Sustainable Smart City Assistant Powered by IBM Granite LLM Generative AI

The Urban Challenge: Growing Cities, Complex Needs

- By 2050, 68% of the world’s population will live in urban areas (UN Habitat).
- Cities face rising demands: energy, waste, infrastructure, and climate resilience.
- Smart, sustainable solutions are critical to improve quality of life and reduce environmental impact.

IBM Granite: Enterprise-Grade Generative AI for Real-World Impact

Open & Performant

Granite LLMs are optimized for business use, offering a balance of power and energy efficiency across models ranging from 2B to 8B parameters.

Advanced Capabilities

Features include long-context understanding (128K tokens), instruction-following, and multilingual support for diverse urban needs.

Ethical & Secure

Built with rigorous data curation, ethical safeguards, and IBM’s industry-leading indemnification, ensuring trustworthy AI.

Sustainability at IBM: AI That Cares for the Planet

Renewable Training: IBM’s AI models are trained on renewable-powered supercomputers, minimizing environmental impact.

Efficient Models: Smaller, efficient Granite models reduce energy consumption and carbon footprint, contributing to a greener future.

Optimized Hardware: Innovations like Telum II processors optimize AI workloads for lower power use.

Smart Techniques: Techniques such as data pruning and iterative scaling cut training energy

Smart City Assistant: AI-Powered Urban Operations & Citizen Engagement

Real-time Data Integration

Seamlessly integrates real-time data from sensors, cameras, and citizen reports for comprehensive insights.

Predictive & Prescriptive AI

Granite LLM analyzes data, predicts urban challenges, and recommends actions for city services.

Optimized City Services

Examples: waste management, energy forecasting, infrastructure maintenance alerts.

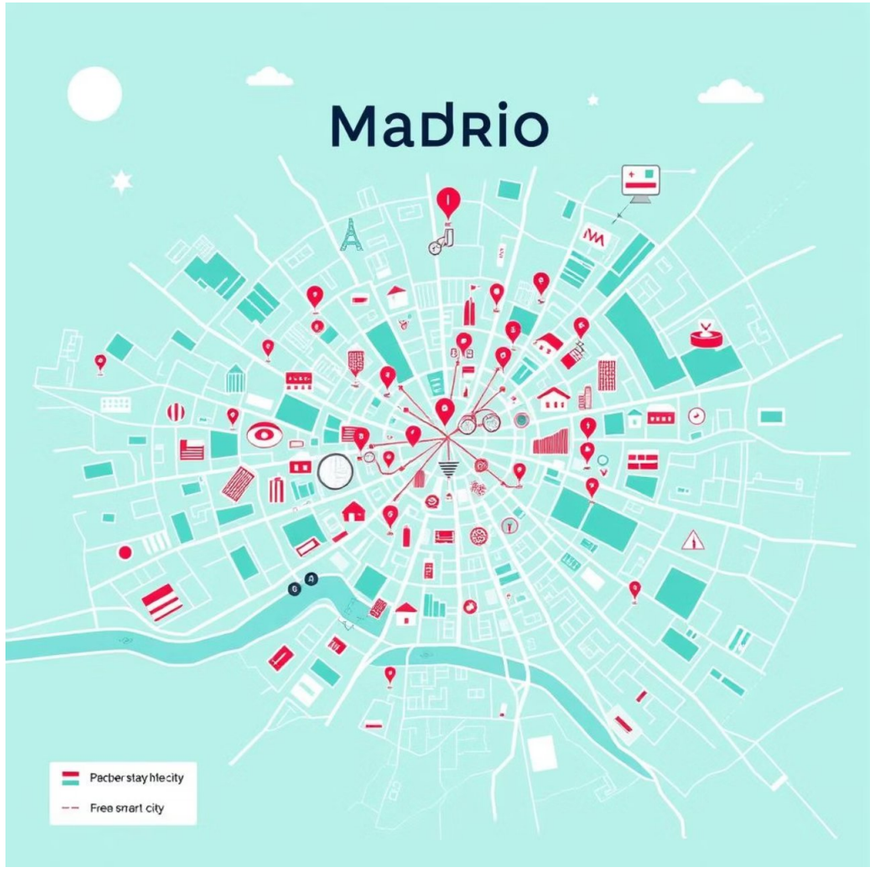
Enhanced Citizen Communication

Enables two-way communication, allowing citizens to report issues and receive instant updates via mobile apps.

