CS410 Text Information Systems
Project Progress Report

Team Members:

Name	Netid	
Manasa Gangaiah	manasag3@illinois.edu	Captain
Sudhir Ponnachana	sudhirp2@illinois.edu	

Overview:

Our project is "***SeekFrame***". 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.

Which tasks are completed and pending, any challenges?

Tasks that are completed:

1. In memory hash table(database): Preprocess the web.vtt (video subtitles) files and create a corpus data. Create a framework such that all the files stored under Data/Subtitles will be properly parsed into a corpus data. We found maintaining in-memory database is good enough and faster than maintaining No SQL database.
2. Basic search engine using BM25 algorithm is implemented. Added logic to get top 5 searches.
3. Added logic to get the right seek time for each video based on the search words.
4. Built web interface for the search website for Coursera videos for CS410 Text Information Systems.
5. Integration of Front-end web interface and backend python server is done.
6. Testing is done with sample videos and subtitles for basic prototyping from CS410 Coursera videos and subtitles. End to end logic required for the video seek with search option is working.

Tasks that are pending:

1. Add all the subtitles files into the git repo and create a complete corpus data.
2. Add all videos to the google drive as git cannot maintain such huge files. TA confirmed to upload the videos to google drive.
3. Enhance the BM25 algorithm by adding stop words, stemming, try other ranking algorithms.
4. Test on the complete datasets.

5. Update all the code to the git repo.
6. Documentation (Project Report) and review.
7. Project power point and Video of the project.

Challenges:

1. Integrating the front-end web with the backend server was a challenge as we were running into issues with the server and web connection failures. This issue is resolved now.
2. Adding correct seek logic to the search videos was challenging as well as the web interface was not properly loading the videos. This issue is resolved now.