Web Scraping from Wikipedia pages

Amirhossein Najafizadeh

May 9, 2024

Abstract

This project aims to gather information from Wikipedia pages and analyze it. Specifically, we'll use web scraping to extract data from these pages. Then, we'll employ techniques to determine what is the page talking about based on its content. To put it in simple words, we want to categorise our pages.

1 Introduction

Our project focuses on extracting and analyzing data from Wikipedia pages using web scraping techniques. Moreover, we want to categorise our pages based on their content. For example, take a look at this page in Wikipedia about Helium. As you can see, this page explains about Helium. What we need as our objective is to scrape this page data, and say that this page can be categoriesd as chemistry, periodic table, and elements.

2 Objectives

Our main objectives are to:

- Gather data from Wikipedia pages.
- Use web scraping techniques to extract relevant information.
- Apply information retrieval techniques to identify page category based on their content.

3 Methodology

We'll begin by selecting Wikipedia pages for analysis. Then, we'll employ web scraping methods to extract data from these pages. Finally, we'll use information retrieval techniques to determine the appropriate titles for the extracted content.

3.1 Gathering Data

Since our input is a web-page address, we need to open that link. Therefore, as our first step we are going to use Python **requests** library to get a Wikipedia page as a HTML data.

```
1 import requests
2
3 page = requests.get("https://en.wikipedia.org/wiki/Helium")
4 print(page.content)
```

Listing 1: Example of getting a page content

In the code above, we openend a link and extracted its content as a string. Now I want you to use this example in order to get an input link and extract its content into a string variable. Note that if the input link is incorrent, you will get an error. Make sure to handle these types of errors by using page status code field.

3.2 Scraping Page Content

3.3 Creating Index Table

4 Expected Outcomes

We anticipate achieving the following outcomes:

- Successful extraction of data from Wikipedia pages using **requests** module.
- Accurate determination of page titles based on content analysis using **beautiful soup**.
- Contribution to the understanding of web scraping and information retrieval techniques.

5 Conclusion

In conclusion, this project aims to extract and analyze data from Wikipedia pages using web scraping and information retrieval techniques. By achieving our objectives, we hope to contribute valuable insights to the field of data analysis.