Case Study: Madrid’s Intelligent Madrid (MiNT) Platform

IBM-powered platform managing over 5 million assets including lamp posts, trees, vehicles, and waste systems across Madrid.

- Citizen Engagement: Citizens upload photos and locations of issues, which the AI prioritizes for resolution.
- Operational Efficiency: Over 300 KPIs are monitored daily, improving service quality and resource dispatch.
- Tangible Results: Faster problem resolution, enhanced transparency, and increased citizen trust.

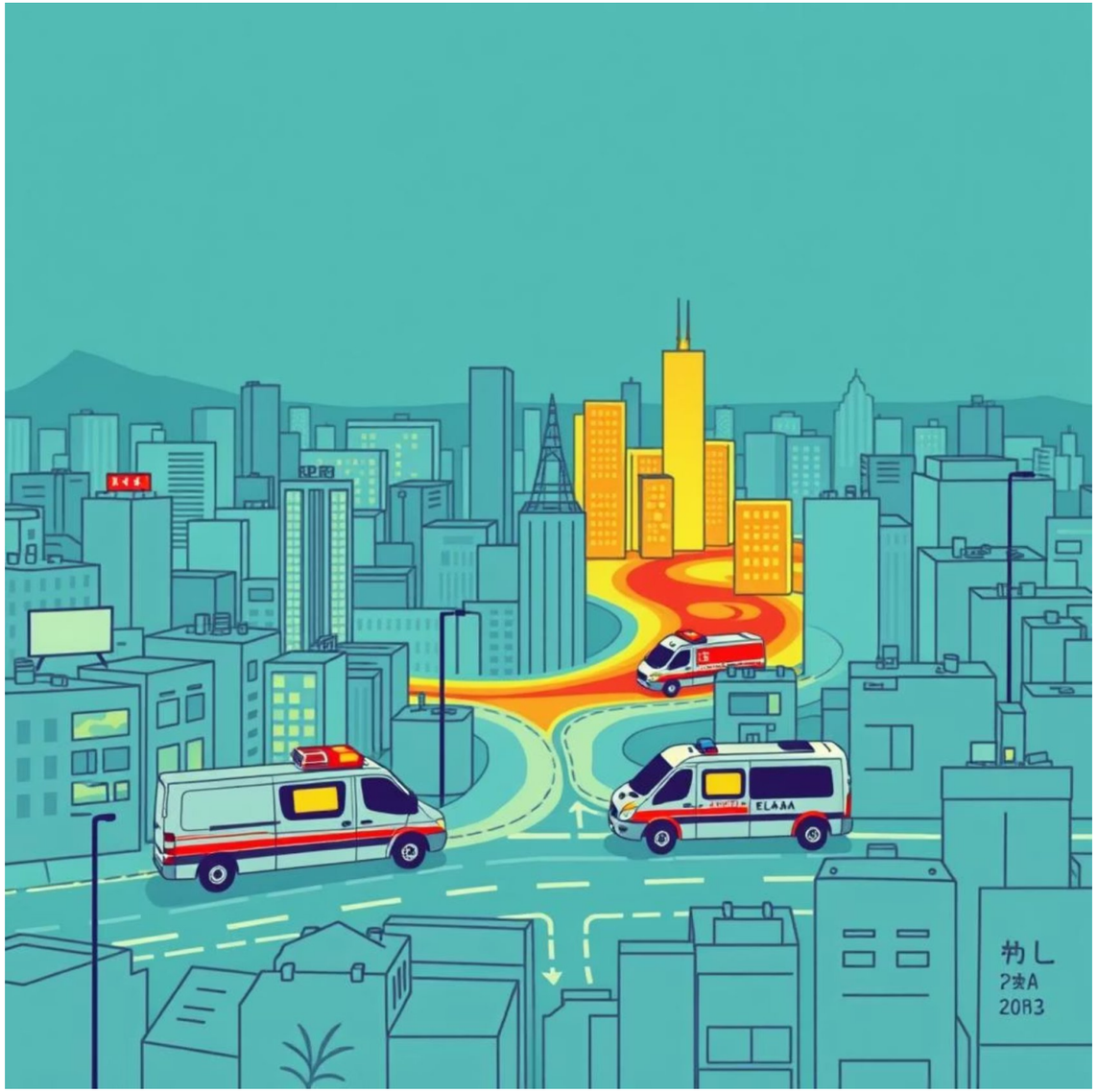


AI for Climate Resilience & Public Health

Through collaboration with C40 Cities, IBM AI predicts urban heat risks and helps mitigate their impacts on residents.

