Title:
First Task Summary Report – Big Data Analytics Intern

Date:
5/11/24

Objective:
Perform exploratory data analysis (EDA) on the provided dataset to understand its structure, identify missing or inconsistent values, and prepare it for further analysis.

Dataset Description:
The dataset contains details of AI-generated art, including artwork ID, artist information, style, creation date, and popularity scores.

---

1. Tasks Completed

Environment Setup:

- Configured Jupyter Notebook and Python IDE.
- Verified access to data repositories.
- Ensured connectivity to required data sources.

Exploratory Data Analysis:

- **Dataset Structure and Data Types:**
  - The dataset contains **12 columns** and **10,000 rows**.
  - Data types include object, float, and datetime formats.
  - Key columns:
    - **Artwork_ID, Artist_Name, Art_Style, Creation_Date, Medium, Popularity_Score.**
- **Summary Statistics:**
  - **Categorical Data:**
    - Most frequent artist: *MidJourney* (appears 694 times).
    - Most common style: *Cubism* (1,050 artworks).
    - Popular tools: *DeepDream* (used 2,032 times).
  - **Numerical Data:**
    - Average Popularity Score: **2,508.19**
    - Score Range: **50.85 to 4,999.62**
- **Missing or Inconsistent Values:**
  - No missing values identified in any columns.

Data Cleaning and Transformation:

- Converted the **Creation_Date** column from string to datetime format for consistency.
- Ensured all columns are clean and appropriately typed.

2. Key Findings

1. **Dataset Observations:**
   - Comprehensive dataset without missing values or inconsistencies.
   - Clear trends in art styles, tools used, and popularity scores.
2. **Insights Gained:**
   - *Cubism* is the most prominent art style.
   - South America is the most frequently represented region.
   - Popular tools like *DeepDream* dominate AI-generated art trends.

3.Learnings

Learnings:

- Improved understanding of EDA workflows, including type conversions and summary report generation.
- Strengthened Python and Pandas skills for large-scale data analysis.

4.PDF file link

https://drive.google.com/file/d/15FBF0m-EUMpREJdbjvsjq8n8UBGFTvP0/view?usp=drive_link

Title: Task 2 -Setting Up Web Scraping Tools

Date:
6/11/24-7/11/24

---

1. Task Overview

Objective:
To install, configure, and demonstrate the functionality of web scraping tools BeautifulSoup and Scrapy, along with documenting the process and dependencies.

Key Deliverables:

1. Three separate test scripts for each tool (BeautifulSoup and Scrapy).
2. A setup guide with step-by-step instructions and screenshots of command-line output

2. Tasks Completed

Library Installation:

1. Installed necessary libraries via pip:

```
pip install beautifulsoup4
```

2. Verified installations:

```
pip list
```

Tool Testing:

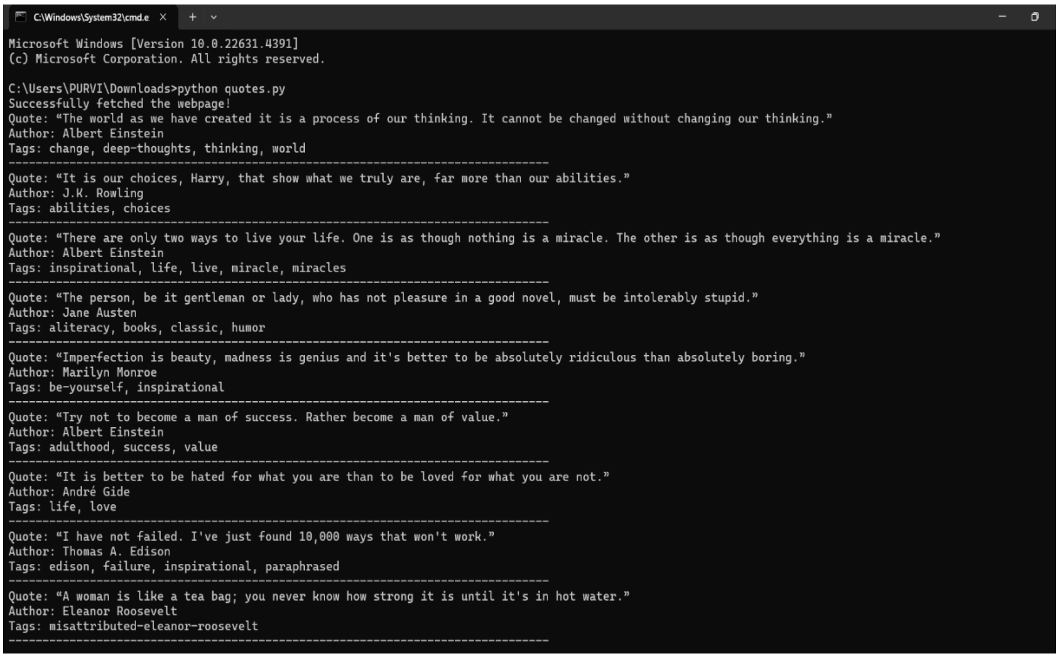
1. BeautifulSoup:

