Building Child-Friendly Educational Datasets

This notebook streamlines the creation of engaging, child-friendly educational datasets. By automating web searches, content extraction, cleaning, and leveraging Gemini for structured outputs, including Learning Objectives, we transform raw web content into valuable learning resources.

The Purpose: A Kid-Friendly Dataset at Scale

Automated Content Creation

The notebook autonomously scours the web to build a dataset tailored for young learners.

Child-Friendly Focus

Content is carefully cleaned and adapted to be appropriate and engaging for children.

Gemini-Powered Generation

Leveraging the Gemini API, the notebook generates structured educational outputs.

Learning Objectives (LOs)

Crucially, it extracts and organises Learning Objectives, making content truly educational.

Unpacking the Main Workflow

The process is broken down into clear, sequential steps to ensure efficiency and accuracy in dataset generation.

01

Environment Setup

Initial configuration, including file uploads and library installations for web scraping, Google Search, Gemini API, and data handling.

02

Library & API Key Import

Loading necessary Python modules and authenticating with Google Custom Search API and Gemini (Google Generative AI).

03

Topic Bank Creation

Defining a structured bank of child-friendly educational topics across various categories.

04

Google Search Step

Retrieving relevant webpages for each topic using the Google Custom Search API.

05

Web Scraping & Cleaning

Extracting, cleaning, and preparing raw text from webpages for Gemini processing.

06

Gemini Model Setup

Connecting to and verifying the chosen Gemini model and API key.

07

Checkpointing & CSV Management

Implementing mechanisms to save progress and resume dataset generation seamlessly.

08

Dataset Generation Loop

The core process of searching, scraping, processing with Gemini, and saving to CSV.

