Notebook

December 28, 2024

#Embedding-Based Retrieval with Activeloop and OpenAI

This first component of the RAG pipeline collects data and prepares it.

1 Environment

[1]:

```
#!pip install beautifulsoup4==4.12.3
#!pip install requests==2.31.0
```

2 DATA COLLECTION

2.1 Collecting the data

[6]:

```
import os
os.chdir("D:\RAG_Rothman\Chapter02")

import requests
from bs4 import BeautifulSoup
import re

# URLs of the Wikipedia articles
urls = [
    "https://en.wikipedia.org/wiki/Space_exploration",
    "https://en.wikipedia.org/wiki/Apollo_program",
    "https://en.wikipedia.org/wiki/Hubble_Space_Telescope",
    "https://en.wikipedia.org/wiki/Mars_rover",  # Corrected link
    "https://en.wikipedia.org/wiki/International_Space_Station",
    "https://en.wikipedia.org/wiki/SpaceX",
    "https://en.wikipedia.org/wiki/Juno_(spacecraft)",
    "https://en.wikipedia.org/wiki/Voyager_program",
    "https://en.wikipedia.org/wiki/Galileo_(spacecraft)",
    "https://en.wikipedia.org/wiki/Kepler_Space_Telescope",
    "https://en.wikipedia.org/wiki/James_Webb_Space_Telescope",
    "https://en.wikipedia.org/wiki/Space_Shuttle",
    "https://en.wikipedia.org/wiki/Artemis_program",
    "https://en.wikipedia.org/wiki/Skylab",
    "https://en.wikipedia.org/wiki/NASA",
    "https://en.wikipedia.org/wiki/European_Space_Agency",
    "https://en.wikipedia.org/wiki/Ariane_(rocket_family)",
    "https://en.wikipedia.org/wiki/Spitzer_Space_Telescope",
    "https://en.wikipedia.org/wiki/New_Horizons",
    "https://en.wikipedia.org/wiki/Cassini%E2%80%93Huygens",
    "https://en.wikipedia.org/wiki/Curiosity_(rover)",
    "https://en.wikipedia.org/wiki/Perseverance_(rover)",
    "https://en.wikipedia.org/wiki/InSight",
    "https://en.wikipedia.org/wiki/OSIRIS-REx",
    "https://en.wikipedia.org/wiki/Parker_Solar_Probe",
    "https://en.wikipedia.org/wiki/BepiColombo",
    "https://en.wikipedia.org/wiki/Juice_(spacecraft)",
    "https://en.wikipedia.org/wiki/Solar_Orbiter",
    "https://en.wikipedia.org/wiki/CHEOPS_(satellite)",
    "https://en.wikipedia.org/wiki/Gaia_(spacecraft)"
]
```

2.2 Preparing the data

[7]:

```
def clean_text(content):
    # Remove references that usually appear as [1], [2], etc.
    content = re.sub(r'\[\d+\]', '', content)
    return content

def fetch_and_clean(url):
    # Fetch the content of the URL
    response = requests.get(url)
    soup = BeautifulSoup(response.content, 'html.parser')

    # Find the main content of the article, ignoring side boxes and headers
    content = soup.find('div', {'class': 'mw-parser-output'})

    # Remove the bibliography section which generally follows a header like␣
↪"References", "Bibliography"
    for section_title in ['References', 'Bibliography', 'External links', 'See␣
↪also']:
        section = content.find('span', id=section_title)
        if section:
            # Remove all content from this section to the end of the document
            for sib in section.parent.find_next_siblings():
                sib.decompose()
            section.parent.decompose()

    # Extract and clean the text
    text = content.get_text(separator=' ', strip=True)
    text = clean_text(text)
```

```python
    return text

# File to write the clean text
with open('llm.txt', 'w', encoding='utf-8') as file:
    for url in urls:
        clean_article_text = fetch_and_clean(url)
        file.write(clean_article_text + '\n')

print("Content written to llm.txt")
```

```
Content written to llm.txt
```