- Wrote a simple script to fetch and parse static HTML from a sample website:

```
import requests

from bs4 import BeautifulSoup

# URL of the webpage to scrape

url = 'http://quotes.toscrape.com/'

# Send GET request to fetch the webpage

response = requests.get(url)

# Check if the request was successful

if response.status_code == 200:

    print("Successfully fetched the webpage!")

    # Parse the HTML content of the page
    soup = BeautifulSoup(response.text, 'html.parser')

    # Find all quote elements on the page

    quotes = soup.find_all('div', class_='quote')

    # Loop through each quote and extract the text, author, and tags

    for quote in quotes:

        # Extract the text of the quote

        text = quote.find('span', class_='text').text

        # Extract the author of the quote

        author = quote.find('small', class_='author').text

        # Extract the tags associated with the quote

        tags = [tag.text for tag in quote.find_all('a', class_='tag')]

        # Print the extracted information

        print(f"Quote: {text}")

        print(f"Author: {author}")

        print(f"Tags: {', '.join(tags)}")

        print('-' * 80)

else:

    print(f"Failed to fetch the webpage. Status code: {response.status_code}")
```

Execution Command:

```
Python quotes.py
```



```
Microsoft Windows [Version 10.0.22631.4391]
(c) Microsoft Corporation. All rights reserved.

C:\Users\PURVI\Downloads>python quotes.py
Successfully fetched the webpage!
Quote: “The world as we have created it is a process of our thinking. It cannot be changed without changing our thinking.”
Author: Albert Einstein
Tags: change, deep-thoughts, thinking, world
--------------------------------------------------------------------------------
Quote: “It is our choices, Harry, that show what we truly are, far more than our abilities.”
Author: J.K. Rowling
Tags: abilities, choices
--------------------------------------------------------------------------------
Quote: “There are only two ways to live your life. One is as though nothing is a miracle. The other is as though everything is a miracle.”
Author: Albert Einstein
Tags: inspirational, life, live, miracle, miracles
--------------------------------------------------------------------------------
Quote: “The person, be it gentleman or lady, who has not pleasure in a good novel, must be intolerably stupid.”
Author: Jane Austen
Tags: aliteracy, books, classic, humor
--------------------------------------------------------------------------------
Quote: “Imperfection is beauty, madness is genius and it's better to be absolutely ridiculous than absolutely boring.”
Author: Marilyn Monroe
Tags: be-yourself, inspirational
--------------------------------------------------------------------------------
Quote: “Try not to become a man of success. Rather become a man of value.”
Author: Albert Einstein
Tags: adulthood, success, value
--------------------------------------------------------------------------------
Quote: “It is better to be hated for what you are than to be loved for what you are not.”
Author: André Gide
Tags: life, love
--------------------------------------------------------------------------------
Quote: “I have not failed. I've just found 10,000 ways that won't work.”
Author: Thomas A. Edison
Tags: edison, failure, inspirational, paraphrased
--------------------------------------------------------------------------------
Quote: “A woman is like a tea bag; you never know how strong it is until it's in hot water.”
Author: Eleanor Roosevelt
Tags: misattributed-eleanor-roosevelt
--------------------------------------------------------------------------------
```

