# Exploratory Data Analysis (EDA) Report

## 1. Dataset Description

The dataset used for this EDA is derived from NCERT Class X Science book content. It contains extracted text data from chapters in the form of paragraphs, headings, and sub-headings. The dataset was processed to support contextual search and LLM-driven responses.

## 2. Objectives

- Understand the structure and distribution of chapter content  
- Identify frequent science topics  
- Check word count and keyword distribution  
- Detect data quality issues such as missing or irrelevant content

## 3. Summary Statistics

Each chapter was tokenized and analyzed for word count, sentence count, and topic frequency. Top recurring science terms included 'photosynthesis', 'acid', 'base', 'friction', etc.

## 4. Key Insights

- The content is clean and well-structured with minimal noise.  
- Certain chapters contain dense scientific terminology.  
- Average paragraph length is moderate, ensuring compatibility with LLM prompt limits.

**🧰 Libraries Used & Their Versions**

| **Library** | **Version** |
| --- | --- |
| pandas | 2.2.2 |
| numpy | 1.26.4 |
| matplotlib | 3.8.4 |
| seaborn | 0.13.2 |
| wordcloud | 1.9.3 |
| nltk | 3.8.1 |
| scikit-learn | 1.4.1 |