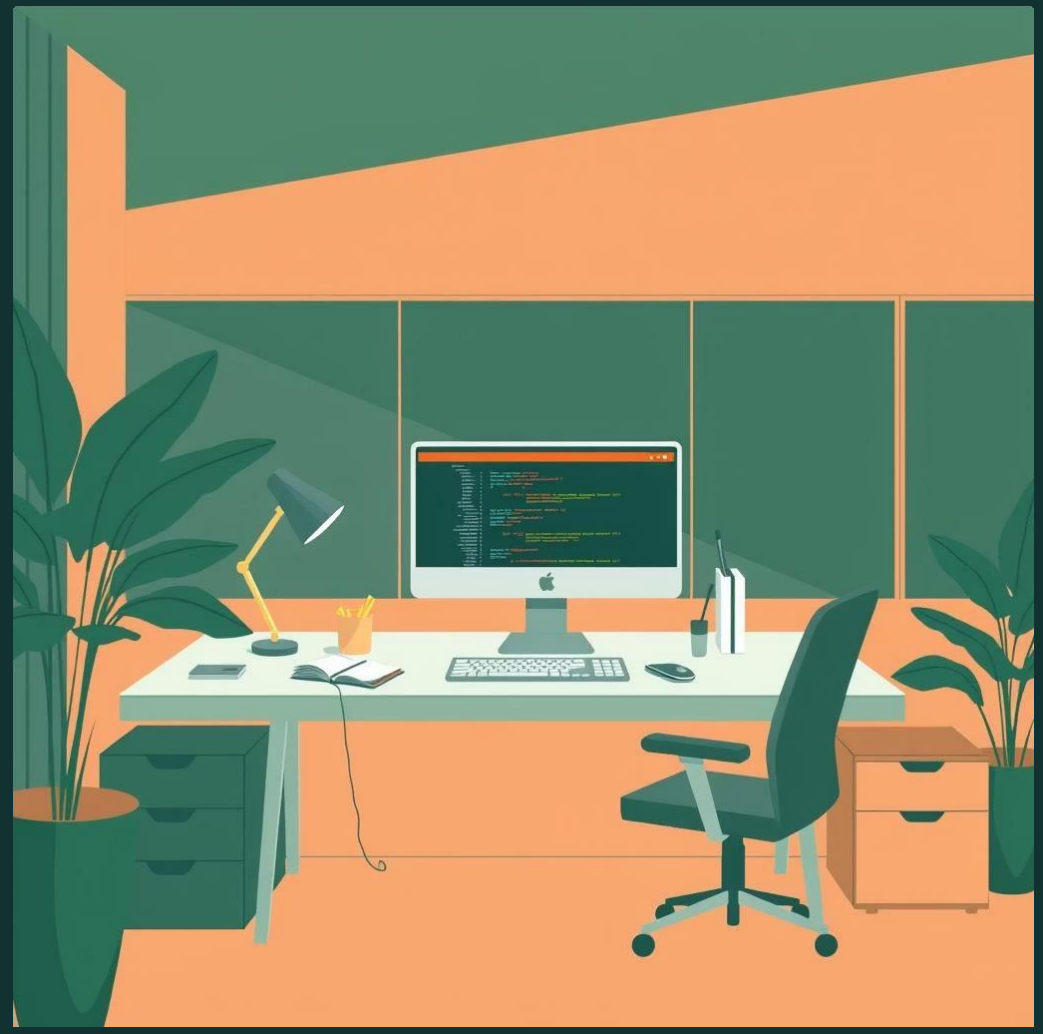
Initial Setup: Preparing the Ground

1. Environment Configuration

Before any data processing begins, the notebook establishes its operational environment.

- Uploading essential files, such as API keys and configuration settings.
- Installing crucial libraries for web scraping, Google Search integration, Gemini API access, and efficient data handling.

This foundational step ensures all necessary tools are in place for automated dataset creation.



API Integration: Powering the Search and Generation

2. Importing Libraries & Loading API Keys

Seamless connectivity is vital for data acquisition and content generation.

- Python modules for HTTP requests, text processing, and DataFrame management are loaded.
- API keys for Google Custom Search and Gemini (Google Generative AI) are securely loaded.
- Robust error handling notifies users if keys are missing or invalid, ensuring smooth operation.

The Curriculum Framework: Children's Educational Topic Bank

A meticulously curated topic bank serves as the foundation for content generation.

1	**Science & Nature** Explore the natural world and its wonders.
2	**Animals** Learn about diverse species and their habitats.
3	**Space** Journey through the cosmos, planets, and stars.
4	**History** Discover tales from the past and important events.
5	**Human Body** Understand how our bodies work.
6	**Environment** Explore our planet and how to protect it.
7	**Technology** Discover innovations shaping our future.

Each category contains several child-friendly topics, ensuring a broad and engaging learning experience.

Content Acquisition & Refinement

4. Google Search Step

For each topic, a dedicated search function uses the Google Custom Search API to find highly relevant webpages, extracting titles, URLs, and snippets. Only the most appropriate links are selected for further processing.

