Text Information Systems
Project Progress Report

Content:

1. Name and netid:

Name: Nita Agarwal
NETID: nitaa2
TeamName: TeamNA (individual participant)

2. Project Summary:

Develop an application to link the MP descriptions to relevant TIS course lectures provided on Coursera.

Task List:

- **Task A: Collecting data (5 hours):** Gather MP assignment and lecture transcript from Coursera for the complete course duration
- **Task B: Preprocess data (8 hours):** Data preprocessing may include removing stopwords and words with less than min length, stemming etc. This step will help in creating the corpus and query data set.
- **Task C: Build Ranking logic (5 hours):** Build the logic for indexing the corpus data set and determine the appropriate metapy ranker for the application
- **Task D: Evaluation and parameter fine tuning (5 hours):** Evaluate the ranking and fine tune the parameter settings based on the human evaluation of the ranked list**.**

Stretch Goals:

- **Task E:** Build User Interface for the project. This will help the user select the MP and see the corresponding lecture video links on a UI instead of command prompt.
- **Task F:** Build Web based API driven interface for the project. This will help in accessing the application over web

3. Tasks completed:

I have been able to successfully complete **Task A** (collecting data), **Task B** (preprocess data), **Task C** (build ranking logic). I have gathered MP assignment and lecture transcript from Coursera for the complete course duration. I have been able to successfully complete the data preprocessing. It included removing stopwords and words with less than min length, stemming etc. This helped in creating the corpus and query data set. I have been able to successfully build the logic for indexing the corpus data set and determine the appropriate metapy ranker for the application. I have done the programming in python language.

4. Tasks remaining:

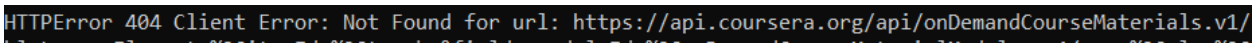
I am currently working on **Task D** (evaluation and parameter fine tuning). Evaluate the ranking and fine tune the parameter settings based on the human evaluation of the ranked list**.**

If time permits, I will work on **Task E (**Build User Interface for the project) & **Task F (**Build Web based API driven interface for the project)

5. Challenges faced:

When I started working on the project, downloading all the Coursera content took some time. I used the coursera downloader coursera-dl (https://github.com/coursera-dl/coursera-dl) to download the content.

I faced the below HTTP error (highlighter in yellow) while doing so



After referring to many online forums I came to know that there were some issues with latest source code of this coursera downloader application. I modified one of the python files of my local copy as per online forums for onDemandCourseMaterialItems.v1. Finally, after few hours of triaging the issue I was able to successfully run the code & download the course content.