AI tools identify vulnerable populations and optimize resource allocation during heat waves.

Integration with healthcare systems provides automated warnings and preventive care guidance, improving public health outcomes.



Developer Empowerment with IBM Granite.Code

1

AI-Assisted Coding

The Granite.Code extension offers advanced AI assistance for city tech teams.

2

Broad Language Support

Supports 116 programming languages with context-aware code generation and testing.

3

Rapid Development

Enables quick development and customization of smart city applications.

4

Data Privacy & Control

Self-hosting option ensures data privacy and operational control for sensitive urban data.

The Future: Scalable, Transparent, and Ethical AI for Cities

- Continuous Innovation: Ongoing model updates with enhanced multimodal and industry-specific capabilities.
- Transparent & Ethical AI: Built-in explainability and bias mitigation ensure trustworthy and fair AI systems.
- Open Architecture: Allows cities to tailor AI assistants to local needs and regulations, fostering adaptability.
- IBM's Commitment: AI that advances sustainability, equity, and resilience for urban environments.

Platform Login Credentials - Pr | Project - Student | colab.google | Health AI.ipynb - Colab

colab.research.google.com/drive/1eYsR57NJRbeTSIRkz51Qm46IgoyDJUbz#scrollTo=kUv-ecpzJppr

Health AI.ipynb

File Edit View Insert Runtime Tools Help

Share Gemini

Commands + Code + Text Run all

RAM Disk

[1]

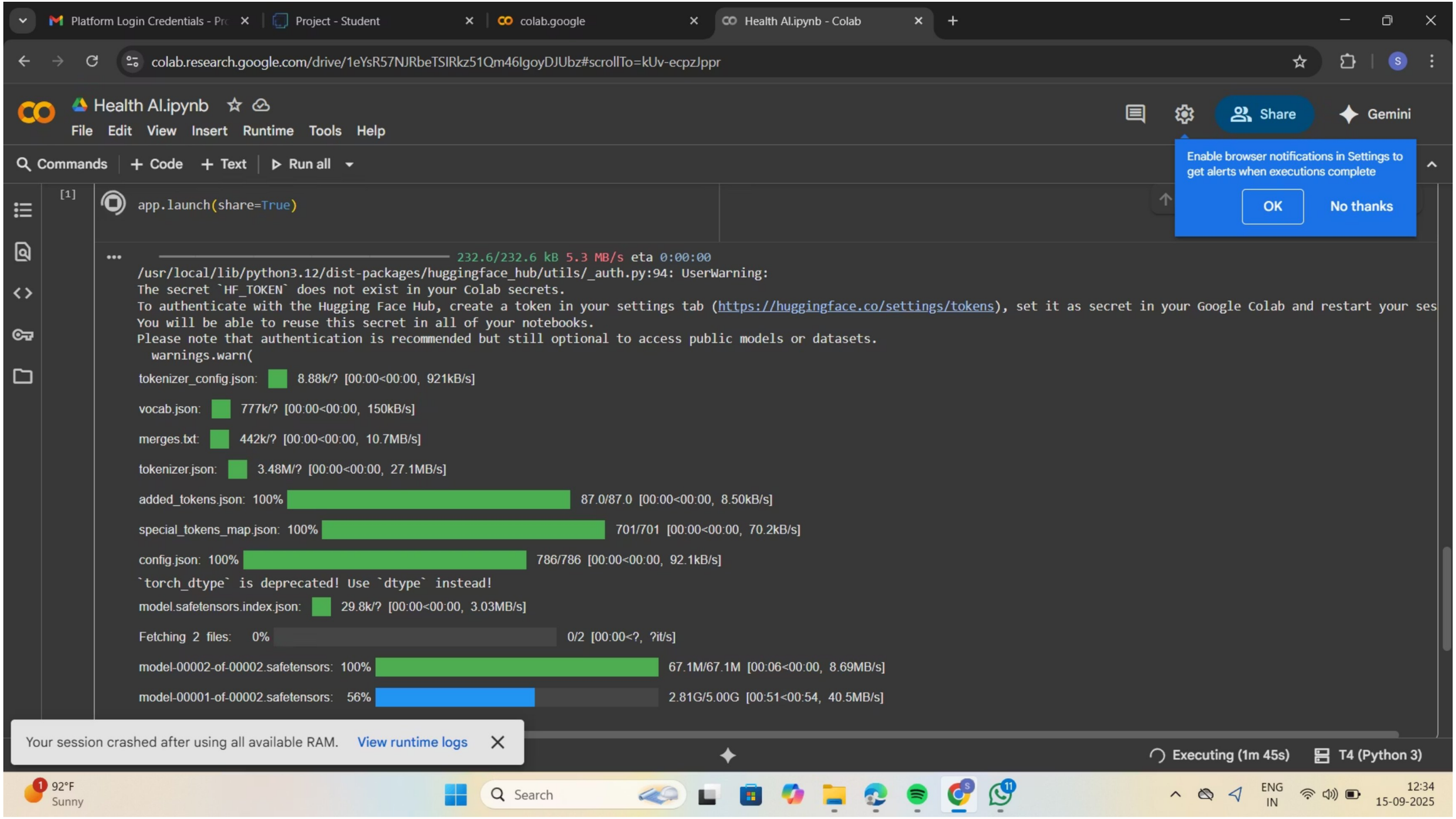
```
!pip install transformers torch gradio PyPDF2 -q

import gradio as gr
import torch
from transformers import AutoTokenizer, AutoModelForCausalLM
import PyPDF2
import io

# Load model and tokenizer
model_name = "ibm-granite/granite-3.2-2b-instruct"
tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModelForCausalLM.from_pretrained(
    model_name,
    torch_dtype=torch.float16 if torch.cuda.is_available() else torch.float32,
    device_map="auto" if torch.cuda.is_available() else None
)

if tokenizer.pad_token is None:
    tokenizer.pad_token = tokenizer.eos_token

def generate_response(prompt, max_length=1024):
    inputs = tokenizer(prompt, return_tensors="pt", truncation=True, max_length=512)

    if torch.cuda.is_available():
        inputs = {k: v.to(model.device) for k, v in inputs.items()}

    with torch.no_grad():
        outputs = model.generate(
```

