IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

YOUTUBE DATA ANALYSIS

[1]Ch Suhas,[2]Akshith Reddy, [3]K Rahul

[123]IV - CSE

[1]St Peter's Engineering College,

[1]Hyderabad, India

Abstract: There is a tremendous growth and popularity of YouTube. It has the potential to touch billions of lives globally as the no. of YouTube users is growing day by day.Almost billions of videos are watched on YouTube every single day, generating a mammoth amount of data daily. Since YouTube data is generally in unstructured form, there is an increased demand to store, process and analyze such real time Big Data.These analysis will help in discovering how competitors are performing on YouTube. One can easily identify what content works best on YouTube. The primary purpose of this project is to find how real YouTube time data can be analyzed to get the latest analysis and trends.

I. INTRODUCTION

YouTube is the world's most popular online video site, with users watching 4 billion hours' worth of video each month, and uploading 72 hours' worth of video every minute (YouTube, 2013). YouTube began in February 2005 and was founded by Chad Hurley, Steve Chen, and Jawed Karim who named it "YouTube.com". Through the YouTube platform, people started to create a video-sharing website on which users could upload, share, and view videos. Since then, YouTube has gained an audience of billions of users including educators and scholars.Data Analysis and Mining are becoming indispensable parts of every major organization to find recent trends and statistics and formulate business strategies, planning and marketing.However, most of the Data generated is generally in Huge Size and comes in unstructured format.Big Data cannot be analyzed by traditional database systems and processes. To resolvethis issue, many new tools that implement Parallel Processing are being deployed in theseorganizations. As part of the Advanced Databases Project, we propose to perform Data Analysisof YouTube data. We extracted data of Video records from YouTube API andperformed Data Analysis on the data to insight into latest trends and user engagement inYouTube with respect to Categories and Year.Data Analysis and Visualization was done using Google Colaboratory. Analysis of structured data has seen tremendous success in the past. However, analysis of large-scale unstructured data in the form of video format remains a challenging area. YouTube, a Google company, has over a billion users and generates billions of views. Since YouTube data is getting created in a very huge amount and with an equally great speed, there is a huge demand to store, process and carefully study this large amount of data to make it usable.

2. LITERATURE SURVEY

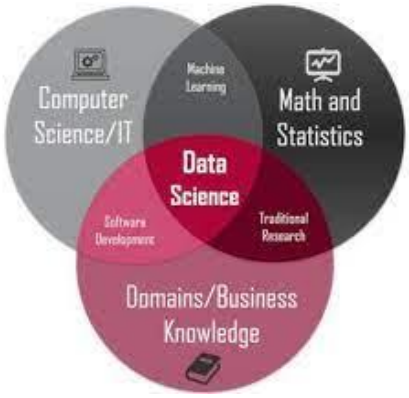
2.1 DATA SCIENCE

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains.

Data science is the domain of study that deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions. Data science uses complex machine learning algorithms to build predictive models. The data used for analysis can be from multiple sources and present in various formats.

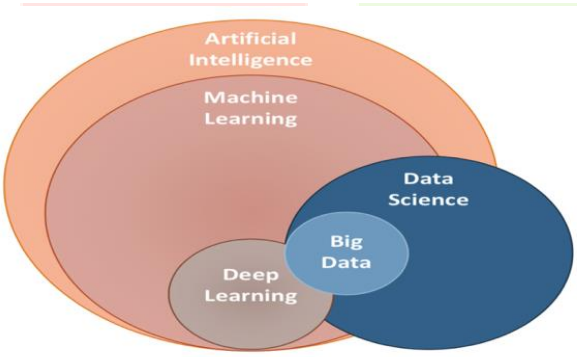
Data science is an essential part of any industry today, given the massive amounts of data that are produced. Data science is one of the most debated topics in the industry these days. Its popularity has grown over the years, and companies have started implementing data science techniques to grow their business and increase customer satisfaction.Data science or data-driven science enables better decision making, predictive analysis, and pattern discovery. It lets you:

- Find the leading cause of a problem by asking the right questions
- Perform exploratory study on the data
- Model the data using various algorithms



In practice, data science is already helping the airline industry predict disruptions in travel to alleviate the pain for both airlines and passengers. With the help of data science, airlines can optimize operations in many ways, including:

