PDV Assignment

Name : Karthik. C
Reg No: 24 1058016
Big Data Analytics

1) Describe the primary purpose of a User-Agent header in web Scraping

→ User-Agent header in web scrapping plays a crucial role in mimicking the behavior of a real web browser

Primary purpose of the User-Agent Header

1. Simulate a Real User/Browser

The User-Agent header tells the server what type of device or browser is making the request. when scrapping, you can set this header to resemble a real browser (like Chrome, Firefox etc) to prevent the server from blocking the request as suspicious or coming from an automated tool.

2. Bypass Anti-Scrapping Mechanisms

- Many websites use anti-bot measures that detect and block requests with non-standard or missing User-Agent headers, as these are often associated with web scrapers or bots.

3. Handle Different Device or Browser-Specific Content:

→ Some websites deliver different content depending on the browser or device requesting it. By including a User-Agent header that mimics a specific browser.

4. Respect Website's Terms and Conditions:

→ while not a guarantee of respect, using a valid User-Agent header that identifies your scraper as a specific browser can indicate to the server that the request is coming legitimate.

2) what is the purpose of a robots.txt file on a website?

-> The robot.txt file is a text file placed in the root directories of a website that provides guidline for web crawlers access to website. It uses the robots Exclusion protocol & can include rules to restrict specific crawlength or path on the site.

Ex:- user-agent: *

Disallow: /admin/

Disallow: /private/

*Disallow:- Indicates which pages or directories should not be crawled by the specified user-agent. If the value is left empty (i.e. Disallow), it means the crawler is allowed to access all parts of the site.

3). Scrape data contents from a single web page using Beautiful soup

-> Here's an example of how to scrape data from a webpage

```
from bs4 import BeautifulSoup
import requests

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

response = requests.get(url)

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

quotes = soup.find_all("div", class_="quotes")

for quote in quotes:

    text = quote.find("span", class_="text").get_text()

    author = quote.find("small", class_="author").get_text()

    print(f"Quote: {text} | Author: {author}\n")
```

4. Scrape data content navigating through multiple web page using Beautiful soup library.

→ from bs4 import BeautifulSoup

import Requests

base_url = "http://quotes.toscrape.com/page1"

for page in range (1,6):

url = f("{base_url}{page}/")"

response = request.get (url)

soup = BeautifulSoup (response.cont, "html.parser")

quotes = soup.find_all ("div", class = "quote")

for quote in quotes:

text = quote.find ("span", class = "text").get.text()

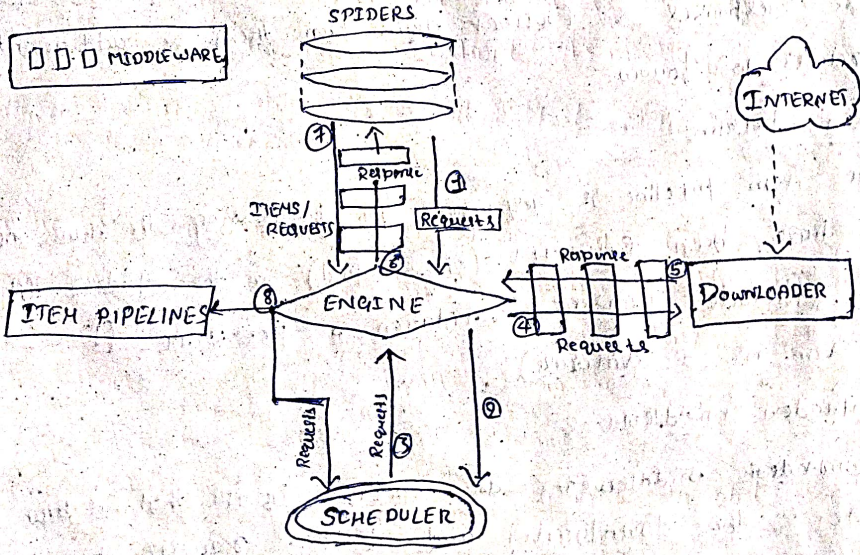
author = quote.find ("small", class_ = "author").get.text().

print (f"page {page}. Quote : {text.} \n

author : {author} \n") ;

5. Describe the architecture of scrapy framework.

→



The data flow in Scrapy is controlled by the Execution Engine, and goes like this:

* Scrapy Engine:
The Engine is responsible for controlling the data flow btw all components of the system, and triggering events when certain action occurs. See the Data flow section above for more details.

* Scheduler :-
The scheduler receives requests from the Engine and Enqueues them for feeding them later (also to the Engine) when the Engine requests them.

* Downloader :-
The Downloader is responsible for fetching web pages and feeding them to the Engine which, in turn, feeds them to the Spiders.

* Spiders
Spiders are custom classes written by Scrapy users to parse responses & Extract items from them or additional requests to follow.

* Item Pipeline
The item pipeline is responsible for processing the items once they have been Extracted (or Scraped) by the Spiders. Typical tasks include cleansing, validation & persistence (like storing the item in a database).

Downloader middlewares
* Downloader middlewares are specific hooks that sit btw the Engine & the downloader and process requests when they pass from the Engine to the Downloader & responses that pass from Downloader to the Engine.