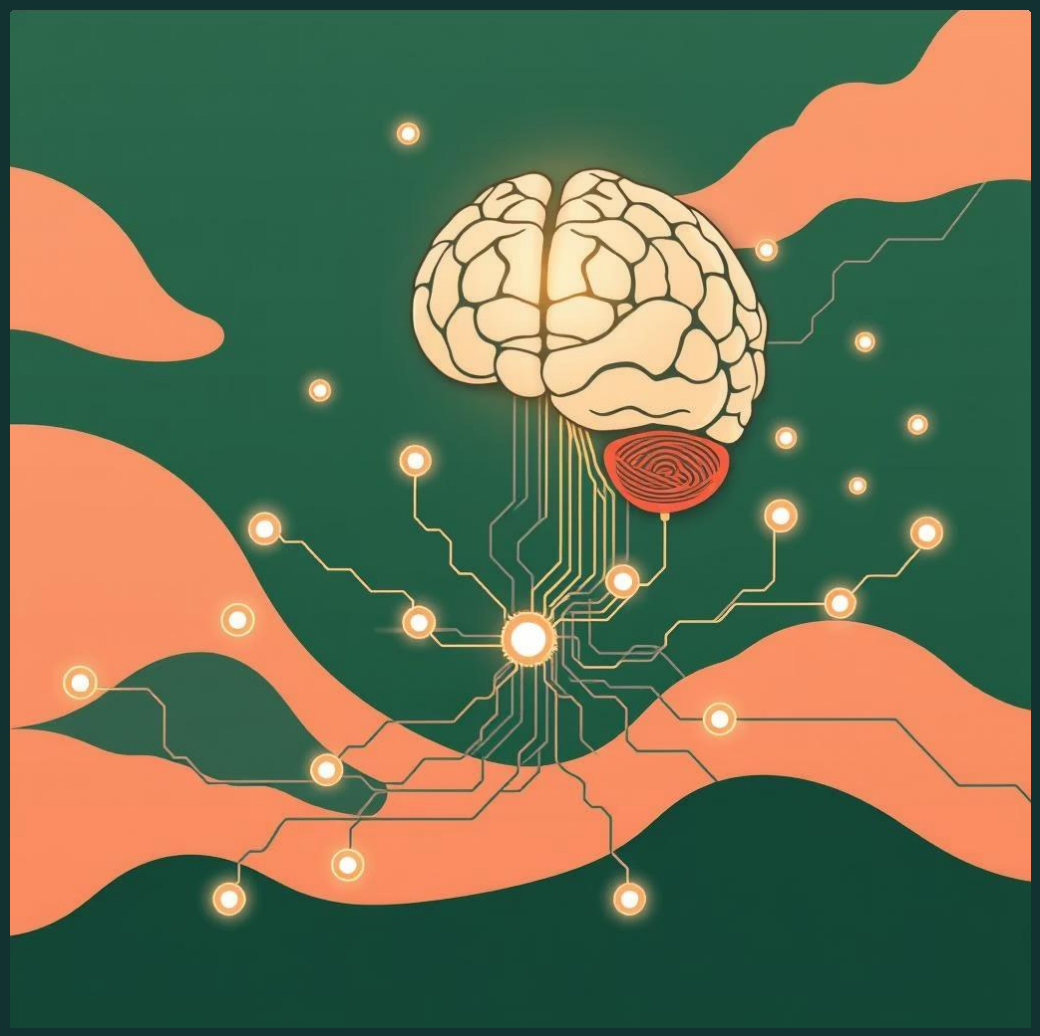
5. Web Scraping & Text Cleaning

Once suitable URLs are identified, the notebook downloads each webpage. It then meticulously extracts readable text, removing irrelevant elements like HTML tags, advertisements, and navigation menus. The text is checked for length requirements and ensures it's clean and educational for children, serving as the raw material for Gemini.

Gemini Integration & Robust Data Management

6. Gemini Model Setup

The notebook establishes a connection with the chosen Gemini model (e.g., gemini-flash), verifying its availability and confirming the API key's functionality. This ensures a reliable link for content generation.



7. Checkpointing & CSV Management

To prevent re-processing and enhance efficiency, a dynamic CSV file (`dynamic_dataset.csv`) is used. It tracks completed topics, allowing the notebook to resume operations after any interruptions. Each newly processed topic is safely appended, making the dataset generation stable.

The Core: Main Dataset Generation Loop

This updated loop iterates through every topic in the Topic Bank, performing a critical sequence of actions:

Google Search

Identifying relevant educational websites.

Scrape & Clean

Extracting and preparing high-quality, kid-friendly text.

Send to Gemini

Gemini generates a rewritten explanation, 3310 Learning Objectives, and other structured outputs.

Save to CSV

All results are appended to `dynamic_dataset.csv`.

The Final Product: A Rich Educational Dataset

The culmination of this process is a comprehensive educational dataset. Each row is a complete learning unit, featuring:

- **Topic & Category:** Clear classification of content.
- **Raw Scraped Text:** The original, cleaned web content.
- **Gemini-Generated Explanation:** A clear, structured, and child-friendly summary.
- **Learning Objectives (LOs):** Specific, measurable goals for learners.
- **Additional Educational Components:** Summaries, key ideas, main concepts, and other lesson elements.

This dataset is perfectly suited for model fine-tuning, RAG systems, curriculum generation, and various educational AI applications.