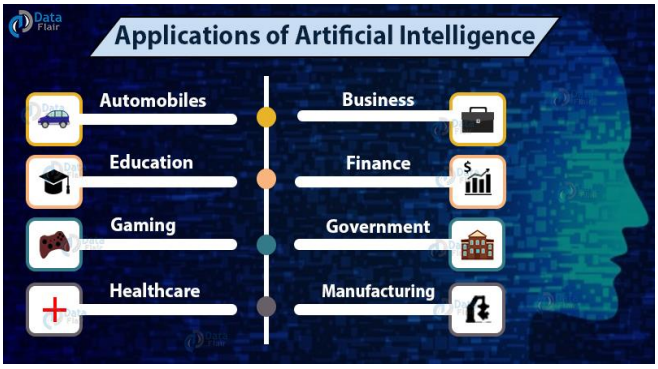
- Plan routes and decide whether to schedule direct or connecting flights
- Build predictive analytics models to forecast flight delays
- Offer personalized promotional offers based on customers booking patterns
- Decide which class of planes to purchase for better overall performance



2.2 ARTIFICIAL INTELLIGENCE

Artificial intelligence (AI) refers to the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions. The term may also be applied to any machine that exhibits traits associated with a human mind such as learning and problem-solving. The ideal characteristic of artificial intelligence is its ability to rationalize and take actions that have the best chance of achieving a specific goal.

A subset of artificial intelligence is machine learning, which refers to the concept that computer programs can automatically learn from and adapt to new data without being assisted by humans. Deep learning techniques enable this automatic learning through the absorption of huge amounts of unstructured data such as text, images, or video.



Artificial intelligence is based on the principle that human intelligence can be defined in a way that a machine can easily mimic it and execute tasks, from the most simple to those that are even more complex. The goals of artificial intelligence include mimicking human cognitive activity. Researchers and developers in the field are making surprisingly rapid strides in mimicking activities such as learning, reasoning, and perception, to the extent that these can be concretely defined.AI is continuously evolving to benefit many different industries. Machines are wired using a cross-disciplinary approach based on mathematics, computer science, linguistics, psychology, and more.

The function and popularity of Artificial Intelligence are soaring by the day. Artificial intelligence is the ability of a system or a program to think and learn from the experience. AI has significantly evolved over the past few years and has found its applications in almost every business sector

1. AI in E-Commerce

2. AI in Navigation

3. AI in Robotics

4. AI in Human Resource

5. AI in Healthcare

6. AI in Gaming

2.3 MACHINE LEARNING

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves.

The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers to learn automatically without human intervention or assistance and adjust actions accordingly.

Machine learning algorithms are used in a wide variety of applications, such as in medicine, email filtering, and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

A subset of machine learning is closely related to computational statistics, which focuses on making predictions using computers; but not all machine learning is statistical learning. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning. In its application across business problems, machine learning is also referred to as predictive analytics.

2.4 NATURAL LANGUAGE PROCESSING:

Natural Language Processing, usually shortened as NLP, is a branch of artificial intelligence that deals with the interaction between computers and humans using the natural language.

The ultimate objective of NLP is to read, decipher, understand, and make sense of the human languages in a manner that is valuable.Most NLP techniques rely on machine learning to derive meaning from human languages.

Natural Language Processing is the driving force behind the following common applications:

- Language translation applications such as Google Translate
- Word Processors such as Microsoft Word and Grammarly that employ NLP to check grammatical accuracy of texts.
- Interactive Voice Response (IVR) applications used in call centers to respond to certain users' requests.
- Personal assistant applications such as OK Google, Siri, Cortana, and Alexa.

Natural language processing helps computers communicate with humans in their own language and scales other language-related tasks. For example, NLP makes it possible for computers to read text, hear speech, interpret it, measure sentiment and determine which parts are important.

It involves the reading and understanding of spoken or written language through the medium of a computer. ... Through natural language processing, computers learn to accurately manage and apply overall linguistic meaning to text excerpts like phrases or sentences.

computer programs that can access data and use it to learn for themselves.

The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide.

Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.

Supervised learning is the type of machine learning in which machines are trained using well "labelled" training data, and on the basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output.

3.REQUIREMENTS SPECIFICATION

3.1 SOFTWARE REQUIREMENTS:

The software requirements are descriptions of features and functionalities of the target system. Requirements convey the expectations of users from the software product.

Operating System: Windows/Mac

IDE:Jupyter Notebook/Google Colaboratory

Programming Language: Python 3

Packages: Pandas ,Scikit-Learn, Seaborn, Matplotlib,Textblob,Wordcloud

3.2 HARDWARE REQUIREMENTS:

The hardware requirements are the requirements of a hardware device. Most hardware only has operating system requirements or compatibility.

