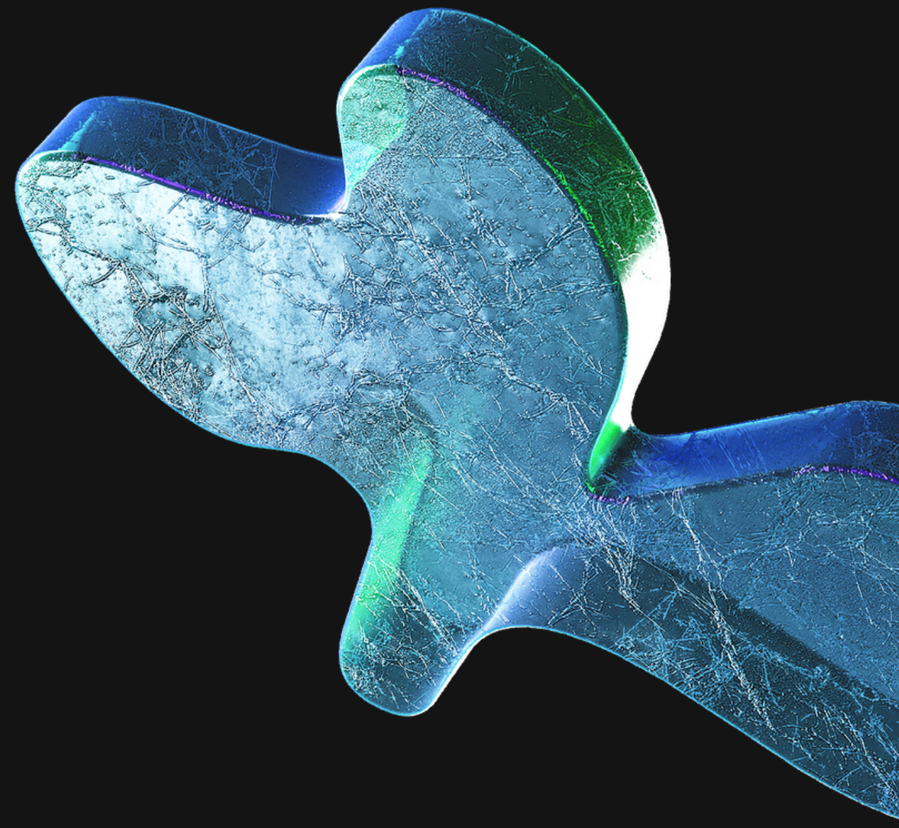
Internship

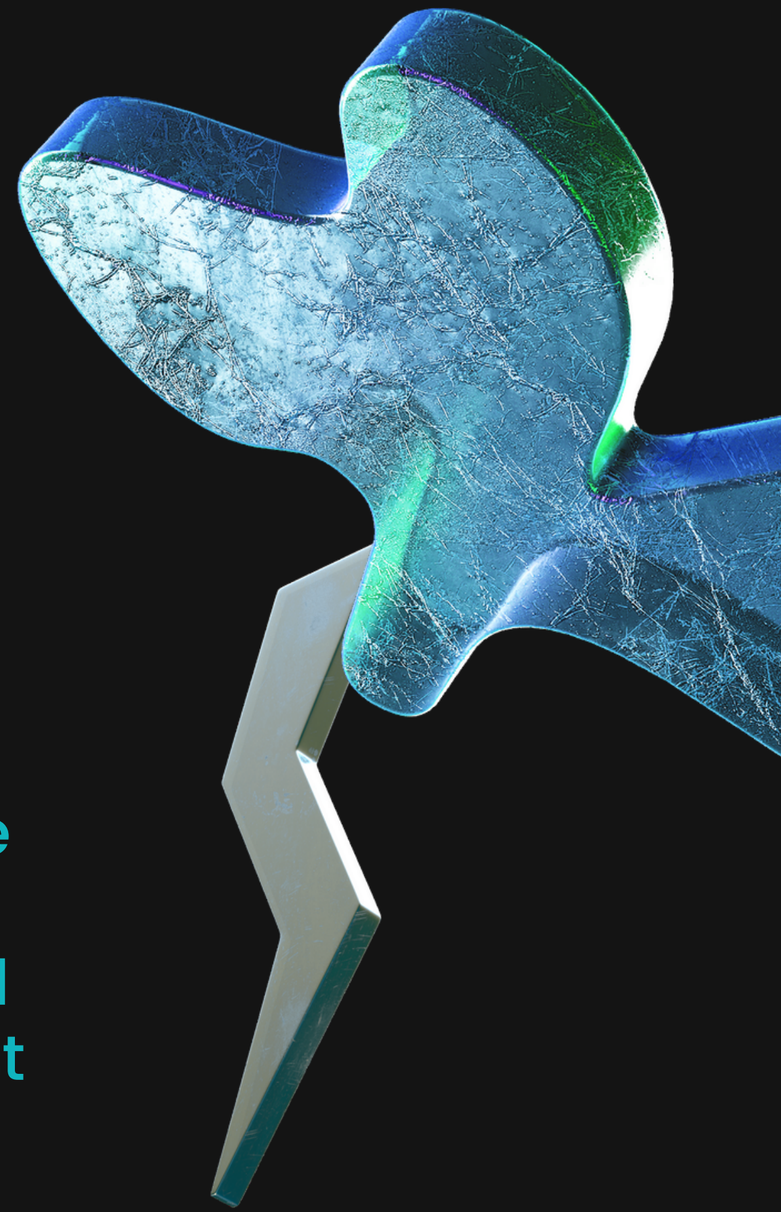
Data Science

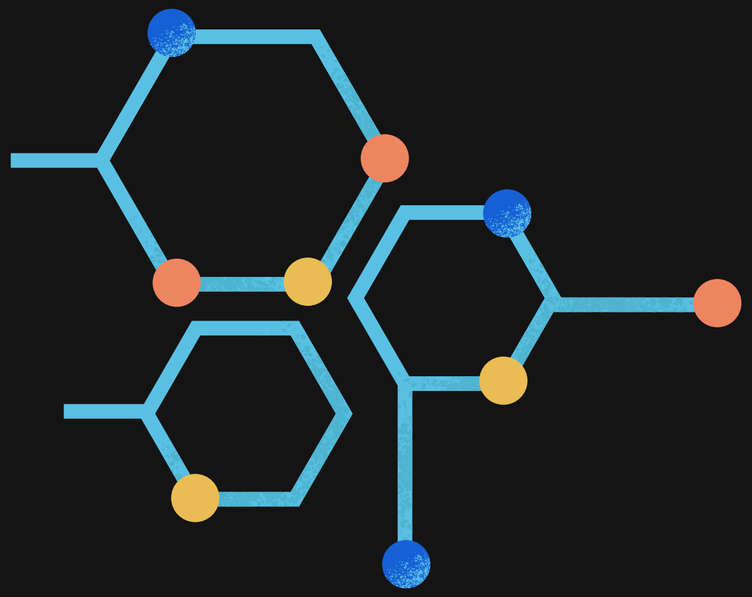


Instructor : Mayank Aggarwal
Organisation : Technophilia
Duration: 4 Weeks
Made By : Harsh Kumar
Roll No : 1900270120026

About Organisation

- Technophilia is in the Industry from past 10 years focusing and delivering skill-building training and workshops.
