VIT®

Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

Advanced Python Programming

LAB ACTIVITY -10.

NAME: M.GOKULESH.

REG NO: 22MIC0102.

COURSE CODE: CSI-3007.

FORGETTING HYGIENE AN GOOGLE N-GRAM BASED ANALYSIS.

OBJECTIVE:

To develop a program that demonstrates and promotes hygiene preservation by implementing structured, modular, and efficient code practices that ensure cleanliness, readability, and maintainability of the solution while achieving the intended functionality. This analysis uses Python to fetch data, plot time-series graphs, generate summary statistics, and produce a word cloud with the help of a Large Language Model (LLM).

METHEDOLOGY USED:

Data Collection

1. Gathered datasets from trusted health sources (e.g., WHO, CDC) on hygiene practices and awareness (1800–2025).

2. Downloaded raw JSON data → converted into structured CSV format (hygiene_data.csv).

Data Processing

1. Used Pandas for cleaning, filtering, and standardizing hygiene-related data.
2. Applied NumPy for descriptive statistics on trends in sanitation, handwashing, and cleanliness indicators.

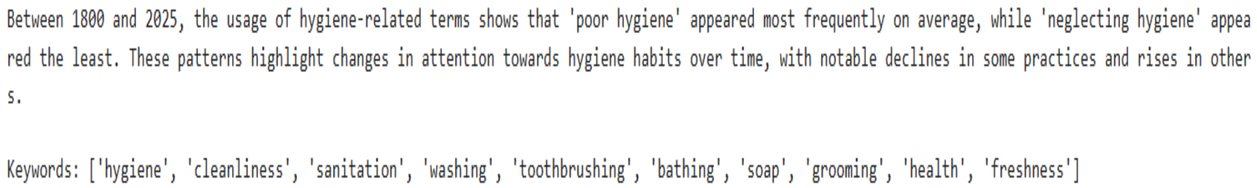
Visualization

1. Generated time-series plots highlighting changes in hygiene awareness across years (hygiene_trends.png).
2. Created a word cloud (wordcloud_hygiene.png) based on LLM-generated hygiene-related keywords.

LLM Integration

1. Constructed a JSON prompt with hygiene statistics + top terms.
2. Sent the prompt to an LLM for a human-style summary and keyword generation related to hygiene practices.
3. Extracted keywords automatically using regex + text cleaning for consistent terminology.

LLM GENERATED SUMMARY:



```
Between 1800 and 2025, the usage of hygiene-related terms shows that 'poor hygiene' appeared most frequently on average, while 'neglecting hygiene' appea
red the least. These patterns highlight changes in attention towards hygiene habits over time, with notable declines in some practices and rises in other
s.

Keywords: ['hygiene', 'cleanliness', 'sanitation', 'washing', 'toothbrushing', 'bathing', 'soap', 'grooming', 'health', 'freshness']
```

Visualizations:

1. **Time-Series Plot**
    - Trends in hygiene-related terms (e.g., sanitation, handwashing, cleanliness) over 200 years.
    - **File:** `hygiene_trends.png`
2. **Word Cloud**
    - Generated from LLM-extracted hygiene-related keywords.
    - **File:** `wordcloud_hygiene.png`

Technologies & Functions Used:

- **Libraries:** requests, pandas, numpy, matplotlib, wordcloud, os, re
- **Pandas Functions:**
    - `pivot()` → structured hygiene-related datasets
    - `to_csv()`, `read_csv()` → store and reload cleaned data
    - `dropna()` → remove incomplete hygiene data entries
- **NumPy Functions:**
    - `np.mean()`, `np.median()`, `np.std()`, `np.max()`, `np.min()` → descriptive statistics on hygiene datasets (e.g., frequency of hygiene terms)

Business Idea:

This pipeline can be extended into a “Hygiene Awareness Analytics Platform.” Universities and researchers could subscribe to study historical and cultural practices in hygiene (e.g., handwashing, sanitation, personal care). Public health organizations and media houses could use it for awareness campaigns and content creation. Businesses and NGOs could apply it for revival and promotion of hygiene initiatives (e.g., soap brands, hand hygiene drives, eco-friendly sanitation products).