System : Pentium IV 2.4GHz.

Hard Disk : 2 TB(7200 RPM)+512 GB SSD.

Floppy Drive : 1.44Mb.

Mouse : Optical Mouse

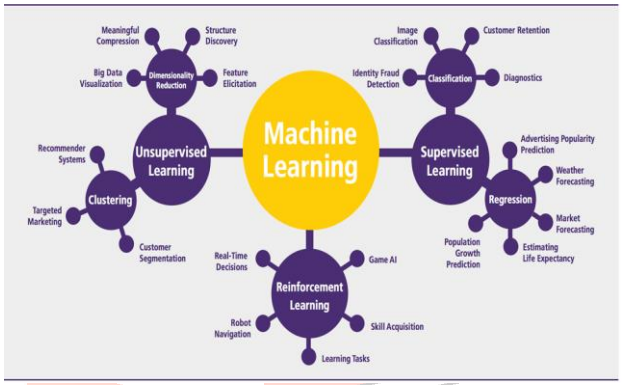
Ram : 128 GB.

4.TECHNOLOGY STACK

4.1 MACHINE LEARNING

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.

Machine learning focuses on the development of



4.2 GOOGLE COLABORATORY

Google is quite aggressive in AI research. Over many years, Google developed an AI framework called TensorFlow and a development tool called Colaboratory. Today TensorFlow is open-sourced and since 2017, Google made Colaboratory free for public use. Colaboratory is now known as Google Colab or simply Colab.

Another attractive feature that Google offers to the developers is the use of GPU. Colab supports GPU and it is totally free. The reasons for making it free for the public could be to make its software a standard in the academics for teaching machine learning and data science. It may also have a long term perspective of building a customer base for Google Cloud APIs which are sold on a per-use basis.

Irrespective of the reasons, the introduction of Colab has eased the learning and development of machine learning applications.

Colab is a free notebook environment that runs entirely in the cloud. It lets you and your team members edit documents, the way you work with Google Docs. Colab supports many popular machine learning libraries which can be easily loaded in your notebook.

What Colab Offers You?

As a programmer, you can perform the following using Google Colab.

- Write and execute code in Python
- Document your code that supports mathematical equations
- Create/Upload/Share notebooks
- Import/Save notebooks from/to Google Drive
- Import/Publish notebooks from GitHub
- Import external datasets e.g. from Kaggle
- Integrate PyTorch, TensorFlow, Keras, OpenCV

4.3 PYTHON

Python is an interpreted, high-level and general-purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant whitespace.

Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly, procedural), object-oriented and functional programming.

Python is often described as a "batteries included" language due to its comprehensive standard library.

1. Python can be used to handle big data and perform complex mathematics

2. Python can be used on a server to create web applications.

3. Python can be used alongside software to create workflows.

4. Python can connect to database systems. It can also read and modify files.

Python is a multi-paradigm programming language. Object-oriented programming and structured programming are fully supported, and many of its features support functional programming and aspect-oriented programming (including by metaprogramming and metaobjects (magic methods)). Many other paradigms are supported via extensions, including design by contract and logic programming.

Python uses dynamic typing and a combination of reference counting and a cycle-detecting garbage collector for memory management. It also features dynamic name resolution (late binding), which binds method and variable names during program execution.

Why Python?

1. Python works on different platforms (Windows, Mac, Linux, Raspberry Pi, etc).

2. Python has a simple syntax similar to the English language. Python has syntax that allows developers to write programs with fewer lines than some other programming languages.

3. Python runs on an interpreter system, meaning that code can be executed as soon as it is written. This means that prototyping can be very quick.

4. Python can be treated in a procedural way, an object-oriented way or a functional way.

5. Python is a general purpose language, used by data scientists and developers, which makes it easy to collaborate across your organization through its simple

syntax.

4.3 LIBRARIES

Pandas: Pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool,built on top of the Python programming language.

pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.

pandas is a Python package that provides fast, flexible, and expressive data structures designed to make working with structured (tabular, multidimensional, potentially heterogeneous) and time series data both easy and intuitive.

It aims to be the fundamental high-level building block for doing practical, real world data analysis in Python. Additionally, it has the broader goal of becoming the most powerful and flexible open source data analysis / manipulation tool available in any language. It is already well on its way toward this goal.

The two primary data structures of pandas, Series (1-dimensional) and DataFrame (2-dimensional), handle the vast majority of typical use cases in finance, statistics, social science, and many areas of engineering.