- Technophilia is a Skills Development Platform Providing Summer Training, Winter Camp, and workshops certified with B-Club IIT Kharagpur.
- Technophilia aims to address Students with all career related queries and giving practical exposure to trending technologies. To help the students along in their transition from students to professionals and answer any career-related questions, to their utmost satisfaction.

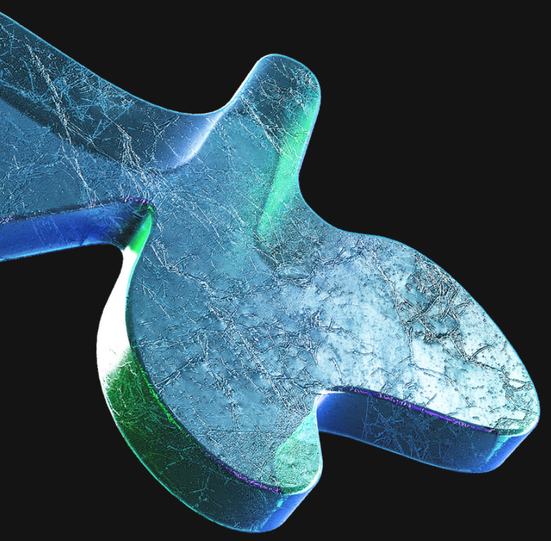




What is Data Science?