[ ]:

```python
# Open the file and read the first 20 lines
with open('llm.txt', 'r', encoding='utf-8') as file:
    lines = file.readlines()
    # Print the first 20 lines
    for line in lines[:20]:
        print(line.strip())
```

```
Exploration of space, planets, and moons "Space Exploration" redirects here. For
the company, see SpaceX . For broader coverage of this topic, see Exploration .
Buzz Aldrin taking a core sample of the Moon during the Apollo 11 mission Self-
portrait of Curiosity rover on Mars 's surface Part of a series on Spaceflight
History History of spaceflight Space Race Timeline of spaceflight Space probes
Lunar missions Mars missions Applications Communications Earth observation
Exploration Espionage Military Navigation Settlement Telescopes Tourism
Spacecraft Robotic spacecraft Satellite Space probe Cargo spacecraft Crewed
spacecraft Apollo Lunar Module Space capsules Space Shuttle Space stations
Spaceplanes Vostok Space launch Spaceport Launch pad Expendable and reusable
launch vehicles Escape velocity Non-rocket spacelaunch Spaceflight types Sub-
orbital Orbital Interplanetary Interstellar Intergalactic List of space
organizations Space agencies Space forces Companies Spaceflight portal v t e
Space exploration is the use of astronomy and space technology to explore outer
space . [ 1 ] While the exploration of space is currently carried out mainly by
astronomers with telescopes , its physical exploration is conducted both by
uncrewed robotic space probes and human spaceflight . Space exploration, like
its classical form astronomy , is one of the main sources for space science .
While the observation of objects in space, known as astronomy , predates
reliable recorded history , it was the development of large and relatively
efficient rockets during the mid-twentieth century that allowed physical space
exploration to become a reality. Common rationales for exploring space include
advancing scientific research, national prestige, uniting different nations,
ensuring the future survival of humanity, and developing military and strategic
advantages against other countries. [ 2 ] The early era of space exploration was
driven by a " Space Race " between the Soviet Union and the United States . A
driving force of the start of space exploration was during the Cold War. After
the ability to create nuclear weapons, the narrative of defense/offense left
land and the power to control the air became the focus. Both the Soviet and the
```

Kosmos 559 Unnamed Kosmos 560 Unnamed Skylab 2 Kosmos 561 Nauka-9KS No.1 Meteor-M No.27 Kosmos 562 Kosmos 563 Kosmos 564 Kosmos 565 Kosmos 566 Kosmos 567 Kosmos 568 Kosmos 569 Kosmos 570 Kosmos 571 Kosmos 572 Explorer 49 OPS 6157 Kosmos 573 Kosmos 574 Kosmos 575 OPS 4018 Kosmos 576 Unnamed Molniya-2-6 OPS 8261 ITOS-E Mars 4 Kosmos 577 Mars 5 Skylab 3 Kosmos 578 Mars 6 Mars 7 OPS 8364 Kosmos 579 OPS 7724 Kosmos 580 Intelsat IV F-7 Kosmos 581 Kosmos 582 Molniya-1-24 Kosmos 583 Kosmos 584 Kosmos 585 Kosmos 586 Unnamed Kosmos 587 Soyuz 12 OPS 6275 Kosmos 588 Kosmos 589 Kosmos 590 Kosmos 591 Kosmos 592 Kosmos 593 Kosmos 594 Kosmos 595 Kosmos 596 Kosmos 597 Kosmos 598 Kosmos 599 Kosmos 600 Kosmos 601 Molniya-2-7 Kosmos 602 Explorer 50 Kosmos 603 Kosmos 604 Transit-O 20 Interkosmos 10 Kosmos 605 Kosmos 606 Mariner 10 NOAA-3 Kosmos 607 OPS 6630 OPS 6630/2 OPS 7705 Molniya-1 No.32 Skylab 4 Kosmos 608 Kosmos 609 Kosmos 610 Kosmos 611 Kosmos 612 Kosmos 613 Molniya-1-26 Kosmos 614 Kosmos 615 OPS 9433 OPS 9434 Explorer 51 Kosmos 616 Soyuz 13 Kosmos 617 Kosmos 618 Kosmos 619 Kosmos 620 Kosmos 621 Kosmos 622 Kosmos 623 Kosmos 624 Kosmos 625 Molniya-2-8 Oreol 2 Kosmos 626 Kosmos 627 Payloads are separated by bullets (·), launches by pipes (|). Crewed flights are indicated in underline . Uncatalogued launch failures are listed in italics . Payloads deployed from other spacecraft are denoted in (brackets). Authority control databases : National Germany Israel United States