**Installation/Usage Instructions**

1. The following files must be available in the current working directory:
   1. coursicle.py
   2. employment\_stats.py
   3. etl.py
   4. heinz\_courses.py
   5. course\_similarity.py
   6. docsim.py
   7. scourse.py
   8. handshake.csv (*downloaded directly from Handshake website, Career Services Resources section*) – DATA SOURCE 1
   9. Smartevals-2019-ALL.csv (*downloaded directly from SmartEvals website, export to csv option*) – DATA SOURCE 2

The below files are generated by the python code on the fly, and should be placed in the project directory if they don’t need to be scraped fresh.

1. coursicle\_data\_dump.txt – DATA SOURCE 3
2. heinz\_courses.txt – DATA SOURCE 4
3. There are two dependencies outside of Anaconda that need to be installed. These are:
   1. Tabula (used for scraping PDFs)   
      Run the following command on the terminal: **pip install tabula**
   2. Gensim (used for semantic search)  
      Run the following command on the terminal: **pip install genism**
4. From the terminal, the application can be launched in two ways. One where it scrapes the needed web data afresh, and the other where it uses older scraped data (described in Step 1).   
   For running the script with older data, use the command:

**python scourse.py -performETL False**  
For running the script and also performing the web scraping, use the command:  
**python scourse.py -performETL True**

1. Once the script runs, it will take 4-5 minutes for generating a corpus to be used by the document similarity checking algorithm for semantic search. A progress will be shown for some of this period.
2. Once the corpus is generated, a welcome text will be printed on the console. The user can now provide an input of the job roles they want to target. Press the enter key when you are done entering the job role. For example: **Data Scientist**
3. Next, the user can type in a comma separated list of skills the user wants to target. For example: **Statistics, Machine Learning**
4. Once you press enter, the algorithm looks through the data set and performs a semantic search for the query terms entered. The top results (courses that match your search criteria the most) are returned. Data from SmartEvals is also taken into consideration before recommending a course, and historically highly rated courses are more likely to show up in the query results.
5. The program also displays a list of historically recorded career outcomes for the recommended courses. These have been provided by the Heinz Career Services office, and will give you an idea of where people who have taken this courses have ended up.