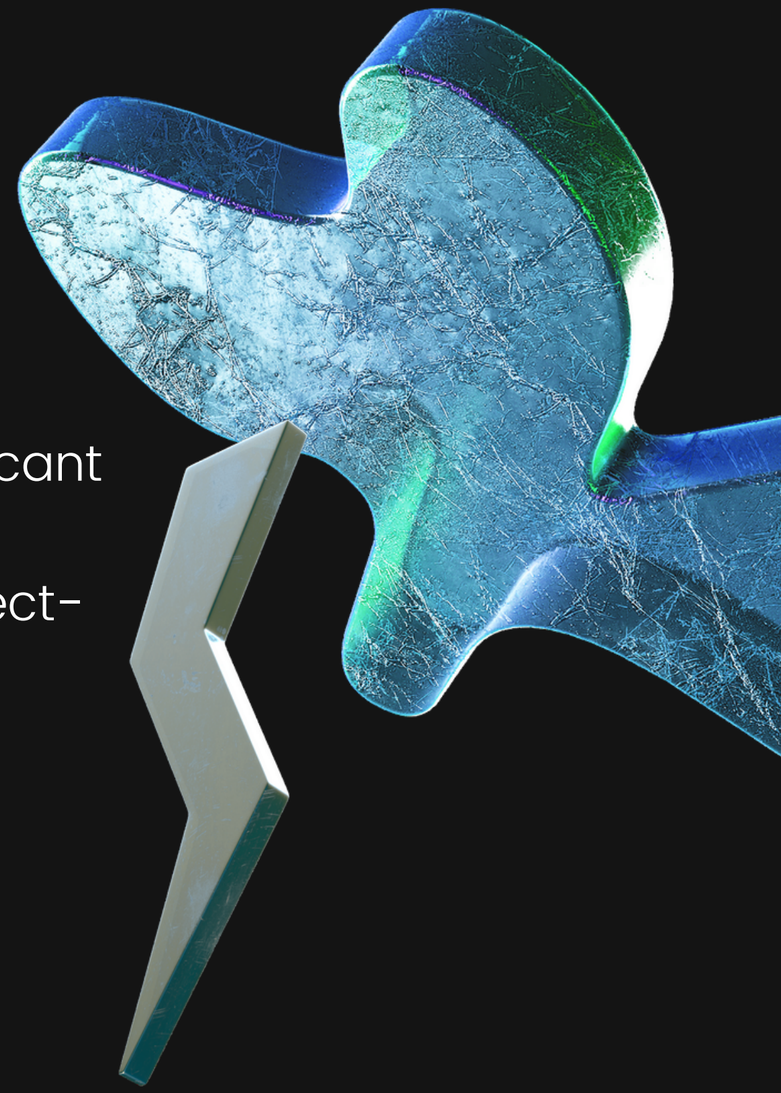
Data science, also known as data-driven science, is an interdisciplinary field of scientific methods, processes, algorithms and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining

Technologies Learned

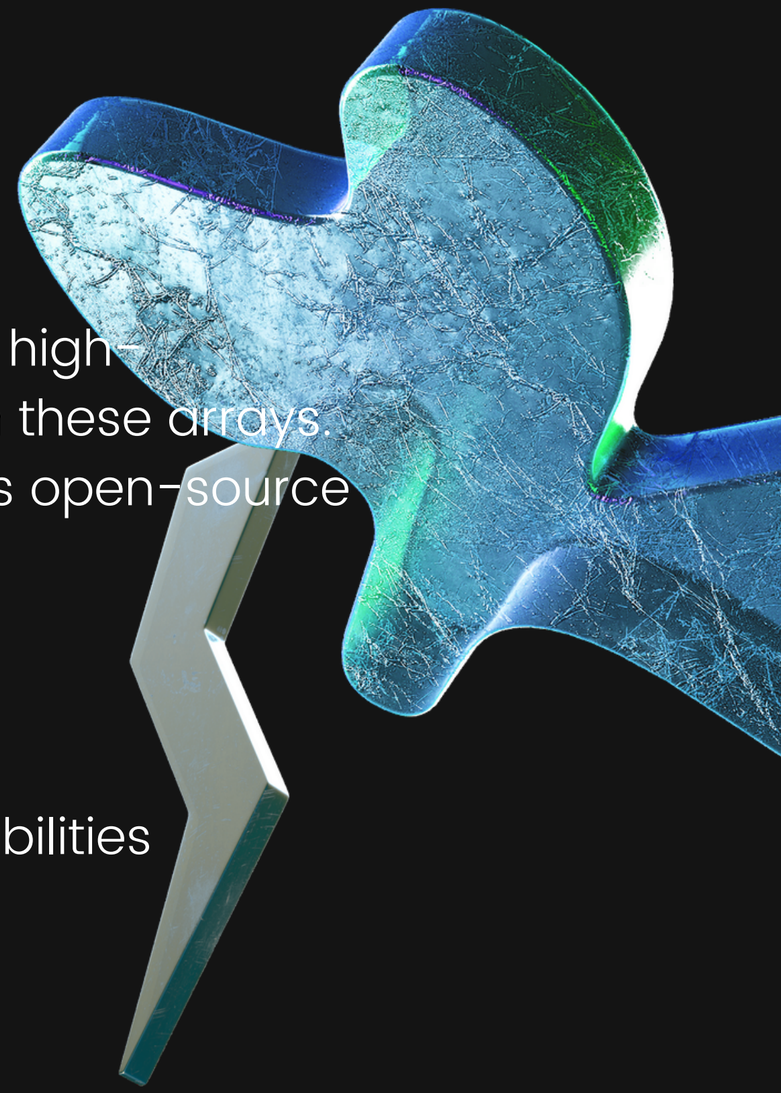


Python

Python is a high-level, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation. Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured, object-oriented and functional programming.