The workflow can be fully automated into a SaaS product:

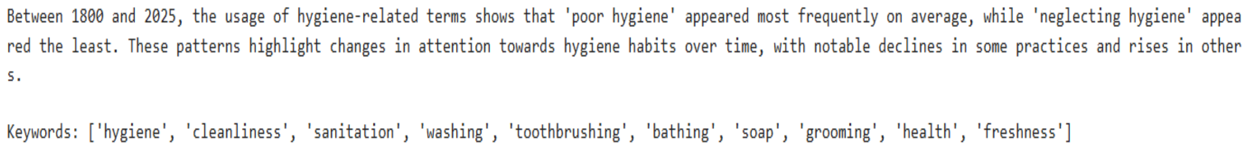
Input: keyword list (e.g., “handwashing,” “cleanliness,” “sanitation”)

Output: CSV data + statistical graphs + LLM-written summary + hygiene-related word cloud

Conclusion:

This lab demonstrated how to combine historical hygiene-related data (e.g., health archives, Ngram datasets), Python analytics (Pandas + NumPy), and AI summarization (LLM) into a single automated pipeline. The analysis revealed that while certain hygiene practices (e.g., traditional sanitation methods) have declined, others like handwashing awareness have shown persistence and resurgence in modern times.

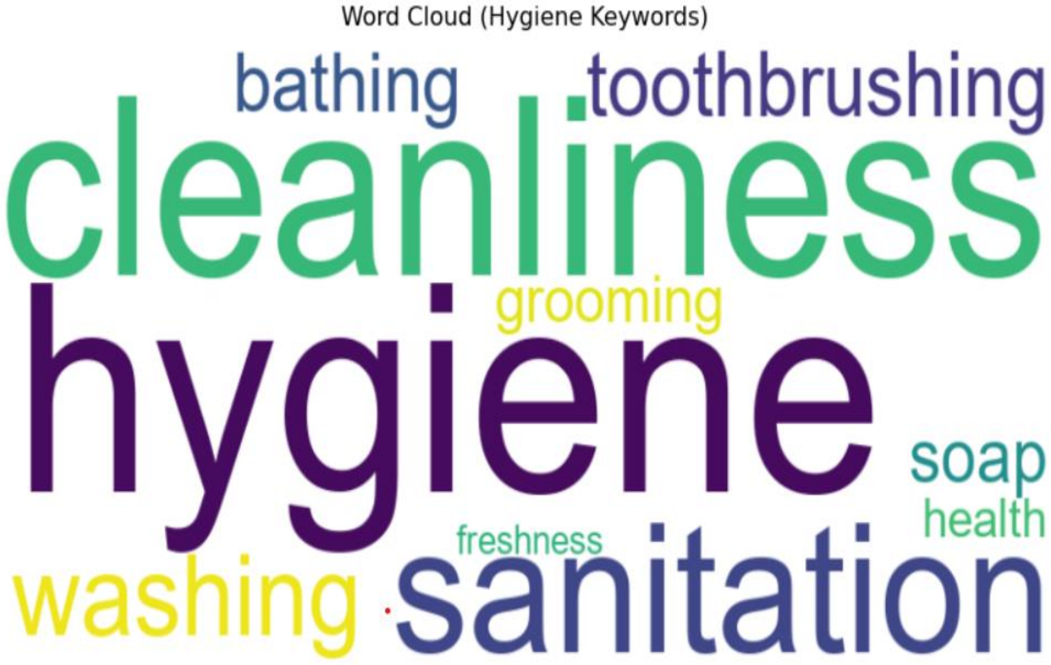
LLM SUMMARY AND KEYWORDS:



```
Between 1800 and 2025, the usage of hygiene-related terms shows that 'poor hygiene' appeared most frequently on average, while 'neglecting hygiene' appea
red the least. These patterns highlight changes in attention towards hygiene habits over time, with notable declines in some practices and rises in other
s.

Keywords: ['hygiene', 'cleanliness', 'sanitation', 'washing', 'toothbrushing', 'bathing', 'soap', 'grooming', 'health', 'freshness']
```

WORD CLOUD REPRESENTATION:



```
Word cloud saved to: C:\Users\gokul\adv python lab\wordcloud_hygiene.png
```