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.

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 titles 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.

4 Timeline

Our project will proceed according to the following timeline:

- Month 1: Data gathering and initial web scraping.
- Month 2: Refining web scraping techniques and beginning information retrieval analysis.
- Month 3: Completing information retrieval analysis and finalizing project report.

5 Expected Outcomes

We anticipate achieving the following outcomes:

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

6 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.