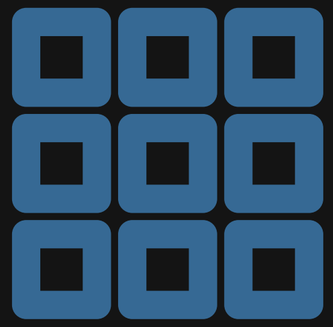


Numpy



NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays.
It is the fundamental package for scientific computing with Python. It is open-source software. It contains various features including these important ones:

- A powerful N-dimensional array object
- Sophisticated (broadcasting) functions
- Tools for integrating C/C++ and Fortran code
- Useful linear algebra, Fourier transform, and random number capabilities



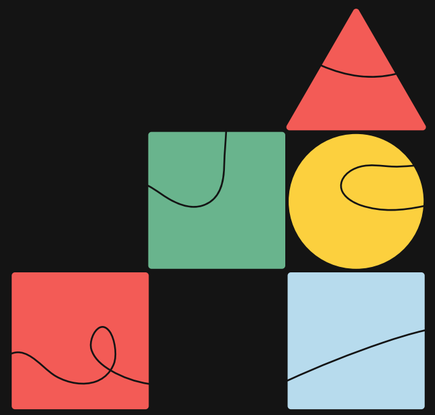
Pandas

Pandas is an open-source library that is made mainly for working with relational or labeled data both easily and intuitively. It provides various data structures and operations for manipulating numerical data and time series. This library is built on top of the NumPy library. Pandas is fast and it has high performance & productivity for users.

MatPlot Library

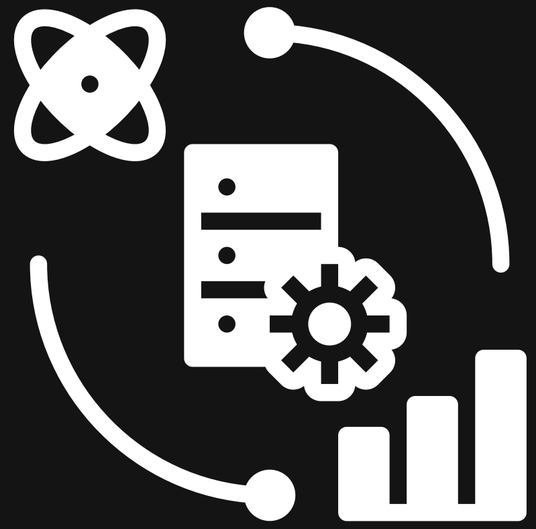
Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002.

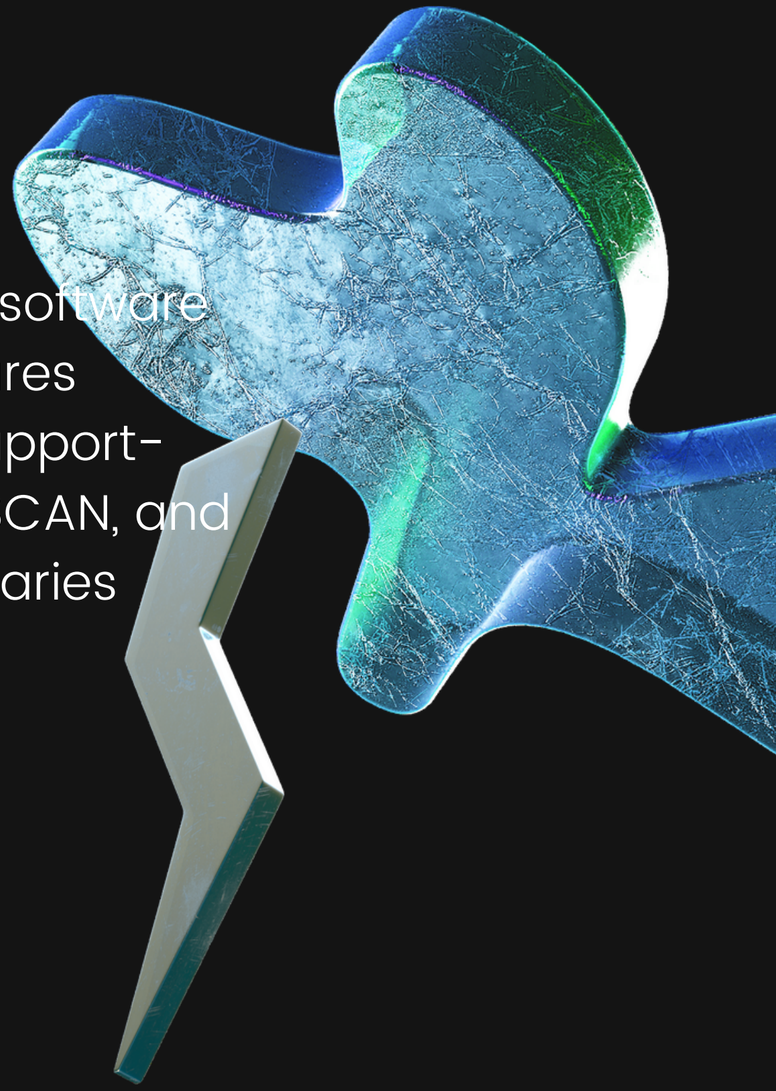
One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.