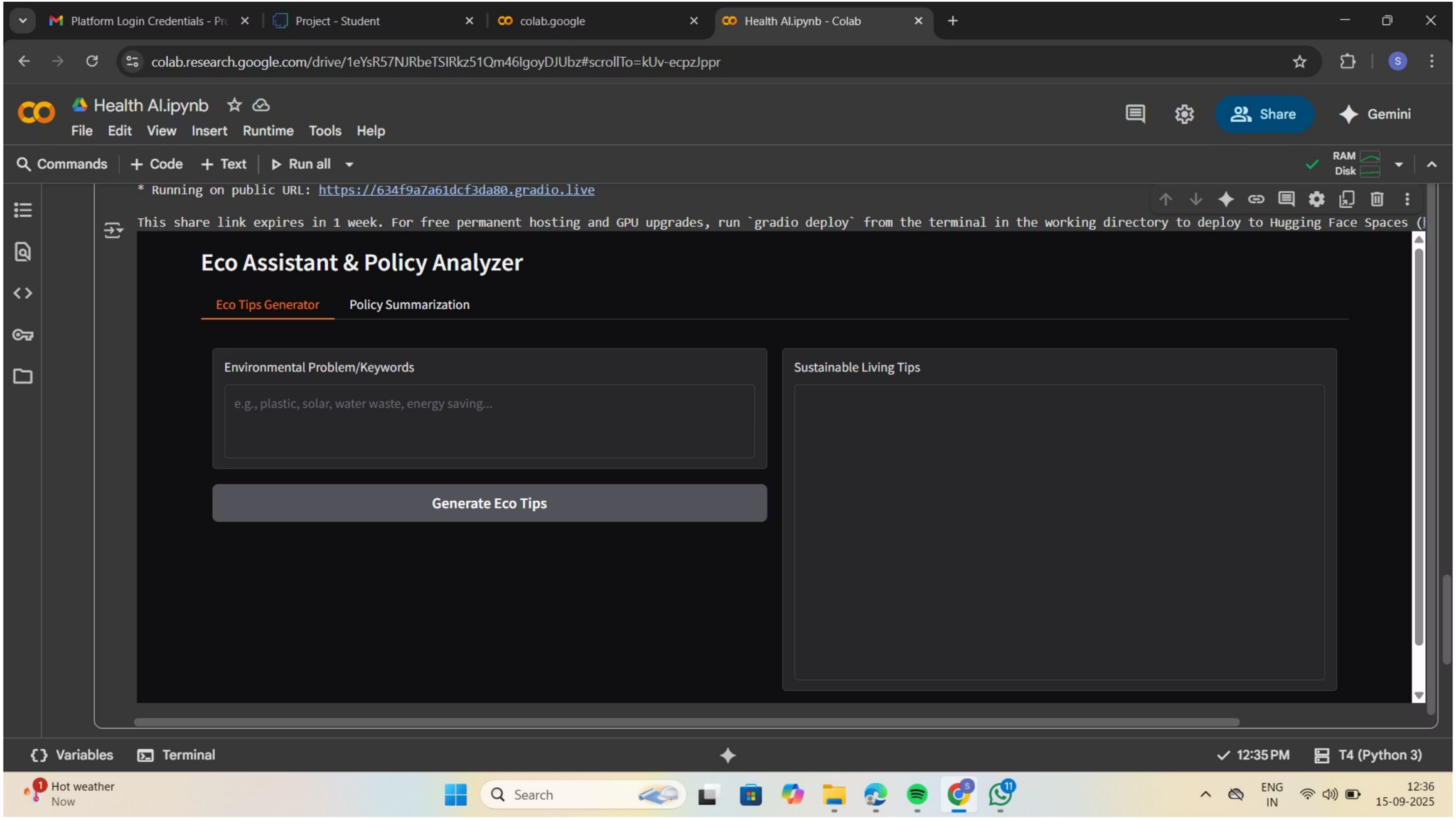
Variables Terminal

Executing (13s) Python 3

Search

ENG IN 12:27 15-09-2025


Platform Login Credentials - Pr
Project - Student
colab.google
Health AI.ipynb - Colab
colab.research.google.com/drive/1eYsR57NJRbeTSIRkz51Qm46IgoyDJUbz#scrollTo=kUv-ecpzJppr
Health AI.ipynb
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands + Code + Text Run all
Enable browser notifications in Settings to get alerts when executions complete
OK
No thanks
[1]
app.launch(share=True)
232.6/232.6 kB 5.3 MB/s eta 0:00:00
/usr/local/lib/python3.12/dist-packages/huggingface_hub/utils/_auth.py:94: UserWarning:
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your settings tab (https://huggingface.co/settings/tokens), set it as secret in your Google Colab and restart your ses
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to access public models or datasets.
warnings.warn(
tokenizer_config.json: 8.88k/? [00:00<00:00, 921kB/s]
vocab.json: 777k/? [00:00<00:00, 150kB/s]
merges.txt: 442k/? [00:00<00:00, 10.7MB/s]
tokenizer.json: 3.48M/? [00:00<00:00, 27.1MB/s]
added_tokens.json: 100% 87.0/87.0 [00:00<00:00, 8.50kB/s]
special_tokens_map.json: 100% 701/701 [00:00<00:00, 70.2kB/s]
config.json: 100% 786/786 [00:00<00:00, 92.1kB/s]
