**Get Pageviews for Wikipedia Dump Articles**

**(note:** **cannot get historical pageview counts before July 2015 via this API.)**

1. Get Article Titles from Dump(extract titles After running WikiExtractor)
2. Format Article Titles for API Use(Replace spaces with underscores)
3. Use the Wikimedia Pageviews API
4. Save to CSV

**Code:**

import requests

import time

import csv

from urllib.parse import quote

from bs4 import BeautifulSoup

import glob

# STEP 1: Extract titles from WikiExtractor files

titles = set()

for filename in glob.glob("extracted\_text/\*/\*.txt"):

with open(filename, 'r', encoding='utf-8') as file:

content = file.read()

soup = BeautifulSoup(content, "html.parser")

for doc in soup.find\_all("doc"):

titles.add(doc.get("title"))

titles = list(titles)

# STEP 2: Set up output file and API date range

start\_date = "20240101"

end\_date = "20240601"

with open("pageviews.csv", "w", newline='', encoding="utf-8") as f:

writer = csv.writer(f)

writer.writerow(["title", "date", "views"])

for i, title in enumerate(titles):

safe\_title = quote(title.replace(" ", "\_")) # URL encode title

url = f"https://wikimedia.org/api/rest\_v1/metrics/pageviews/per-article/en.wikipedia/all-access/all-agents/{safe\_title}/daily/{start\_date}/{end\_date}"

try:

response = requests.get(url)

if response.status\_code == 200:

data = response.json()["items"]

for item in data:

writer.writerow([title, item["timestamp"], item["views"]])

else:

print(f"[{i}] Skipped: {title} - {response.status\_code}")

except Exception as e:

print(f"[{i}] Error: {title} - {e}")

time.sleep(0.5) # Politeness delay to avoid rate limiting