2. Scrapy:

- Created a basic Scrapy project and crawler:

To get started, first install Scrapy:

```
pip install scrapy
```

Then create a Scrapy project:

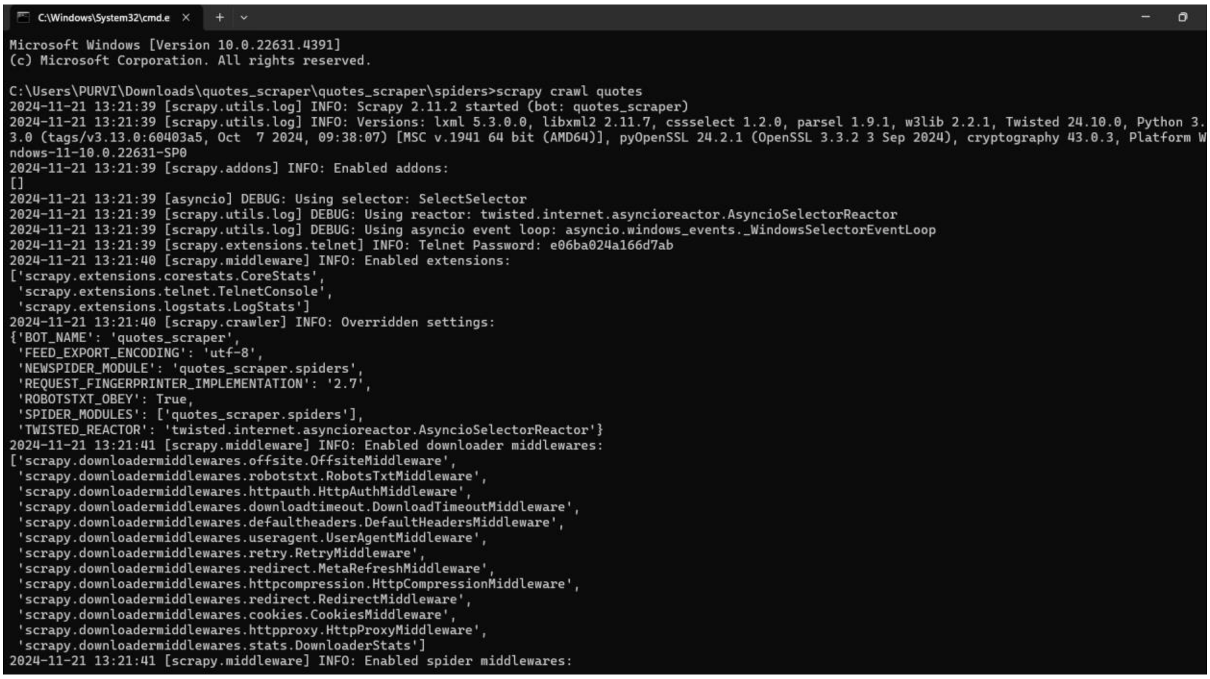
```
scrapy startproject quotes_scraper
```

Edited the spider to scrape titles from a static website:

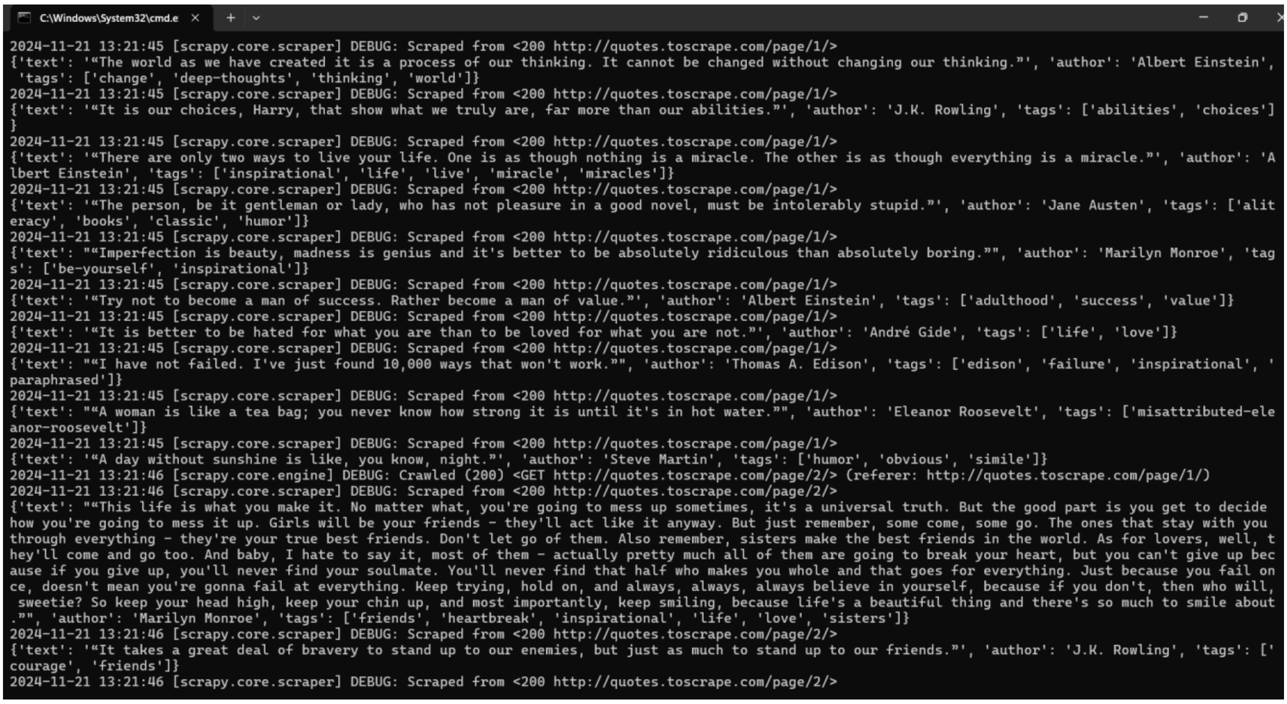
```
import scrapy

class QuotesSpider(scrapy.Spider):

  name = "quotes"  # Name of the spider

  start_urls = [

    'http://quotes.toscrape.com/page/1/',  # Starting URL

  ]

  def parse(self, response):

    # Loop through each quote block

    for quote in response.css('div.quote'):

      # Extract text, author, and tags

      yield {

        'text': quote.css('span.text::text').get(),

        'author': quote.css('small.author::text').get(),

        'tags': quote.css('div.tags a.tag::text').getall(),

      }

    # Follow the "Next" page link if it exists

    next_page = response.css('li.next a::attr(href)').get()

    if next_page is not None:

      yield response.follow(next_page, self.parse)
```