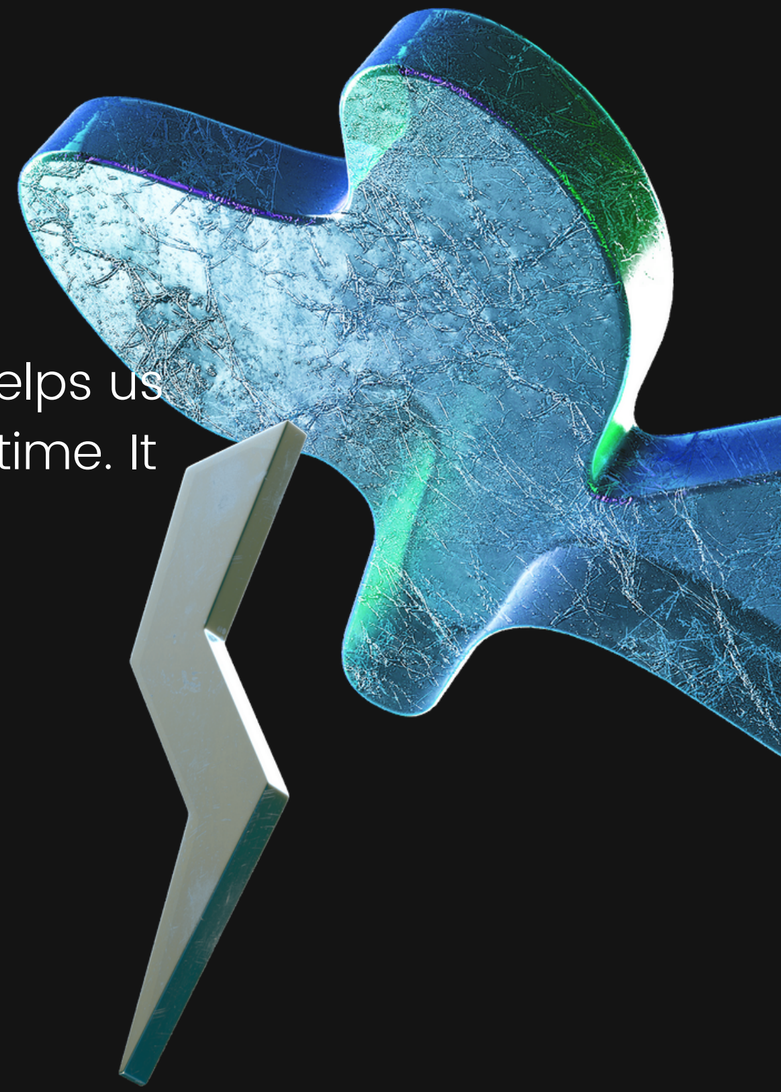
Scikit-Learn Library

Scikit-learn (formerly scikits.learn and also known as sklearn) is a free software machine learning library for the Python programming language.It features various classification, regression and clustering algorithms including support-vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

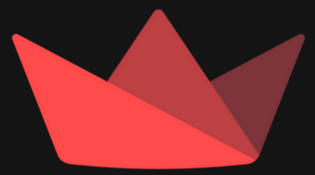




Streamlit



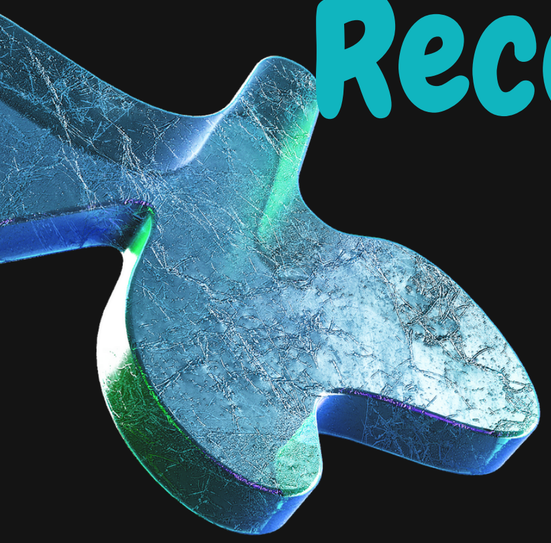
Streamlit is an open source app framework in Python language. It helps us create web apps for data science and machine learning in a short time. It is compatible with major Python libraries such as scikit-learn, Keras, PyTorch, SymPy(latex), NumPy, pandas, Matplotlib etc.



