Tho Le

CS 1030 -003

May 5, 2023

Data tools

These days, we can hardly imagine life without applications that use databases built on structured query language (SQL). When saving information that you don’t care about statistics, we usually use a simple text file. Using tools like SQL databases or Excel, we can efficiently perform functions like counting, averaging, and summing up data in rows and columns to make sense of everything. Datasets help us uncover hidden patterns, track trends, and even pinpoint that one specific case we're after. It's no exaggeration to say that databases and datasets are a huge part of our modern lives, making countless tasks so much easier. It's challenging to find an app that doesn't use a database. With technology ever-evolving, we can expect databases to stick around and keep making our lives better for the foreseeable future.

Big Data

With where we are with technology, it’s important to think about when a computing system needs to store massive amounts of data because we need to consider how the data will be stored and how to process data efficiently. As technology advances and the amount of data generated and collected grows exponentially, efficient storage and processing of massive data, become crucial. Organizations must carefully select scalable storage solutions, employ efficient algorithms and parallel computing techniques, and implement proper data management to ensure data quality. Security, privacy, and ethical considerations, such as data encryption, access controls, and transparency, should also be considered to maintain data integrity and user trust. By addressing these concerns, organizations can leverage big data to drive innovation, enhance

decision-making, and unlock new opportunities while adhering to ethical standards and ensuring data security and privacy.

Bias in Machine Learning

This module offers a comprehensive introduction to machine learning, encompassing fundamental concepts such as reinforcement, unsupervised, and supervised learning. It delves into neural networks, a widely used technique in supervised learning, outlining their architecture and training processes. The significance of diverse and high-quality training data is underscored, emphasizing the need to tackle potential biases in the data to guarantee equitable and precise outcomes. Artificial intelligence enables us to accomplish remarkable tasks like identity recognition, ethnicity detection, and emotion recognition. However, current AI systems grapple with biases and fail to operate fairly for all individuals. AI faces significant accuracy issues in the criminal justice system, leading to the wrongful identification and arrest of many people of color due to flawed facial recognition technology.

Unit testing

The unit test was a great way to review my learning in the three modules. I scored 9 out of 9. It was pretty easy because throughout reading the modules, Khan Academy gives you practice questions to reinforce what I've just read. Some of the questions caught me off guard because it seemed like there were multiple right choices, and there was only one right choice. The unit test also does a good job of telling you whether you got the answer right before moving on to the next. It also gives you a reason why all the other answers are wrong.

Data Analysis - Application and Presentation

Question

What are the top 20 countries with the most content on Netflix?

Process

I started this process by finding a Netflix data set on Kaggle that gives me the necessary information, like the country in which the content is created in. This data contains 6234 rows and 12 columns. After I download the dataset as a .csv file. I used Python and the pandas module to process data on a Kaggle notebook. After I imported the dataset as a data frame, I used the matplotlib module to draw a visual presentation of how many movies are created in each country and limit it to the highest 20.

Here's the link to the data set on Kaggle:
https://www.kaggle.com/datasets/shivamb/netflix-shows?datasetId=434238&searchQuery=data+analysis

Result

The data shows that the United States has the most content on Netflix, with over 4000 movies and shows. Although India comes in second place with the most content, most countries in the top 20 have similar amount of contents.