**CS410 Text Information Systems**

**Final Project Proposal**

**Team Members:**

|  |  |  |
| --- | --- | --- |
| **Name** | **Netid** |  |
| Manasa Gangaiah | [manasag3@illinois.edu](mailto:manasag3@illinois.edu) | Captain |
| Sudhir Ponnachana | [sudhirp2@illinois.edu](mailto:sudhirp2@illinois.edu) |  |

**What topic have you chosen? Why is it a problem? How does it relate to the theme and to the class?**

**Problem Statement/Background and relation to the class:**

If a speaker presents a major topic, usually it will be presented in multiple video series as chapters like in Coursera. Each week has multiple videos which are typically 10 minutes or more. For a student to listen/watch a particular video segment of their interest, they must manually review video series, identify the video and use seek bar to identify video segment. This is time consuming.

This project helps to reduce the time taken by the user to seek to the right video segment based on text query. This involves the concepts of text mining and using the rank list algorithms to identify the most related video segments by exploiting the companion text data. We plan to provide a search engine website for the user to search a topic and list top ranked search results. Then user can select the link in searched results, which will display video properly seeked to the right segment of interest. Particularly, this will be helpful for the user to prepare for the exam or refresh the topic of interest.

**Topic Selection and Project Information:**

Name of the project is “***SeekFrame***” under the main topic (2): “***Intelligent Learning Platform***”. In this project we plan to build a search engine that can seek to right video segment/location based on query words. This would allow user to type in a query and provide a ranked list of videos and when user selects the link, the video will be seeked to the right segment/location based on the query words.

**Briefly describe any datasets, algorithms or techniques you plan to use.**

**Dataset**

Data will be obtained from the Coursera. CS410 Text Information Systems videos and corresponding subtitles (English) webVTT documents will be used as the companion text data for the video.

Sample webVTT text data is provided below,

**00:00:09.625 --> 00:00:12.226 >> This lecture is about Natural Language**

WebVTT documents contains the subtitles of the video in English along with timestamp of the video segment. This data will be used for the searching the video and extract the time segment information from the data to point to video segment based on the query words.

**Algorithms/Techniques**

We will use web interface to provide the front end so that user can enter query words of their interest to search the video segments.NoSQL will be used as database to hold the information of the video segment id and corresponding timestamp. Preprocessing of the data will be done to clean up and build corpus data using the subtitles from webvtt txt file. BM25 algorithm will be used for ranking the video segments. This will be used for search engine to search query words and list the ranked list of video segments. So when the user selects the searched link, seeked video segment will be displayed.

**How will you demonstrate that your approach will work as expected? Which programming language do you plan to use?**

We will provide web interface for the user to make the query and based on the search results which will be ranked. User can select the link from the searched list to view the video what would be seeked at the right segment that’s been queried. We plan to use noSQL database, python, html, css, javascript.

**Please justify that the workload of your topic is at least 20\*N hours, N being the total number of students in your team. You may list the main tasks to be completed, and the estimated time cost for each task.**

**Workload**

Planned activities for the project,

1. Prepare the data by downloading the required videos and text. Webvtt is organized with 4 seconds break in the sentence and does not represent a proper video segment or a proper sentence. So, we need to organize it into meaningful video segments. For instance, grouping into 4 sentences per video segment. Estimate effort for this task is around 6 hours.
2. Building Database and create predefined index to map to time of the video segment. We plan to use NoSQL and the effort would be around 4 hrs.
3. Build a search engine: Create stop words. Preprocessing the data (like tokenization, removing special characters, stemming, removing stop words.), creating a required data corpus and use ranking algorithms to create a rank list. We plan to use BM25 algorithm. Coding and Testing, and effort would be around 15 hours.
4. Build web interface for the search website for Coursera videos for CS410 Text Information Systems and effort would be around 6 hours.
5. Integrate database, search engine, search website. Effort would be 2 hours.
6. Documentation and review. Effort would be around 6 hours.
7. Project power point and Video of the project. 2 hours.