Streamlit

PROJECT

A Content Based Movie Recommendation System using Cosine Similarity



PROBLEM STATEMENT

- How to recommend movies?
- What kind of movie features can be used for the recommender system.
- How to calculate the similarity between two movies?

What is Recommendation System ?

Recommender System

Recommender systems are the systems that are designed to recommend things to the user based on many different factors. These systems predict the most likely product that the users are most likely to purchase and are of interest to. Companies like Netflix, Amazon, etc. use recommender systems to help their users to identify the correct product or movies for them.

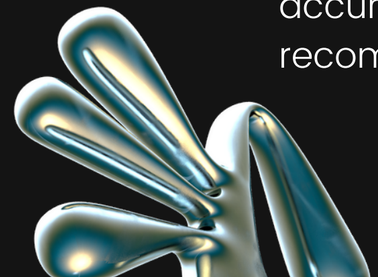
Types of Recommendation System

COLLABORATIVE FILTER SYSTEMS

The collaborative filtering method is based on gathering and analyzing data on user's behavior. This includes the user's online activities and predicting what they will like based on the similarity with other users.

CONTENT BASED SYSTEMS

Content-based filtering methods are based on the description of a product and a profile of the user's preferred choices. In this recommendation system, products are described using keywords, and a user profile is built to express the kind of item this user likes.



HYBRID RECOMMENDATION SYSTEMS

In hybrid recommendation systems, products are recommended using both content-based and collaborative filtering simultaneously to suggest a broader range of products to customers. This recommendation system is up-and-coming and is said to provide more accurate recommendations than other recommender systems.

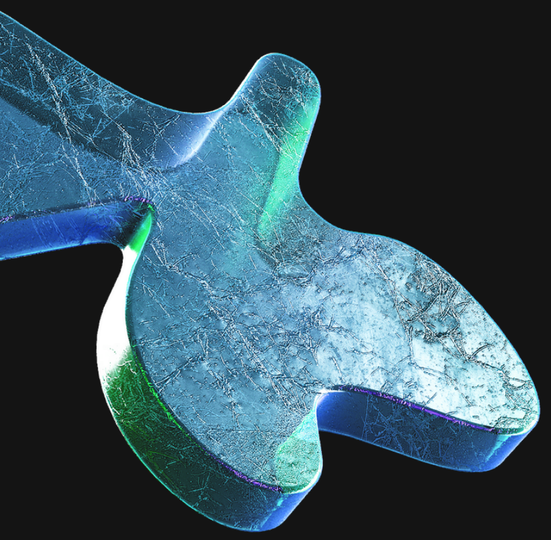
API and Dataset Used

API USED : TMDB API (For Poster Generation)
DATASET USED : TMDB_5000_MOVIES.CSV AND TMBD_5000_CREDITS