Spider middlewares

Spider middlewears are specific hooks that sit blw the Engine & the spiders and are able to process spider input (responses) and output (items and requests).

6. Demonstrate Scrapy framework with an example to store information about a product having attributes.

→ Set up Scrapy project

Scrapy startproject shopping_scrapers

It will create a folder structure

Inside the shopping_scrapers/items.py

```
import scrapy
class ProductItem (Scrapy.Item)
    name = scrapy.Field()
    cost = scrapy.Field()
  manufacturer = scrapy.Field()
    rating = scrapy.Field()
```

shopping_scrappers/spiders -> create new python file. product_spider.py.

```
import scrapy
from shopping_scrapers.items import ProductItem
class ProductSpider (Scrapy.Spider):
    name = 'product_spider'
    start_urls = ['https://www.shopping-tv.com']
    def parse (self, response):
        for product in response.css ("product"):
            yield {
                "name": product.css ("name::text").get()
                "cost": product.css ("cost::text").get()
                "manufacture": product.css ("manufacture::text").get()
```

```
        "rating" : product.css ("rating ::text "). get()
    }

    next_page = response.css ('a.next-page::attr(href)'). get()
    if next_page:
        yield response. follow (next_page, self.parse)
```

7) Create a pandas dataframe from Super Market Invoice Data, and Execute following operations.

i) Load the data

```
import pandas as pd.
df = pd. read_csv ("supermarket_invoice_data.csv")
```

a) Count the number of customers per product category or department.

```
customer_count = df. groupby ('Description') ['customer ID']
                 . nunique()
Print (customer_count).
```

b) Calculate the average customer visit per product category.

```
avg_visits = df. groupby ('Description') ['customer ID']. value_counts().
             groupby () mean()
Print (avg_visits)
```

c) Identify the product bought by the most customers

```
Product_bought = df. groupby ('Description') ['customer ID']. nunique().
                 idxmax()
Print (Product_bought)
```

d) Sort customers based on age, products purchased, or products cost.

```
Sorted_customer = df. sort_values (by = ["sales", "Quantity"],
                  ascends = [False, False])
Print (sorted_customer)
```

Explain categories of Exploratory visualizations based on the relationships b/w the three necessary players.

1. Univariate visualizations: (Single Variable)

These visualizations explore the characteristics of a singl Variable. The relationship here is b/w a single variable & its visual representation, where the users (observers) interpret patterns.

Examples:

- Histogram : Display the distribution of a continuous variable
- Bar charts : Show frequency counts or categories for a discrete variable

2. Bivariate Visualizations (Two Variables)

These visualizations focus on the relationship b/w two variables, either numerical or categorical the key relationship here is between two variables & their visual mapping.

3. Multivariate visualizations (multiple variables)

When exploring the relationships b/w three or more variable the visualization often becomes more complex. the user is tasked with interpreting multiple data dimensions simultaneously. here the visual representation could employ colors, shape, size or even animation to represent more than

Define DAX. Demonstrate any 10 DAX formulas with a context.

1. Count :- Counts the number of non-blank rows in a column Typically used to count the number of rows with valid entries in a dataset.

2. Distinct Count :-

Counts the distinct (unique) values in a column

Useful for identifying the number of unique entries, such as distinct customers, products, or transactions.

3. Average :-

Calculates the average of a column.

Used to find the mean value, such as average sales or average quantities.

4. MIN :-

Returns the smallest value from a column.

Useful for identifying the minimum numerical or date value, such as the lowest sales amount or earliest date.

5. MAX :-

Returns the largest value from a column.

Helps identify the maximum value in a dataset, such as highest sales or latest date.

6. SUMMARIZE :-

Returns a summary table for the requested columns and aggregated data.

Used to group data and apply aggregations like SUM or COUNT.

7. CALCULATE :-

Modifies the context of a calculation with filters.

Used to calculate results under specific conditions, such as sales for a specific product category.

8. IF

Returns one value if a condition is true, & another if it's false.

Enables conditional calculations, such as classifying sales as high or

9. SUM :-

Sums the values in a column used to add up numerical values such as sales amounts, quantities, or profits.

10. ~~IFERROR~~ :- SUMX :-

Returns a specified result if an expression for each row summing the results. Enables summation of complex expressions across rows, such as calculating total profits for each transaction.

11. SWITCH :-

Evaluates an expression against a list of values and returns the corresponding result. provides a simplified way to test multiple conditions, like a case statement.

Discuss ten differences b/w calculated measure & calculated column.

Calculated Measure	Calculated column.
* Calculated on the fly based on context	* Row-wise calculation stored in the model.
* Based on filter and slicer context	* Calculated for each row, independent of filters.
* It not stored: calculated dynamically	* Stored in the data model.
* Used in aggregations and summaries	* Used for row-level calculations or new dimensions
* Performance impact recalculated each time it's used	* Calculated once, stored at load time.
* Filter use cannot be used directly in filters	* Can be used directly in filters & slicers.

* No impact on model structure	* Adds new columns to the model
* It is Dynamic and changes with context	* Static once calculated.
* It operates on aggregates or summaries	* Operates on individual rows.
* use function like SUM() Average(), etc.	* often use IF(), concatenation(), etc.