Indian Institute of Science Education and Research(IISER)Mohali

Instructor: Dr. Vishal Bhardwaj

KEYWORD IDENTIFICATION

Group 8

Anas K P (MS21224)

G Indrajith (MS21092)

Fayiz M (MS21078)

Date: November 7, 2023

1 Contributions

Throughout the duration of this project, the entire team exhibited exceptional teamwork and a shared commitment to the project's success. Each team member actively participated, contributed, and played an equal role in various aspects of the project.

2 Aim

- To perform keyword identification process on the Indian Institute of Science Education and Research (IISER) Pune website
- To perform web scraping, data preprocessing, keyword extraction, and data visualization.
- To cross check the analysis being done using a site (`https://web.iisermohali.ac.in/`) of known data.

3 Library and Packages Used

We utilized essential libraries to achieve project goals:

Requests: For collecting web data from IISER Pune's website.

Beautiful Soup: To parse HTML content and prepare it for analysis.

spaCy: To tokenize text and identify keywords, while removing stopwords.

sqlite3: For managing the database, although data insertion was in progress.

Matplotlib: To create bar charts and word clouds for data visualization.

Wordcloud: Collaborating with Matplotlib to enhance word cloud visualizations.

4 Summary

This project aimed to extract and analyze textual data from the IISER Pune website. Key steps included web scraping to gather content, text processing to remove stopwords and non-alphabetic characters, and keyword identification using spaCy. Data visualization was planned but not implemented in the code. Improvements were needed for SSL certificate verification, SQLite database interaction, and generating visualizations. Future work could enhance data storage, resolve SSL issues, and create meaningful visualizations for insights from the web content.

5 Application

Enhancing online content personalization, improving search engine results, shaping media decisions, streamlining research, enabling chatbots, aiding e-learning, gauging sentiment, analyzing healthcare data, content moderation, and using data for business intelligence.

6 Challenges

Challenges included SSL and similar security certificate issues, SQLite data insertion problems, data shortage for visualizations

7 References

1. https://realpython.com/beautiful-soup-web-scraper-python/
2. https://www.scrapingbee.com/blog/python-web-scraping-beautiful-soup/
3. https://pythonprogramming.net/introduction-scraping-parsing-beautiful-soup-tutorial/