- **Execution Command:**

```
scrapy crawl quotes
```



```
Microsoft Windows [Version 10.0.22631.4391]
(c) Microsoft Corporation. All rights reserved.

C:\Users\PURVI\Downloads\quotes_scraper\quotes_scraper\spiders>scrapy crawl quotes
2024-11-21 13:21:39 [scrapy.utils.log] INFO: Scrapy 2.11.2 started (bot: quotes_scraper)
2024-11-21 13:21:39 [scrapy.utils.log] INFO: Versions: lxml 5.3.0.0, libxml2 2.11.7, cssselect 1.2.0, parsel 1.9.1, w3lib 2.2.1, Twisted 24.10.0, Python 3.
3.0 (tags/v3.13.0:60403a5, Oct  7 2024, 09:38:07) [MSC v.1941 64 bit (AMD64)], pyOpenSSL 24.2.1 (OpenSSL 3.3.2 3 Sep 2024), cryptography 43.0.3, Platform W
ndows-11-10.0.22631-SP0
2024-11-21 13:21:39 [scrapy.addons] INFO: Enabled addons:
[]
2024-11-21 13:21:39 [asyncio] DEBUG: Using selector: SelectSelector
2024-11-21 13:21:39 [scrapy.utils.log] DEBUG: Using reactor: twisted.internet.asyncioreactor.AsyncioSelectorReactor
2024-11-21 13:21:39 [scrapy.utils.log] DEBUG: Using asyncio event loop: asyncio.windows_events._WindowsSelectorEventLoop
2024-11-21 13:21:39 [scrapy.extensions.telnet] INFO: Telnet Password: e06ba024a166d7ab
2024-11-21 13:21:40 [scrapy.middleware] INFO: Enabled extensions:
['scrapy.extensions.corestats.CoreStats',
 'scrapy.extensions.telnet.TelnetConsole',
 'scrapy.extensions.logstats.LogStats']
2024-11-21 13:21:40 [scrapy.crawler] INFO: Overridden settings:
{'BOT_NAME': 'quotes_scraper',
 'FEED_EXPORT_ENCODING': 'utf-8',
 'NEWSPIDER_MODULE': 'quotes_scraper.spiders',
 'REQUEST_FINGERPRINTER_IMPLEMENTATION': '2.7',
 'ROBOTSTXT_OBEY': True,
 'SPIDER_MODULES': ['quotes_scraper.spiders'],
 'TWISTED_REACTOR': 'twisted.internet.asyncioreactor.AsyncioSelectorReactor'}
2024-11-21 13:21:41 [scrapy.middleware] INFO: Enabled downloader middlewares:
['scrapy.downloadermiddlewares.offsite.OffsiteMiddleware',
 'scrapy.downloadermiddlewares.robotstxt.RobotsTxtMiddleware',
 'scrapy.downloadermiddlewares.httpauth.HttpAuthMiddleware',
 'scrapy.downloadermiddlewares.downloadtimeout.DownloadTimeoutMiddleware',
 'scrapy.downloadermiddlewares.defaultheaders.DefaultHeadersMiddleware',
 'scrapy.downloadermiddlewares.useragent.UserAgentMiddleware',
 'scrapy.downloadermiddlewares.retry.RetryMiddleware',
 'scrapy.downloadermiddlewares.redirect.MetaRefreshMiddleware',
 'scrapy.downloadermiddlewares.httpcompression.HttpCompressionMiddleware',
 'scrapy.downloadermiddlewares.redirect.RedirectMiddleware',
 'scrapy.downloadermiddlewares.cookies.CookiesMiddleware',
 'scrapy.downloadermiddlewares.httpproxy.HttpProxyMiddleware',
 'scrapy.downloadermiddlewares.stats.DownloaderStats']
2024-11-21 13:21:41 [scrapy.middleware] INFO: Enabled spider middlewares:
```



```
2024-11-21 13:21:45 [scrapy.core.scraper] DEBUG: Scraped from <200 http://quotes.toscrape.com/page/1/>
{'text': '“The world as we have created it is a process of our thinking. It cannot be changed without changing our thinking.”', 'author': 'Albert Einstein',
 'tags': ['change', 'deep-thoughts', 'thinking', 'world']}
2024-11-21 13:21:45 [scrapy.core.scraper] DEBUG: Scraped from <200 http://quotes.toscrape.com/page/1/>
{'text': '“It is our choices, Harry, that show what we truly are, far more than our abilities.”', 'author': 'J.K. Rowling', 'tags': ['abilities', 'choices']
}
2024-11-21 13:21:45 [scrapy.core.scraper] DEBUG: Scraped from <200 http://quotes.toscrape.com/page/1/>
{'text': '“There are only two ways to live your life. One is as though nothing is a miracle. The other is as though everything is a miracle.”', 'author': 'A
lbert Einstein', 'tags': ['inspirational', 'life', 'live', 'miracle', 'miracles']}
2024-11-21 13:21:45 [scrapy.core.scraper] DEBUG: Scraped from <200 http://quotes.toscrape.com/page/1/>
