TAIYŌAI INC.

Data Engineering Trial Task

Objective:
Find, standardize, and continuously update data regarding construction and infrastructure projects and tenders in the state of California.

Part 1: Research and Data Sourcing
Task: Research and identify 5-10 reliable data sources about construction and infrastructure projects and tenders in California.

Methodology: Use a combination of online research and language models (e.g., OpenAI's GPT models) to identify these sources. Explicitly state how and why you used GPT or similar models in your research process.

Part 2: Data Extraction and Standardization
Task: From the provided Table 1 and your own list, suggest methods to scrape data using language model-based tools like OpenAI API, Mistral 7B, Llama2, or other open-source models.

Requirements:
Demonstrate how you can build data products (DPs) to scrape data from multiple sources.
Standardize the scraped data according to the guidelines provided in Table 2.

Part 3: Automation and Continuous Updating
Task: Propose a system for automating the data scraping and standardization processes.

Details:
Explain how the data sources will be continuously updated.
Describe the use of cron jobs or similar scheduling tools for ongoing data updates.
Ensure your methodology adheres to a production environment's standards.

Evaluation Criteria

- Scalability: Ability to scrape multiple sources effectively.
- Adherence to Standards: Conformity with the provided data standards; penalties for deviation.

- Automation and Continuity: Quality of the proposal for continuous data updating, including details on cron monitoring and production environment suitability.

Deliverables

Candidates should share a Google Drive folder containing:

1. Python Scripts: The actual code used for data scraping and standardization.
2. Documentation: Detailed explanations of the scripts and methodologies.
3. Sample Datasets: Examples of the data extracted and standardized.
4. Production Environment Plan: A document detailing the implementation of cron monitoring and how the system will operate in a production environment.

Notes to Candidates

- Pay close attention to the data standards and ensure your methods are scalable and suitable for a production environment.
- Clearly articulate your use of AI or machine learning models, specifically in the context of data sourcing and any preprocessing tasks.
- Demonstrate a thoughtful approach to continuous data updating and monitoring.

Task Submission: Kindly fill out this form to submit your task-
https://forms.gle/BGNFXQr4VeJed7ug8

Suggested Data Sources

City	Sour URL
Richmond	https://www.ci.richmond.ca.us/1404/Major-Projects
bakersfield	https://www.bakersfieldcity.us/518/Projects-Programs
wasco	https://www.cityofwasco.org/311/Current-Projects
eureka	https://www.eurekaca.gov/744/Upcoming-Projects
Arcata	https://www.cityofarcata.org/413/Current-City-Construction-Projects
Mckinleyville	https://www.mckinleyvillecsd.com/news-and-project-updates
sanrafael	https://www.cityofsanrafael.org/major-planning-projects-2/
novato	https://www.novato.org/government/community-development/planning-division/planning-projects?locale=en
Millivalley	https://www.cityofmillvalley.org/258/Projects
Riverside	https://riversideca.gov/utilities/projects
moreno	https://www.moval.org/cdd/documents/about-projects.html
Corona	https://www.coronaca.gov/government/departments-divisions/department-of-water-and-power/construction
sacramento	http://www.cityofsacramento.org/public-works/engineering-services/projects
citrus heights	https://www.citrusheights.net/292/Current-Projects
elk grove	https://www.elkgrovecity.org/southeast-policy-area/development-projects

