods used in NLP to summarize texts: abstraction-based and extraction-based.

- Extraction-based summarization: a summary is created by combining a small group of words or phrases from the lengthy text that best capture the key ideas. The outcomes might not be grammatically correct. Types of this technique are - TextRank, Sentence embeddings, Word Frequencies
- Advanced deep learning approaches : (mostly in seq-to-seq models) are used to paraphrase and condense the original content, exactly like humans do. This is known as abstraction-based summarization. The grammatical errors of the extraction-based approaches can be overcome because abstractive machine learning algorithms can produce new phrases and sentences that convey the most crucial information from the source text.Types of this technique are seq2seq.

BeautifulSoup

A well-known Python module called Beautiful Soup is frequently used for web scraping and processing HTML and XML data. It is a useful tool for data extraction, data mining, and web automation activities because it offers a straightforward and intuitive interface for accessing and altering the content of online pages. With the help of Beautiful Soup, developers can easily extract structured data from websites even when the HTML is badly constructed.

Web developers and data scientists love the library because of its versatility and durability. The key advantage of Beautiful Soup is its capacity to convert unstructured online data into organized, usable formats, enabling users to quickly extract certain data from web sites, such as text, links, and characteristics. The main advantage of this library is its capacity to convert unstructured online data into an organized and practical format, enabling users to easily extract certain information from web pages, such as text, links, and characteristics. Beautiful Soup is a popular option for web scraping chores because to its versatility, thorough documentation, and interoperability with other Python programmes. Beautiful Soup makes it easier to access and extract data from the web, enabling users to effectively use web resources for a variety of purposes, whether they are aggregating material, performing research, or automating data collecting.

C. Methodology :

Let' break down the methodology in the following below steps,

I. Locking the target webpage(i.e https://en.wikipedia.org/wiki/Alexander_the_Great)
II. Choosing BeautifulSoup as web scraping library in Python
III. Starting the scraping and extracting the headings , content
IV. Text summarization using extractive based method
V. I followed the above steps in a synchronous order like a pipeline containing a list of functions.
VI. I used the tech stack of Python , OpenAI .

D. Performance Metrics :

Below are the list of performance metrics that I chose to measure the capability of my assignment and they are as follows -

1. **Perplexity** - How well a probability distribution predicts a sample is measured by perplexity. Perplexity may be calculated in the context of text summarizing to evaluate how accurately the summary model **anticipates** the terms in the original text. Better performance is indicated by lower perplexity values.

2. **Fluency -** Fluency evaluates the readability and flow of the resulting summary. Evaluating the summary's naturalness and coherence requires human judgment. The fluency of summaries can be rated on a scale by human assessors.

3. **Jaccard Similarity -** Calculate how semantically similar the resulting summary is to the source text. To evaluate the overlap of words or vectors between the summary and the reference text, Jaccard similarity is frequently used.

4. **Gunning Fog Index or Flesch-Kincaid Grade Level -** Measures of readability rate how simple it is for a reader to comprehend the summary. Summaries with lower readability ratings are simpler to comprehend.

E. Challenges & Solutions :

1. Uncertainty in choosing a right performance metric as there are many metrics related to text summarization.
    a. So the solution that I thought is , Perplexity, Fluency, Semantic Similarity Metrics like Jaccard Similarity and Readability Metrics like FK Score
2. I wasn't so familiar with the BeautifulSoup library.
    a. So the solution is, I went through the documentation for a high level understanding and read a few online articles about the library for better flow of it. As I had knowledge of web scraping previously, I was able to relate with this.
3. I was so perplexed in picking a particular text summarization technique and I brainstormed a solution based on the context of the assignment.
    a. So the method that I chose was, Extractive summarization entails choosing and removing phrases or portions from the original material to generate better context in the summary.
4. I was unable to crack a way to create prompt chains to generate summary based on prompt. But as per the resources that i can use, I took help of OpenAI and gave the solution.

F. Conclusion :

Last but not least, I developed a small tool to generate summary based on the web page given in the input by using the BeautifulSoup for Web Scraping and then followed by Extraction based summary generation for better and instant understanding of the given text using Python language.

G. Bibliography

1. https://youtu.be/e2-nMDDzC5A
2. https://www.octoparse.com/blog/introduction-to-web-scraping-techniques-and-tools?utm_source=medium&utm_medium=blog&utm_campaign=promotion
3. https://www.octoparse.com/blog/what-is-web-scraping-basics-and-use-cases?utm_source=medium&utm_medium=blog&utm_campaign=promotion
4. https://medium.com/google-cloud/how-to-use-llms-to-generate-concise-summaries-of-text-a04966659ed
5. https://txt.generativeailab.org/introducing-cohere-summarize-beta-unlocking-the-power-of-text-summarization-cab84d224acf
6. https://medium.com/nlplanet/two-minutes-nlp-four-different-approaches-to-text-summarization-5a0ce9c06c74
7. https://www.crummy.com/software/BeautifulSoup/bs4/doc/