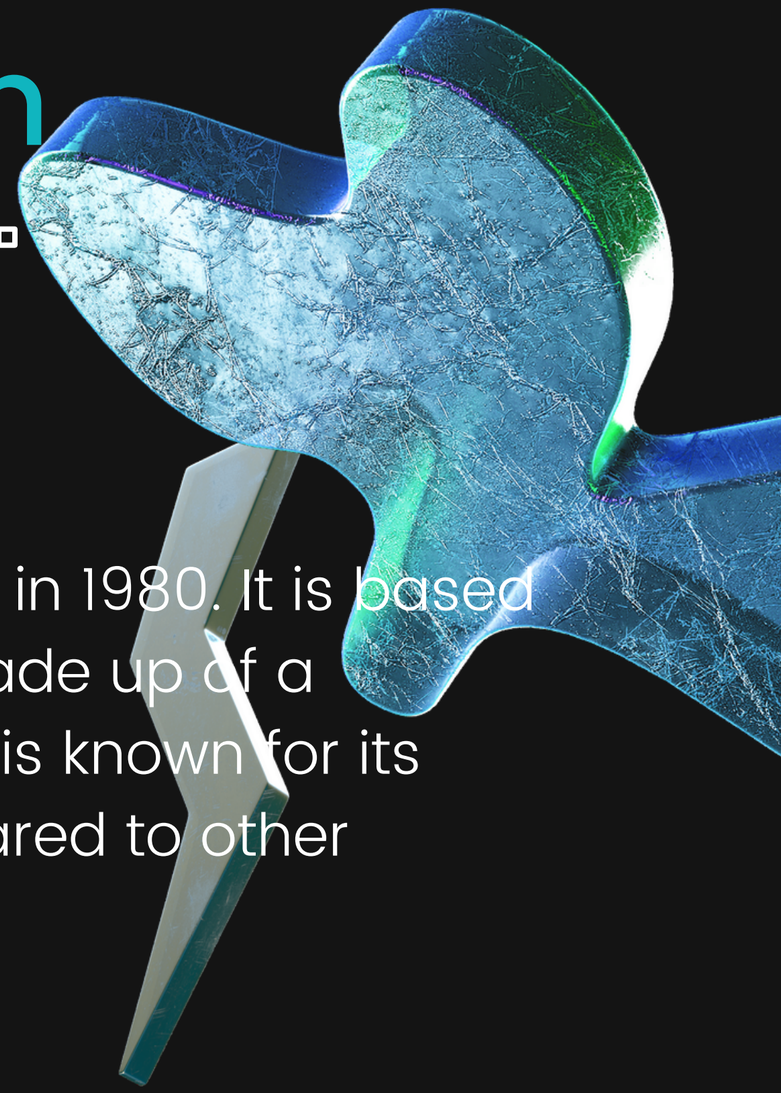
Features Used

FEATURES USED : movie_id, title, overview, genre, cast and crew

TECHNIQUES USED



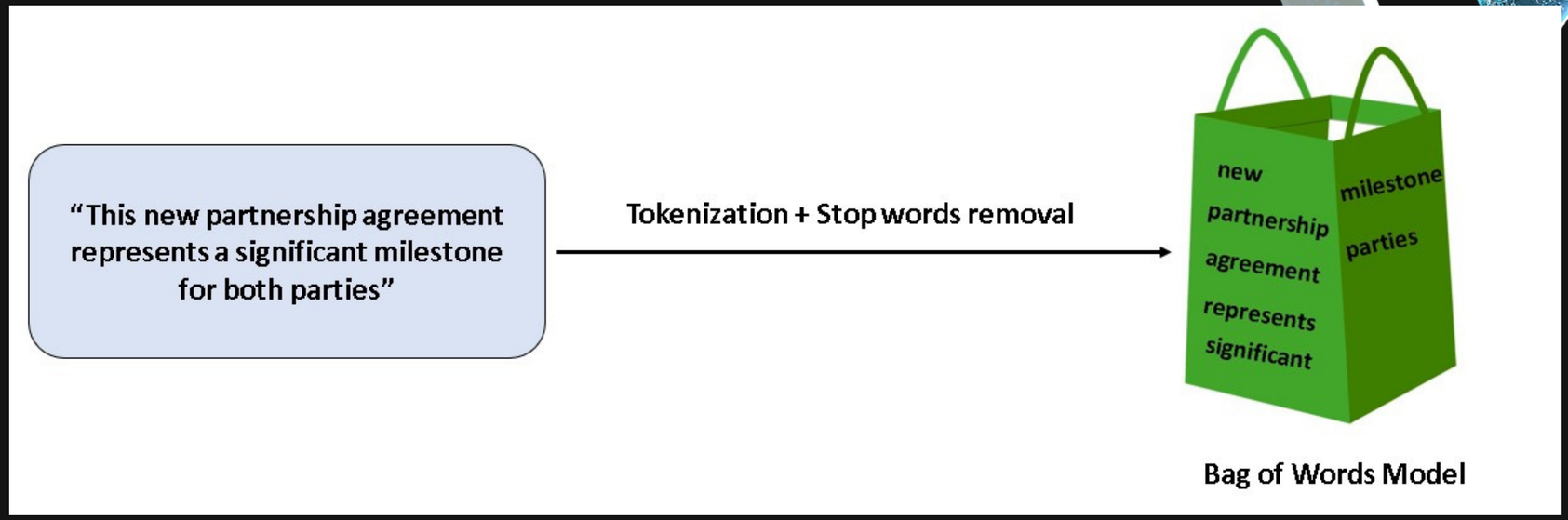
Porter’s Stemmer algorithm



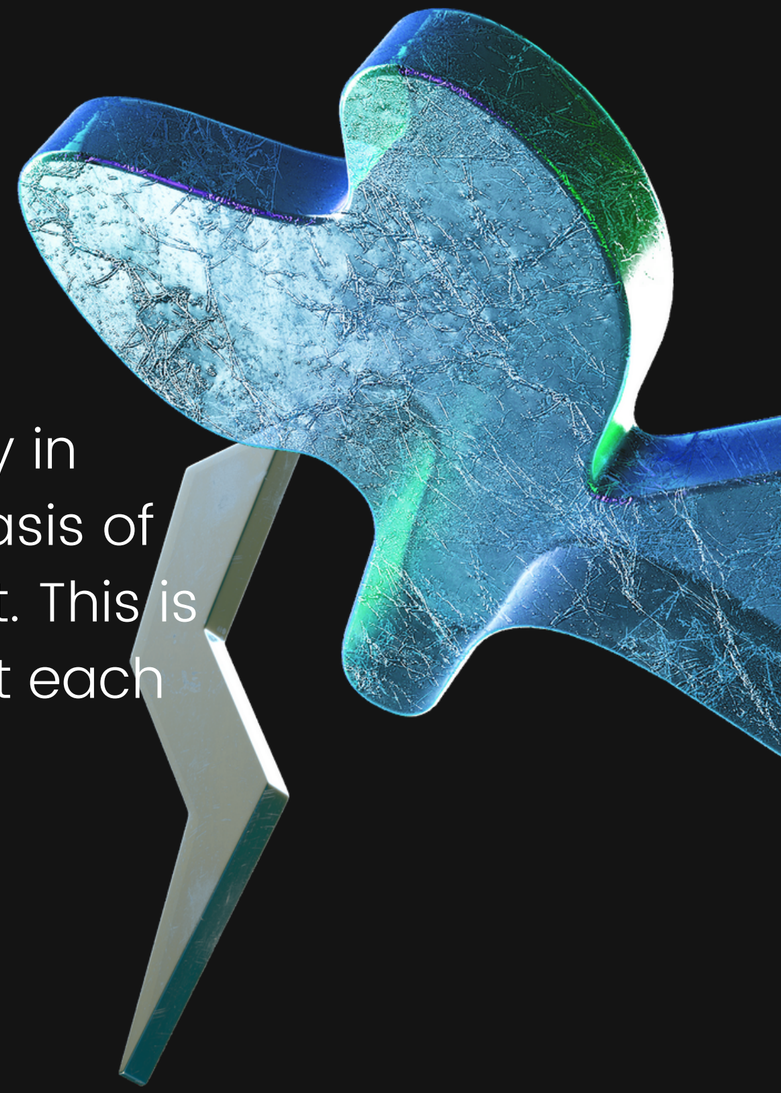
It is one of the most popular stemming methods proposed in 1980. It is based on the idea that the suffixes in the English language are made up of a combination of smaller and simpler suffixes. This stemmer is known for its speed and simplicity.It produces the best output as compared to other stemmers and it has less error rate.

BAG OF WORDS

Bag of Words model is used to preprocess the text by converting it into a bag of words, which keeps a count of the total occurrences of most frequently used words.



COUNT VECTORIZATION

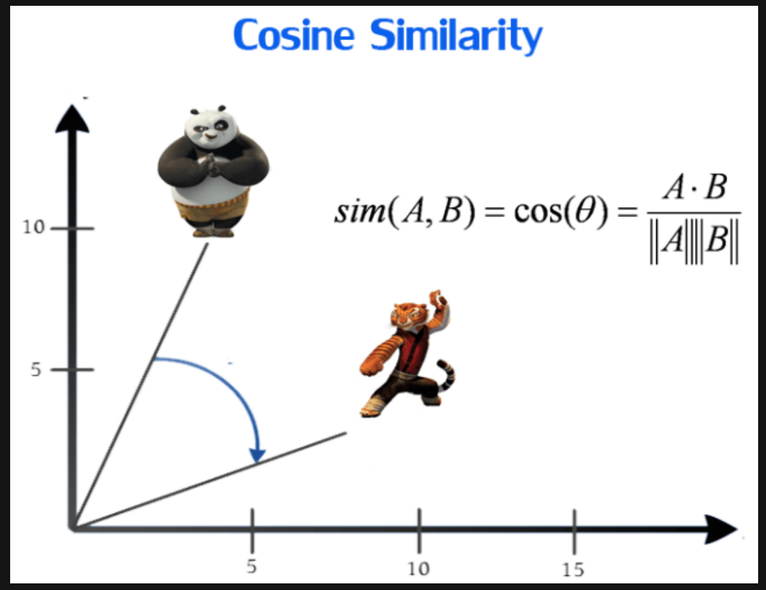


CountVectorizer is a great tool provided by the scikit-learn library in Python. It is used to transform a given text into a vector on the basis of the frequency (count) of each word that occurs in the entire text. This is helpful when we have multiple such texts, and we wish to convert each word in each text into vectors (for using in further text analysis).

COSINE SIMILARITY

Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them.

Similarity = (A.B) / (||A||.||B||) where A and B are vectors.



Project User Interface

Movie Recommendation System

Project User Interface

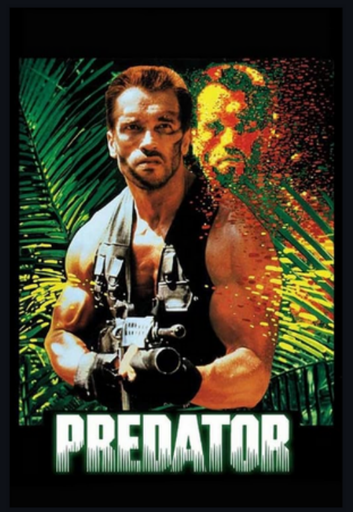
Movie Recommendation System

These movies may be of your interest

Aliens vs Predator: Requiem

Battle: Los Angeles

Predator



Titan A.E.

Lifeforce

Thank you!