Plan Overview

A Data Management Plan created using DMP Tool

DMP ID: https://doi.org/10.48321/D1BF4BE9E7

Title: Topic Tagging in Educational Videos Using Text-Based Search Techniques

Creator: Soongho Han - **ORCID:** 0009-0007-5443-9173

Affiliation: Iie Varsity College

Funder: The Independent Institute of Education (iie.ac.za)

Template: Digital Curation Centre

Project abstract:

This research project delves into the technical intricacies of developing robust algorithms for automatic speech-to-text transcription in educational videos. The primary objective is to employ advanced signal processing techniques and deep learning models to accurately convert spoken language in videos into textual format. These transcriptions will serve as the foundation for implementing sophisticated topic tagging and indexing mechanisms within the educational video content. By leveraging cutting-edge advancements in speech recognition and natural language processing, the project aims to revolutionize the accessibility and navigability of educational video resources through precise text-based search functionalities.

Start date: 04-02-2024

End date: 11-04-2024

Last modified: 07-09-2024

Topic Tagging in Educational Videos Using Text-Based Search Techniques

Data Collection

What data will you collect or create?

I am going to employ secondary data analysis approach. This involves the utilization of existing data sources, such as previously published research articles, datasets available from educational platforms, and publicly accessible video content metadata. This approach is selected to leverage the wealth of existing information and to conduct a comprehensive analysis without the need for primary data collection.

-YouTube-8M

-Journals, reports and other internet resource which are already existing.

I am going to see several educational videos and look for the algorithm about the tagging.

-Extracting metadata, transcripts, and user interaction data from educational video repositories.
- Metadata : Evaluating the completeness and relevance of metadata associated with educational videos.
- Transcription: Using natural language processing (NLP) techniques to analyze video transcripts and identify key topics.

-Published Research: Reviewing and synthesizing findings from previous studies on video tagging and educational content management.

How will the data be collected or created?

The data collection will involve gathering secondary data from multiple sources:

-YouTube-8M: The YouTube-8M has dataset with human-verified segment annotations. They collected human-verified labels on about 237k segments. So, I can get the large amount of label dataset.

-Online learning platforms: Such as Khan academy, Udemy, these videos span various subjects and educational levels, providing a broad basis for analysis.

-Reports, journals and other internet resources: there is a lot of resources about topic detection, tagging and labelling for videos. And even for the algorithms.

Documentation and Metadata

What documentation and metadata will accompany the data?

- Detailed Dataset Descriptions: Describe the structure and contents of your datasets, including the types of data collected (quantitative or qualitative), the nature of the variables, and any specific data formats used. This should also include any file naming conventions or identifiers used for the dataset components.
- Methodology Documentation: Document the methodology used in data collection, analysis, and processing. This includes the specific techniques and tools used for data scraping, natural language processing, and any statistical or thematic analysis methods. It’s crucial to detail the algorithms or models developed, including versioning information of any software or tools used.

- Ethical Considerations: Document any ethical considerations and approvals, including informed consent forms and how participant confidentiality is maintained. This information is crucial for ethical reproducibility.
- Citations and References: Include complete references for any sources cited in your dataset documentation. This allows future researchers to trace back to your foundational sources.

Ethics and Legal Compliance

How will you manage any ethical issues?

Regarding the ethics for this research, the secondary data collection methods will be used. Secondary data is freely available, but some data is often covered by an ‘open data license’ which means in this case, I need to gain permission from the data owner if it is not covered by a license for re-use. It must be checked what data can be used. And all data must be referenced, and the hyperlinks can be provided for some cases. All secondary data must be secured and protected from theft, damage and unintentional use, even if it is freely available. If data is file format, it will be encrypted and stored in a external hard drive for more security. The secondary data must be used for this research.

How will you manage copyright and Intellectual Property Rights (IP/IPR) issues?

I am going reference and cite for any material and data not created by me. So others can see where the data come from. And if I need to use private data, I will ask people who created it for the permission.

Storage and Backup

How will the data be stored and backed up during the research?

It can be stored on computer and external hard drive and also cloud storage.

so If my computer and hard drive are damaged, I can still get data from the cloud storage.

How will you manage access and security?

I can encrypt data, so then no one can open it without the permission. And I will store in the external hard drive and cloud storage.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

All my research documents are retained and shared forever and it will be stored in the school database.

And all data I used for this research project will be destroyed after 1 year.

Everyone and school can use my research but no one can edit my research document. And everyone needs to reference.

What is the long-term preservation plan for the dataset?

I am going to use my computer, external HDD and google drive.

Google drive provides allocated free disk storage, so there is no payment and I can preserve my document as long as it is saved on Google drive.

All data I used will be retained for 1 year.

And there is no cost for storing data because I am using my computer, hard drive and free google drive.

Data Sharing

How will you share the data?

I can send my document through email and whatsapp if someone wants and I can also post it on such as blog.

And I can also share the link from the Google drive, so then people can download. After finishing this research, it can be shared and available to use.

And I am going to make my research document in PDF format. This will be shared instead of word format.

Are any restrictions on data sharing required?

There is no restriction for my project. But when someone use my project, it must be rephrased or referenced.

For 6 months, it will be exclusive because my research project will be completed in November.

Responsibilities and Resources

Who will be responsible for data management?

I am the only one responsible. No one access for the data management.

What resources will you require to deliver your plan?

I need just computer and external hard drive for hardware.

And Microsoft word and adobe acrobat will be needed for software.

And I can use free space of Google drive.

All data can be backed up in Google drive and external hard drive.

Planned Research Outputs

Text - "Topic Tagging in Educational Videos Using Text-Based Search Techniques"

This research project delves into the technical intricacies of developing robust algorithms for automatic speech-to-text transcription in educational videos. The primary objective is to employ advanced signal processing techniques and deep learning models to accurately convert spoken language in videos into textual format. These transcriptions will serve as the foundation for implementing sophisticated topic tagging and indexing mechanisms within the educational video content. By leveraging cutting-edge advancements in speech recognition and natural language processing, the project aims to revolutionize the accessibility and navigability of educational video resources through precise text-based search functionalities.

Planned research output details

Title	Type	Anticipated release date	Initial access level	Intended repository(ies)	Anticipated file size	License	Metadata standard(s)	May contain sensitive data?	May contain PII?
Topic Tagging in Educational Videos Using Text-Bas ...	Text	2024-12-03	Open	None specified		Creative Commons Attribution 4.0 International	None specified	No	No