**Data Sci Mini project**

**Search Engine for YouTube videos**

112103059 - Anvay Joshi

112103058 - Manas Jorvekar

T4 batch Div-1

1 . Introduction:

Traditional search methods often fall short in delivering precise results, necessitating the development of a specialized search engine. This project aims to create a robust search platform for YouTube videos through keyword-based retrieval mechanisms.

1. Problem Statement:

Existing search mechanisms often rely solely on metadata such as titles and descriptions, leading to sub-optimal results. The lack of contextual understanding sometimes leads to incorrect results. Thus, we develop a tool that helps to cluster videos based on the content of its transcripts.

1. Objectives:

* Develop a search engine capable of accurately retrieving YouTube videos based on user query keywords.
* Try to use all the tools and techniques learned throughout the duration of this course.
* Improve the relevance and accuracy of search results by considering video content instead of metadata.
* Provide users with personalized and contextually relevant recommendations based on their search history and preferences.
* Implement LDA (Latent Dirichlet Allocation) using unsupervised classification of documents to find natural groups of topics.

1. Scope:

The project will focus on developing the search engine's core functionality, including keyword-based video retrieval and relevance ranking. Using concepts like normalization, standardization and calculating TF-IDF and document work frequency.

The scope excludes advanced features such as real-time updates, user authentication, and deep learning-based recommendation systems.

1. Methodology:

The project will utilize a combination of the YouTube Data API for metadata retrieval, topic modelling techniques like LDA , and information retrieval algorithms for search result ranking ( TF-IDF). Python libraries such as google-api-python-client, nltk, google-transcript-api and scikit-learn will be used for data extraction, processing, and modeling.

1. Expected Outcomes:

* A functional search engine capable of retrieving YouTube videos based on query keywords.
* Enhanced user engagement and satisfaction with personalized video recommendations.

1. Timeline:

* Day 1-2: Data collection and preprocessing.
* Day 3-4: Development of search engine algorithms and integration with YouTube Data API.
* Day 5-6: Implementation of natural language processing techniques for query understanding.
* Day 7-8: Evaluation and testing of the search engine's performance.
* Day 9-10: Fine-tuning and optimization based on user feedback.

1. Resources Required:

* Python programming environment
* YouTube Data API key
* NLTK and scikit-learn libraries
* Data processing and modeling

9. References:

* <https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2>
* https://towardsdatascience.com/nlp-topic-modeling-to-identify-clusters-ca207244d04f