`torch_dtype` is deprecated! Use `dtype` instead!
model.safetensors.index.json: 29.8k/? [00:00<00:00, 3.03MB/s]
Fetching 2 files: 0% 0/2 [00:00<?, ?it/s]
model-00002-of-00002.safetensors: 100% 67.1M/67.1M [00:06<00:00, 8.69MB/s]
model-00001-of-00002.safetensors: 56% 2.81G/5.00G [00:51<00:54, 40.5MB/s]
Your session crashed after using all available RAM.
View runtime logs
Executing (1m 45s)
T4 (Python 3)
92°F Sunny
Search
ENG IN
12:34 15-09-2025


Platform Login Credentials - Pr
Project - Student
colab.google
Health AI.ipynb - Colab
colab.research.google.com/drive/1eYsR57NJRbeTSIRkz51Qm46IgoyDJUbz#scrollTo=kUv-ecpzJppr
Health AI.ipynb
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands + Code + Text Run all
RAM
Disk
* Running on public URL: https://634f9a7a61dcf3da80.gradio.live
This share link expires in 1 week. For free permanent hosting and GPU upgrades, run `gradio deploy` from the terminal in the working directory to deploy to Hugging Face Spaces (
Eco Assistant & Policy Analyzer
Eco Tips Generator
Policy Summarization
Environmental Problem/Keywords
e.g., plastic, solar, water waste, energy saving...
Generate Eco Tips
Sustainable Living Tips
Variables
Terminal
12:35 PM
T4 (Python 3)
Hot weather
Now
Search
ENG
IN
12:36
15-09-2025