https://github.com/Hanan20/data-science-assignment.git

Number 2
Explain the different stages of data science and advance your research in data Acquisition and explain web scraping in detail citing at least 5 research papers. Publish your work in an online journal and in your paper submission, please include the link to the journal

Data science is the study of data to extract meaningful insights for business through data gathering, analysis and decision making.

Data science encompasses a series of stages that involve problem identification, business understanding, collecting data, pre-processing data ,analyzing data, data modeling, model evaluation, model training , model deployment and interpreting data to extract meaningful insights. While the exact stages may vary, a typical data science workflow includes:

1. Problem identification
This is the most important stage of any Data Science endeavor; The first step is to understand how Data Science is useful in the domain under consideration and to identify appropriate tasks that are useful for the same. Domain specialists and data scientists play critical roles in problem identification. The domain expert is well-versed in the application domain and understands the challenge at hand. Data Scientists understand the area and can assist in the discovery of problems and viable solutions.

2. Business Understanding

The business goals are formed by the customer's need to make predictions, boost sales, minimize losses, or optimize any given process, among other things.

3. Collecting Data:

Data collection is a vital phase since it serves as the foundation for achieving specified business objectives. In general, the information gleaned from surveys is valuable. Data is recorded in various software systems used in the company at various stages, which is vital for understanding the process from product development through deployment and delivery. Historical data from archives is also useful for better understanding the business.

4. Pre-Processing Data:

Large amounts of data are gathered through archives, everyday transactions, and intermediate records. The data is available in a variety of formats and forms. Some data may also be provided in hard copy format. The information is dispersed among multiple servers. All of this data is extracted, transformed, and processed into a single format. Typically, a data warehouse is built to house the Extract, Transform, and Load (ETL) process or processes.

5. Analyzing Data:

Now that the data is available and ready in the required format, the next critical step is to thoroughly grasp the data. This understanding is derived via data analysis utilizing various statistical tools. This is also known as exploratory data analysis (EDA). The data is investigated here by creating various statistical functions and identifying dependent and independent variables or features. Data analysis reveals which data or features are essential, as well as the distribution of data.

6. Data Modelling:

After the data has been analyzed and visualized, the next critical step is data modeling. The key components are kept in the dataset, and therefore the data gets refined. The crucial thing now is to decide how to model the data. What tasks lend themselves well to modeling Which activities, like classification or regression are appropriate is determined by the amount of commercial value required. The Machine Learning engineer generates the result by applying various algorithms to the data. Many times, while modeling data, the models are first tested using dummy data that is similar to the actual data.

7. Model Evaluation/ Monitoring:

Because there are numerous methods for modeling data, it is critical to determine which one is most effective. The model is now being tested with real-world data. When there are few data points, the output is monitored for improvement. While the model is being evaluated or tested, data may change, and the output may alter dramatically as a result.

8. Model Training:

The crucial stage is to train the model once the task and model have been determined, as well as the data drift analysis modeling. Training can be done in steps, with the relevant parameters fine-tuned to obtain the needed accuracy. During the production phase, the model is exposed to actual data and its output is monitored.

9. Model Deployment

The model is now exposed to real-time data entering the system, and output is generated. The model can be deployed as a web service, an embedded application in an edge application, or a mobile application. This is a critical phase since the model is now exposed to the real world.

10.interpretation

Following model deployment in the real world, the next stage is to determine how the model behaves in a real-life setting.this includes Interpreting Results: Translating technical findings into actionable insights.Visualization and Reporting: Communicating findings to stakeholders using clear and concise visuals and reports.

data Acquisition and web scraping

What is data acquisition?

Data acquisition is the process of collecting, measuring, and storing information from various sources. This information can be in the form of physical signals (analog or digital), sensor outputs, or data from other devices and systems. The goal of data acquisition is to convert real-world phenomena into digital data that can be analyzed, processed, and used for various purposes such as monitoring, control, and analysis.

Data acquisition methods

1.direct Measurement:
Description: Involves the direct measurement of physical quantities using sensors or instruments.
Example: Using a thermometer to measure temperature, a pressure sensor to measure pressure, or a voltmeter to measure voltage.

2.Sensor Networks:
Description: Deploying a network of sensors to monitor and collect data from a wide area.
Example: Environmental monitoring using a network of weather sensors to collect temperature, humidity, and air pressure data.

3.Lab Experiments:
Description: Conducting controlled experiments in a laboratory environment to gather data under specific conditions.
Example: Testing the tensile strength of materials under different loads in a materials science laboratory.

4.Remote Sensing:
Description: Collecting data from a distance using sensors mounted on satellites, aircraft, or other platforms.
Example: Satellite imagery for monitoring land use, climate, and environmental changes.

5.Surveys and Questionnaires:
Description: Gathering data through structured interviews, questionnaires, or surveys.
Example: Conducting customer satisfaction surveys or market research to collect opinions and preferences.

6.Web Scraping:
Description: Extracting data from websites by parsing HTML or using web APIs.
Example: Scraping product prices from e-commerce websites for competitive analysis.

