Lab-2 (15/7/25)

1. Source of the Website:

For my scraping task, I used Project Gutenberg – https://www.gutenberg.org.

I specifically scraped the text from the classic book A Tale of Two Cities by Charles Dickens. (Link: https://www.gutenberg.org/files/98/98-0.txt)

2. What Data I Scraped and Why:

I scraped the full plain text of the book.

The main reason was to practice text processing techniques like tokenization, which are used in Natural Language Processing (NLP).

This was done purely for learning purposes, not for any commercial use.

3. Under What Clause It Is Allowed:

Project Gutenberg makes it very clear that their books are in the public domain.

This means the content is free to use, copy, and share legally, especially for personal or educational use like mine.They even mention:

“This eBook is for the use of anyone anywhere… with almost no restrictions whatsoever.”

So scraping this text is completely allowed.

4. Copyright Clause :

Since the book is in the public domain, there is no copyright on it anymore.

That’s why I didn’t need any special permission to use or scrape the content.

It’s open to everyone, which is perfect for students and researchers.