Project Final Report

Project Title: Online Learning Data Set
Theme: Theme 4 - Data Set Creation

1. **Team Member/Captain**: Randi Weston (NetID: rweston2)

2. **Completed Tasks**:

- Research (2 hr): Remaining research needed for Coursera pages and remaining half of research needed for the rest of the EdX pages.
- Development (10 hrs): Write web scrapers for each platform.
- Testing (6 hr): Test web scrapers on a small number of courses.
- Revision (3 hrs): Make any necessary revisions to scrapers to reduce data cleaning.
- Web Scraping and Data Cleaning (10 hrs - This time isn't reflected in the total because most of it was web scraping and that didn't require my involvement): Scrape all courses and clean data.
- Documentation (1 hrs): Document data set.
- Project Administration (2 hrs): Presentation and reports.
- Total: 24 hours

3. **Challenges Faced**:

- Inconsistent URLs and different course detail page layouts
  - Coursera and Edx have different url schemes and detail page layouts depending on the course type (i.e. coure, mini-bachelors, masters, etc.)
- Generic class names
  - Coursera, Edx, and Udacity have specific classes for specific elements on a page, however FutureLearn does not. Generic class names make it harder to find the elements you're looking for with Selenium
- Differing information
  - FutureLearn doesn't contain succinct skills and prerequisites lists like the other online learning sites do, so I had to make do with scraping information from different sections of the FutureLearn details pages that contained data most like the skills and prerequisites data I was looking for.
- Troubleshooting
  - Selenium doesn't generate many error messages when it fails and the error messages it does generate are not very specific.

4. **Outcome**: I was able to successfully scrape 4,583 courses from the four learning platforms which is fairly close to my estimated number of 5,000 courses.

5. **Future Improvements**:

- Include more course types (i.e. mini-bachelors, bachelors, masters, etc.)
- Add a set of default queries and query relevancy calculations

- Add more browser/platform support