README

# URL Extractor for "arcgis/rest/services"

This project extracts URLs containing 'arcgis/rest/services' from Google search results and saves them to a CSV file.

# Prerequisites

Before running the solution, ensure you have the following installed:  
- Python 3.x  
- Requests library  
- BeautifulSoup4 library  
- pandas library  
  
You can install the necessary libraries using pip:

**pip install requests beautifulsoup4 pandas**

# Instructions

## 1. Clone the Repository

Clone the repository or download the script to your local machine.

## 2. Run the Script

Open the Jupyter notebook and open the file I have given to you and run in the new notebook and will see the csv file stored in the folder.

This will execute the script and perform the following steps:  
- Perform a Google search for 'arcgis/rest/services'.  
- Extract URLs containing 'arcgis/rest/services' from the search results.  
- Save the valid URLs to a CSV file named 'urls.csv'.

## 3. Check the Output

After running the script, you will see a confirmation message indicating that the file has been saved successfully:

File 'urls.csv' has been saved successfully.

The 'urls.csv' file will be created in the same directory as the script, containing the extracted URLs.

# Script Overview

## Functions

1. google\_search(query, num\_results=1000000):  
- Performs a Google search with the specified query and returns the HTML content of the search results page.

2. extract\_urls(html):  
- Parses the HTML content to extract URLs containing 'arcgis/rest/services'.

3. is\_valid\_url(url):  
- Validates the extracted URLs to ensure they contain the specified path and are properly formatted.

4. save\_to\_csv(urls, filename='urls.csv'):  
- Saves the cleaned and valid URLs to a CSV file and prints a confirmation message.

## Main Workflow

- Execute the Google search to get the HTML content.  
- Extract the URLs from the HTML content.  
- Save the cleaned and valid URLs to a CSV file and print a success message.

# Notes

The script fetches the Google search results, which may be subject to change based on Google's HTML structure. If the script fails to extract URLs correctly, it may need to be updated to handle changes in the HTML structure.  
Ensure you have a stable internet connection while running the script as it fetches data from the web.