

CS410 Progress Report

Team: Mission Lucy

Captain: Sam Song(samsong2)

Azim Keshwani(azimmk2)

Kavithaa Suresh Kumar(ks64)

Uttam Roy(uroy)

1) Which tasks have been completed?

- A web scraper to scrape for classes from Coursera has been implemented.
 - The web scraper can download the time stamped transcript and video link from each lecture given the course name.
 - The web scraper uses Selenium to get and load the page. BeautifulSoup is then used to scrape the pages.

2) Which tasks are pending?

- Building the web interface to perform the search
- Data preprocessing- removal of stop words, removal of extra characters, lowercase conversion,stemming
- Finding similarity between the query and the segments to retrieve the results.
- Ranking the list of relevant results

3) Are you facing any challenges?

- Due to the fact that Coursera is a dynamically loaded webpage, there is a need to wait for the javascript to fully load the page before parsing. This results in around a 3-5 second delay when fetching for each page. As a result, scraping for an entire course takes a long time.
- Scraping for a course from Coursera requires an user that is already enrolled in the course to provide authentication information.
- Still deciding on approach to build out the web interface (packages needed, design, etc)
- Decision on which python library (whoosh or scikit)to use for searching remains a challenge.