

# **Text Information Systems**

## **Project Progress Report**

### **Content:**

1.	Name and netid:.....	2
2.	Project Summary:.....	2
3.	Tasks completed:.....	2
4.	Tasks remaining:.....	2
5.	Challenges faced:.....	3

## 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

```
C:\Users\nitaj>coursera-dl -ca "FmeR8K...rdCmhKecttaI7HUFHCfAC10bka_n_swZ8-nxVWncS1fmauJg_8t500g50Qje_bEPmZyKz01ZX3Hdd423svtyekhHzFDKX5YuoTm-55cEcB4tPneugZ
9HPPjwi_Mg89050Lr5SAhtyCpUX7S1HddmUJfC...d2w1qZ1r2HSf12U20Yb3TC30K5Y13V7E4eSx3-9gPSbd0A10AQ6u6ZTLlow_agrtgPgJcleRc9s7u10Zt0ptPhahG8Pmbu4y26TX9_Qab2k-_QYHbf
s8Ro0TAS24Vtrg_16h4fLlKtPP1Ae_yggw6-t...3V6r0K5InJ3ztocwJ_rEBR187j6b3PR3vr8XA790wObu-RxV6HLL-d37qqlY6e6hqlBq1-Do1-7o" cs437iot
coursera-dl version 0.11.5
Downloading class: cs437iot (1 / 1)
Parsing syllabus of on-demand course (id=0hzY2EWLEemNIQudd3U40A). This may take some time, please be patient ...
Error 404 Client Error: Not Found for url: https://api.coursera.org/api/onDemandCourseMaterials.v1/?q=slug&slug=cs437iot&includes=moduleIds%2ClessonIds%2CpassableItemGroups%2CpassableItemGroupChoices%2CpassableLessonElements%2CitemIds%2Ctracks&fields=moduleIds%2ConDemandCourseMaterialModules.v1(name%2Cslug%2Cdescription%2CtimeCommitment%2ClessonIds%2Coptional)%2ConDemandCourseMaterialLessons.v1(name%2Cslug%2CtimeCommitment%2CelementIds%2Coptional%2CtrackId)%2ConDemandCourseMaterialPassableItemGroups.v1(requiredPassedCount%2CpassableItemGroupChoiceIds%2CtrackId)%2ConDemandCourseMaterialPassableItemGroupChoices.v1(name%2Cdescription%2CitemIds)%2ConDemandCourseMaterialPassableLessonElements.v1(gradingWeight)%2ConDemandCourseMaterialItems.v1(name%2Cslug%2CtimeCommitment%2Ccontent%2CisLocked%2ClockableByItem%2CitemLockedReasonCode%2CtrackId)%2ConDemandCourseMaterialTracks.v1(passablesCount)&showLockedItems=true getting page https://api.coursera.org/api/onDemandCourseMaterials.v1/?q=slug&slug=cs437iot&includes=moduleIds%2ClessonIds%2CpassableItemGroups%2CpassableItemGroupChoices%2CpassableLessonElements.v1(name%2Cslug%2CtimeCommitment%2CelementIds%2Coptional%2CtrackId)%2ConDemandCourseMaterialPassableItemGroups.v1(name%2Cslug%2Cdescription%2CtimeCommitment%2ClessonIds%2Coptional)%2ConDemandCourseMaterialLessons.v1(name%2Cslug%2CtimeCommitment%2CelementIds%2CtrackId)%2ConDemandCourseMaterialPassableItemGroups.v1(requiredPassedCount%2CpassableItemGroupChoiceIds%2CtrackId)%2ConDemandCourseMaterialPassableItemGroupChoices.v1(name%2Cdescription%2CitemIds)%2ConDemandCourseMaterialPassableLessonElements.v1(gradingWeight)%2ConDemandCourseMaterialItems.v1(name%2Cslug%2CtimeCommitment%2Ccontent%2CisLocked%2ClockableByItem%2CitemLockedReasonCode%2CtrackId)%2ConDemandCourseMaterialTracks.v1(passablesCount)&showLockedItems=true
The server replied: <html>
  <head>
    <title>Coursera - API Route Does Not Exist</title>
  </head>
  <body style="background-color:#e4e4e4">
    <div style="position:absolute; top:0; bottom:0; left:0; right:0; margin:auto; height:200px; width: 600px">
      <div style="text-align:center">
        
      </div>
      <div style="text-align:center; font-family:Helvetica, Arial, sans-serif; font-weight:100; color: #555">
        API Route Does Not Exist
      </div>
      <div style="text-align:center; font-family:Helvetica, Arial, sans-serif; font-weight:300; font-size:13pt; color: #555">
        Edge does not know about this API route. <br>
        Check whether this route is exposed in the routing table.
      </div>
    </div>
  </body>
</html>
HTTPError 404 Client Error: Not Found for url: https://api.coursera.org/api/onDemandCourseMaterials.v1/
```

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.