Project Progress Report

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

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

2. **Progress Made Thus Far**:
   - Research (2 hrs): This task was initially included in my development estimation, but based on the work I've done so far I should have broken it out into its own task. I've been looking through each of the course catalogs I'm using (Coursera, EdX, Udacity, FutureLearn) and documenting the HTML/CSS information needed in order to access the information I intend to scrape. I've completed this documentation for Udacity and FutureLearn, and am about halfway done with EdX.
   - Development (2 hrs): I've written pseudo code for two of my scrapers based on the research I've done so far. This will make development much faster since I won't have to look up the HTML/CSS structure of each page as I'm coding. I've also created a class for the class information I will be scraping that includes methods for writing the class information to a CSV and writing information to the console.
   - Documentation (20 min): I've been working on creating some project documentation/instructions in the instructions.txt file of my project as I go so I don't forget to include any installations/imports future users may need in order to run my code.
   - Project Administration (1 hr): Created progress report

3. **Remaining Tasks**:
   - Research (1 hr): Remaining research needed for Coursera pages and remaining half of research needed for the rest of the EdX pages.
   - Development (6 hrs): Write web scrapers for each platform.
   - Testing (1 hr): Test web scrapers on a small number of courses.
   - Revision (2 hrs): Make any necessary revisions to scrapers to reduce data cleaning.
   - Web Scraping and Data Cleaning (4 hrs): Scrape all courses and clean data.
   - Documentation (1.5 hrs): Document data set.
   - Project Administration (2 hrs): Presentation.

4. **Issues/Challenges Being Faced**: The biggest challenge I've faced so far is that FutureLearn doesn't tend to use ids or unique classes for several of the sections on the course details pages. So, for a couple pieces of information, I've had to resort to looking for specific text content that will put me in the right spot to scrape the information I need to scrape. Additionally, I didn't realize it before I started looking into the HTML/CSS structure of each site, but FutureLearn doesn't have distinct lists for "skills" or "prerequisites". So, I have to take information from two sections ("What will you achieve?" and "Who is this course for?") and let that information stand in for the skills and prerequisites properties for FutureLearn classes. I've also never used Selenium before, so that has been a learning experience.