{'text': '“The person, be it gentleman or lady, who has not pleasure in a good novel, must be intolerably stupid.”', 'author': 'Jane Austen', 'tags': ['alit
eracy', 'books', 'classic', 'humor']}
2024-11-21 13:21:45 [scrapy.core.scraper] DEBUG: Scraped from <200 http://quotes.toscrape.com/page/1/>
{'text': "“Imperfection is beauty, madness is genius and it's better to be absolutely ridiculous than absolutely boring.”", 'author': 'Marilyn Monroe', 'tag
s': ['be-yourself', 'inspirational']}
2024-11-21 13:21:45 [scrapy.core.scraper] DEBUG: Scraped from <200 http://quotes.toscrape.com/page/1/>
{'text': '“Try not to become a man of success. Rather become a man of value.”', 'author': 'Albert Einstein', 'tags': ['adulthood', 'success', 'value']}
2024-11-21 13:21:45 [scrapy.core.scraper] DEBUG: Scraped from <200 http://quotes.toscrape.com/page/1/>
{'text': '“It is better to be hated for what you are than to be loved for what you are not.”', 'author': 'André Gide', 'tags': ['life', 'love']}
2024-11-21 13:21:45 [scrapy.core.scraper] DEBUG: Scraped from <200 http://quotes.toscrape.com/page/1/>
{'text': "“I have not failed. I've just found 10,000 ways that won't work.”", 'author': 'Thomas A. Edison', 'tags': ['edison', 'failure', 'inspirational', '
paraphrased']}
2024-11-21 13:21:45 [scrapy.core.scraper] DEBUG: Scraped from <200 http://quotes.toscrape.com/page/1/>
{'text': "“A woman is like a tea bag; you never know how strong it is until it's in hot water.”", 'author': 'Eleanor Roosevelt', 'tags': ['misattributed-ele
anor-roosevelt']}
2024-11-21 13:21:45 [scrapy.core.scraper] DEBUG: Scraped from <200 http://quotes.toscrape.com/page/1/>
{'text': '“A day without sunshine is like, you know, night.”', 'author': 'Steve Martin', 'tags': ['humor', 'obvious', 'simile']}
2024-11-21 13:21:46 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://quotes.toscrape.com/page/2/> (referer: http://quotes.toscrape.com/page/1/)
2024-11-21 13:21:46 [scrapy.core.scraper] DEBUG: Scraped from <200 http://quotes.toscrape.com/page/2/>
{'text': "“This life is what you make it. No matter what, you're going to mess up sometimes, it's a universal truth. But the good part is you get to decide
how you're going to mess it up. Girls will be your friends - they'll act like it anyway. But just remember, some come, some go. The ones that stay with you
through everything - they're your true best friends. Don't let go of them. Also remember, sisters make the best friends in the world. As for lovers, well, t
hey'll come and go too. And baby, I hate to say it, most of them - actually pretty much all of them are going to break your heart, but you can't give up bec
ause if you give up, you'll never find your soulmate. You'll never find that half who makes you whole and that goes for everything. Just because you fail on
ce, doesn't mean you're gonna fail at everything. Keep trying, hold on, and always, always, always believe in yourself, because if you don't, then who will,
 sweetie? So keep your head high, keep your chin up, and most importantly, keep smiling, because life's a beautiful thing and there's so much to smile about
.”", 'author': 'Marilyn Monroe', 'tags': ['friends', 'heartbreak', 'inspirational', 'life', 'love', 'sisters']}
2024-11-21 13:21:46 [scrapy.core.scraper] DEBUG: Scraped from <200 http://quotes.toscrape.com/page/2/>
{'text': '“It takes a great deal of bravery to stand up to our enemies, but just as much to stand up to our friends.”', 'author': 'J.K. Rowling', 'tags': ['
courage', 'friends']}
2024-11-21 13:21:46 [scrapy.core.scraper] DEBUG: Scraped from <200 http://quotes.toscrape.com/page/2/>
```

4. Key Findings

- **BeautifulSoup:** Ideal for static websites or small-scale scraping tasks.
- **Scrapy:** Best suited for complex projects with multi-page scraping requirements.