CS410 Course Project

1 Team Information

Team name: Torrey Pines
Team member (NetID): Wusi Chen (wusic2)
Team captain: Wusi Chen

2 Topic Information

In this project, I propose to build a search engine that can support the segment search of CS410 course video on Coursera. With this search engine, learners of this course can precisely locate the video segments to watch based on their queries. This can improve the learning experience of this course, since this search engine can save learner's time in watching unrelated course videos before they find the relevant video segments. Although Coursera has the search bar on the top of their webpages which can find the relevant video segments, that search feature will return the only the video segments which match all the query terms. For example, the search feature in Coursera will return no video segment if the query is "bag of word entropy". This unannounced restriction/limitation will deteriorate the learning experience, since the learner may not know if the terms in their queries does not happens at the same time a segment of the transcript. This topic is related to the theme of this class, since this project involves the ranking of the relevant text segments of the course video transcripts, as well as the development of the search engine.

3 Project Technical Plan

The subtitles files (WebVTT files) of CS410 course videos will be downloaded from Coursera, and used as the dataset for this project, since the video timestamps included in WebVTT files can be used to map the text segments to the video segments. The text data (transcripts) and timestamp data in the WebVTT files will be parsed by the webvtt-py Python library. The parsed data will be further processed and stored in a csv file for data retrieval.

The Python library Flask will be used as the framework for the search engine website, in which the learners can enter queries and view the list of relevant video and text segments. The Python library Metapy will be used to rank the relevance of the transcript segments and return the top ranked segments based on the learners' queries. Besides the csv file, the parsed data by webvtt-py library will be also processed and stored in a compatible file to Metapy.

4 Project Demonstration

To demonstrate this search engine can work as expected, a video demonstration will be recorded and provided in the CMT platform. In this video demonstration, a user will enter the website of this search engine, and type the query in the provided search bar. After hitting the search button, the website will return the top-relevant video segments, and each of them has the following information: Name of the lecture, the start and end time of the video segment, the transcript of the video segment, and the hyperlink to the video segment in Coursera.

The programming language for this project is Python.

5 Task Breakdown

The tasks of this project and their expect workload are listed below:

Task	Expected Workload
Parse subtitles files (WebVTT files) of CS410 course videos, and store the parsed data in a CSV file and a compatible file to Metapy.	8 hours
Build a method in Python to return top-relevant transcript segments based on the queries.	4 hours
Build a search engine website for video segment search.	8 hours