San Bernardino	https://www.sbcity.org/city_hall/community_economic_development/development_projects
Fontana	https://www.fontanaca.gov/765/Current-Projects
Ontario	https://www.ontarioca.gov/Planning/CurrentPlanning
Chula Vista	https://www.chulavistaca.gov/departments/development-services/city-projects
OceanSide	https://www.ci.oceanside.ca.us/government/development-services/engineering/capital-improvement-program/current-projects
san luis obispo	https://www.slocity.org/government/department-directory/parks-and-recreation/current-projects
paso robles	https://www.prcity.com/363/City-Projects
atascadero	https://www.atascadero.org/index.php?option=com_content&view=article&id=652&Itemid=1723
sanmateo	https://www.cityofsanmateo.org/1176/Whats-Happening-in-Development
daily city	https://www.dalycity.org/362/Current-Project-List
lompoc	https://www.cityoflompoc.com/government/departments/economic-community-development/planning-division/major-project-updates
santa maria	https://www.santamariagroup.com/projects
santa clara	https://www.santaclaraca.gov/business-development/development-projects/projects-listing
vacaville	https://www.ci.vacaville.ca.us/government/community-development/major-development-projects?locale=en
vallejo	https://www.cityofvallejo.net/our_city/departments_divisions/planning_development_services/economic_development_department/development_projects
fair field	https://www.fairfield.ca.gov/government/city-departments/community-development/planning-division/development-activity?locale=en, https://www.fairfield.ca.gov/government/city-departments/public-works/capital-improvement-projects
Rohnert Park	https://www.rpcity.org/city_hall/departments/development_services/engineering/projects_in_progress
santa rose	https://www.srcity.org/3212/Current-Projects
petaluma	https://cityofpetaluma.org/planning-projects/
thousand oaks	https://www.toaks.org/departments/public-works/construction

simivalley	https://www.simivalley.org/departments/public-works/public-works-engineering/capital-projects/current-capital-projects
shoreline	https://www.shorelinewa.gov/government/projects-initiativess

Table 2. Data Standards List

Attribute	Description	Expectations
Generic		
original_id	Unique from source	expect_column_to_exist, expect_column_values_to_be_of_type, expect_column_values_to_be_unique
aug_id	Generated by Taiyo (UUID functionality)	expect_column_to_exist, expect_column_values_to_be_of_type, expect_column_values_to_be_unique
country_name	Name of the Country	expect_column_to_exist, expect_column_values_to_be_in_set
country_code	ISO 3-letter Country Code	expect_column_to_exist, expect_column_values_to_be_in_set, expect_column_values_to_be_valid_iso_country
map_coordinates	Geo Point of the region. Should be formatted like this. {"type": "Point", "coordinates": [longitude, latitude]}	expect_column_to_exist, expect_column_values_to_be_of_type, expect_column_values_to_be_valid_geojson
url	Url of the website of the source	expect_column_to_exist, expect_column_values_to_be_valid_urls, expect_column_values_to_be_unique

region_name	Region Name for a Country according to World Bank Standards	expect_column_to_exist, expect_column_values_to_be_in_set
region_code	Region code for a Region according to World Bank Standards	expect_column_to_exist, expect_column_values_to_be_in_set
Projects and Tender		
title	A title for this tender/project. This will often be used by applications as a headline to attract interest, and to help analysts understand the nature of this procurement	expect_column_to_exist, expect_column_values_to_be_unique
description	A summary description of the tender/project. This complements any structured information provided using the items array. Descriptions should be short and easy to read. Avoid using ALL CAPS.	expect_column_to_exist

status	The current status of the tender/project, from the closed tenderStatus codelist	expect_column_to_exist, expect_column_values_to_be_in_set
stages		expect_column_to_exist, expect_column_values_to_be_in_set
date	The date on which the information contained in the release was first recorded in, or published by, any system	expect_column_to_exist, expect_column_values_to_match_strftime_format
procurementMethod	The procurement method is the procedure used to purchase the relevant works, goods or services. The procurement method, from the codelist	expect_column_to_exist, expect_column_values_to_be_in_set
budget	The total upper estimated value of the procurement. A negative value indicates that the contracting process may involve payments from the supplier to the buyer (commonly used in concession contracts)	expect_column_to_exist, expect_column_values_to_be_of_type

currency	The currency for each amount must be specified using the uppercase 3-letter currency code from ISO4217	expect_column_to_exist, expect_column_values_to_be_valid_currency_code
buyer	A buyer is an entity whose budget will be used to pay for goods, works or services related to a contract	expect_column_to_exist
sector	A high-level categorization of the main sector this procurement process relates to. Use of UN COFOG codes, with 'COFOG' as the classification scheme, and the numerical COFOG code is recommended for the primary sector classification.	expect_column_to_exist, expect_column_values_to_be_in_set
subsector	A further subdivision of the sector the procurement process belongs to	expect_column_to_exist expect_column_values_to_be_in_set