Project Update

Yi-An Chu

1. **Introduction**

   Ted is a useful open-source tool for people to learn new knowledge from expert presentations which contain not only the basic information on that field but also the summary of their experience. There are too many videos on the website, so it is a challenge for users to search the videos which are satisfied their requirements. This project aims to help users better match ted videos, based on their demand. To make the process more efficient, I built a search engine for videos search. Given the keywords and the user's requirement, it will retrieve relevant videos title and URLs from the database.

2. **Data**

❖ **The main dataset**

The dataset is from Kaggle (TED Talks Transcripts for NLP | Kaggle), and I would like to use ted_talks_en.csv to create the system. The dataset contains 4005 documents, 15164 terms, 120743 postings, and 130774 tokens. I create some data frames, including data, transcript, speaker, and info, which might use in the system.

data	docno	title	topics	related_talks	text
transcript	docno	transcript			
speaker	docno	speaker_1	all_speakers	occupations	about_speakers
info	docno	published_date	views		

❖ **The topics dataset**

I use the title as the topics from the dataset.

❖ **The query dataset**

I use the popular news in 2022 as the query set for the system. It's because I think humans would like to search the background knowledge when they get the news for learning deeply the information. For the annotations (labeling) part of each query, there are four label scores available: -1, 0, 1, 2 . 2 is the most relevant document(with the highest relevance score), followed by 1 and 0. -1 is irrelevant document.

- Measure methods of relevant score:

1. I convert the ted descriptions corpus to vectors. Write a function that given a new query, will convert its text to a vector (using the same vectorizer as you used for the descriptions corpus), and measure the cosine similarity

score between the query and each document. And then use quartile to define the relevant score.

- ➢ cosine similarity score < Q1 : -1
- ➢ cosine similarity score < Q2 : 0
- ➢ cosine similarity score < Q3 : 1
- ➢ cosine similarity score >= Q3 : 2

2. (might use other methods to define the label)

3. Related Work

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 matching videos for users. The transcript is helpful content for searching for suitable videos. The article [1] uses transcripts to rank the best matching videos, and I think it could add other features, such as titles and topics. The article [2] researches the duration that would influence the rate of users. I think it could be used in the system. The article [3] uses K-Nearest Neighbor (KNN) method for query-dependent ranking, and I think the project could be contained different methods and compared the efficacy. The article [4] and [5] research the new algorithms, AdaRank and BoltzRank, which significantly outperforms the baseline methods of BM25, Ranking SVM, and RankBoost. Therefore, I think I also could use this method in the project and test if it also outperforms in this case.

4. Methodology

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 a 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 as a joke for making the presentation more interesting. How to discriminate if the word is important information for the content is also a challenge.

• Outline of Step:

1. I would use the Title, Speaker, and Published Date in the 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.

5. Evaluation and Results

The last part of the project is utilizing the different methods, “map”, “ndcg” and “tf-idf”, to measure the efficacy of the model.

6. Work Plan

I. I have selected the dataset and created the index for the title. In addition, the query set has also been created.

II. I would like to add the transcript information to make the rank much better.

7. References

[1] Design and implementation of an online corpus of presentation transcripts of TED Talks, Yoichiro Hasebe Doshisha University, Kyotanabe 610-0394, Japan

[2] Handling of online information by users: evidence from TED talks, M. Utku Özmen & Eray Yucel (2019)

[3] Query dependent ranking using K-nearest neighbor, Xiubo Geng, Tie-Yan Liu, Tao Qin, Andrew Arnold, Hang Li, Heung-Yeung Shum (2008)

[4] AdaRank: a boosting algorithm for information retrieval, Jun Xu, Hang Li (2007)

[5] BoltzRank: learning to maximize expected ranking gain, M. Volkovs, R. Zemel (2009)