# Project Proposal

1. Introduction

Ted Talks is a useful open source for people to get educational information about technology, education, psychology, etc. With so many educational videos available, finding the most appropriate video may be a challenge if people are looking for a specific topic and content.

This project aims to assist users could get the best video, which fits their requirements based on the keywords they search. I think that improving the search engine is one of the main targets they dedicated to the Ted talks website.

With the improvement of the website's integrity, I think that more and more information could be utilized on search engines. It might only use the title, speaker, and introduction to get the appropriate video for users. The transcript is helpful content for searching the suitable video.

The project would help many people, especially students. They can learn something new through a short talk, or get access to the field and figure out which they are interested in the future direction. In addition, there are also some psychological and inspirational videos, and it is also useful for people who are under heavy pressure or felt confused about their future.

The datasets that might be used:

* [TED – Ultimate Dataset | Kaggle](https://www.kaggle.com/datasets/miguelcorraljr/ted-ultimate-dataset)
* [TED Talks | Kaggle](https://www.kaggle.com/datasets/rounakbanik/ted-talks)
* [TED-talks/transcripts.csv at master · owentemple/TED-talks (github.com)](https://github.com/owentemple/TED-talks/tree/master/data)

The features that might be used:

* Title
* Published Date
* Speaker
* Content

1. Processing

Firstly, I think the information on the Title, Speaker, and Published Date is too short to get the matching video which is good enough. If the transcript is considered, it definitely could get the better result. However, it would be a huge challenge when the transcript is considered in the measure because the information is too much. It would be a huge load and time-consuming for programming to get the answer. The transcript is for presentation, so it may include many unimportant words, such like a joke for making the presentation more interesting. How to discriminate if the word is important information for the content is also a challenge.

I might use “map” and “ndcg” to measure the efficacy of the model.

* **Outline of Step:**

1. I would use the Title, Speaker, and Published Date in first place. Due to the short length of the information, it is easier to calculate the score that measures how well the videos match.
2. And then, rank the score and get the 20 to 30 videos. (It would be parameters)
3. Utilizing the transcript calculates the score to measure the fitness again and get the better one. It would lower the time-consuming and the load of the programming.
4. Utilizing the different methods to

* **Schedule:**

I would like to set the milestones as finishing the first step, which is utilizing the title, speaker, and published date to be the database for search engine.

* **Person who on the team:**

Yi-An Chu