7.Machine Learning and Predictive Modeling:
Description: Training models to predict or estimate data based on historical patterns.
Example: Using machine learning algorithms to predict stock prices or demand for a product.

8.Data Logging:
Description: Continuous recording of data over time using data loggers or logging systems.
Example: Monitoring temperature changes in a warehouse using temperature data loggers.

9.Simulation and Modeling:
Description: Generating data through computer simulations or mathematical models.
Example: Simulating the behavior of a new aircraft design in a virtual environment to gather performance data.

10.Social Media Monitoring:
Description: Extracting and analyzing data from social media platforms to understand trends and public sentiment.
Example: Analyzing Twitter data to gauge public reaction to a specific event or product release.

11.IoT (Internet of Things):
Description: Collecting data from interconnected devices in the IoT ecosystem.
Example: Monitoring and controlling smart home devices or tracking assets in a supply chain.

Key components of a typical data acquisition system include:

Sensors and Transducers: These devices convert physical signals, such as temperature, pressure, voltage, or light, into electrical signals.

Signal Conditioning: The acquired signals often need to be conditioned or modified to meet the requirements of the data acquisition system. This may involve amplification, filtering, or other adjustments.

Analog-to-Digital Conversion (ADC): Analog signals from sensors are converted into digital form so that they can be processed by a computer or other digital devices.

Data Processing: Once the data is in digital form, it can be processed, analyzed, and stored by a computer or other digital processing unit.

Data Storage: The acquired and processed data is typically stored for future analysis or reference.

Data Display and Analysis: The final step involves presenting the data in a meaningful way through graphical displays, reports, or other formats, allowing researchers, engineers, or analysts to draw conclusions from the data.

web scraping

Techniques

1.Human copy-and-paste

The simplest form of web scraping is manually copying and pasting data from a web page into a text file or spreadsheet. Sometimes even the best web-scraping technology cannot replace a human's manual examination and copy-and-paste, and sometimes this may be the only workable solution when the websites for scraping explicitly set up barriers to prevent machine automation.

2.Text pattern matching

A simple yet powerful approach to extract information from web pages can be based on the UNIX grep command or regular expression-matching facilities of programming languages (for instance Perl or Python).

3.HTTP programming

Static and dynamic web pages can be retrieved by posting HTTP requests to the remote web server using socket programming.

4.HTML parsing

Many websites have large collections of pages generated dynamically from an underlying structured source like a database. Data of the same category are typically

encoded into similar pages by a common script or template. In data mining, a program that detects such templates in a particular information source, extracts its content, and translates it into a relational form, is called a wrapper. Wrapper generation algorithms assume that input pages of a wrapper induction system conform to a common template and that they can be easily identified in terms of a URL common scheme.Moreover, some semi-structured data query languages, such as XQuery and the HTQL, can be used to parse HTML pages and to retrieve and transform page content.

5.DOM parsing

By embedding a full-fledged web browser, such as the Internet Explorer or the Mozilla browser control, programs can retrieve the dynamic content generated by client-side scripts. These browser controls also parse web pages into a DOM tree, based on which programs can retrieve parts of the pages. Languages such as Xpath can be used to parse the resulting DOM tree.

6.Vertical aggregation

There are several companies that have developed vertical specific harvesting platforms. These platforms create and monitor a multitude of "bots" for specific verticals with no "man in the loop" (no direct human involvement), and no work related to a specific target site. The preparation involves establishing the knowledge base for the entire vertical and then the platform creates the bots automatically. The platform's robustness is measured by the quality of the information it retrieves (usually number of fields) and its scalability (how quick it can scale up to hundreds or thousands of sites). This scalability is mostly used to target the Long Tail of sites that common aggregators find complicated or too labor-intensive to harvest content from.

7.Semantic annotation recognizing

The pages being scraped may embrace metadata or semantic markups and annotations, which can be used to locate specific data snippets. If the annotations are embedded in the pages, as Microformat does, this technique can be viewed as a special case of DOM parsing. In another case, the annotations, organized into a semantic layer,are stored and managed separately from the web pages, so the scrapers can retrieve data schema and instructions from this layer before scraping the pages.

8.Computer vision web-page analysis

There are efforts using machine learning and computer vision that attempt to identify and extract information from web pages by interpreting pages visually as a human being might.

Web Scraping used for?

1. Price Monitoring
2. Market Research
3. News Monitoring
4. Sentiment Analysis
5. Email Marketing

Different Types of Web Scrapers

1.Self-built Web Scrapers
2.Browser extensions Web Scrapers
3.Cloud Web Scrapers

References:

https://en.wikipedia.org/wiki/Web_scraping
https://www.simplilearn.com/data-acquisition-article
https://www.analyticsinsight.net/what-is-data-science-life-cycle-steps-explained/
https://www.geeksforgeeks.org/data-science-lifecycle/
https://www.codecademy.com/article/intro-to-data-acquisition