**Project 5.21 University Ranking**

In order to research for exchange information, students always need to search for the university ranking from multiple sources. I propose to make a web crawler to crawl top 1,000 universities from each sources.

**Suggested website:**

**Main**

1. QS ranking

* <https://www.topuniversities.com/>

1. US News Ranking

* <https://www.usnews.com/education/best-global-universities/rankings>

**Bonus**

1. The Complete university guide

* <https://www.thecompleteuniversityguide.co.uk/league-tables/rankings>

1. Web metrics

* <http://www.webometrics.info/en/world>

1. Times World University Ranking

* <https://www.timeshighereducation.com/world-university-rankings/2017/world-ranking#!/page/0/length/25/sort_by/rank/sort_order/asc/cols/stats>

1. Shanghai Ranking

* <http://www.shanghairanking.com/ARWU2017.html>

1. Other websites (1500 – 2000 points, to be determined by project tutor)

**Required – Data manipulation**

1. Integrate all the data into one file
2. Data cleansing
   1. Organize all information from different website into unified format
   * University Name
   * Ranking
   * Country
   * Source
   1. The following fields may not appear in all data source
   * Assessment criteria
   * University website
   * Focus
   * Year of foundation
   * No of students
   * Private/public university

Requirement: the crawled data for each website should be cleaned and submitted in csv format. Moreover, a combined data file is expected. Otherwise points reduction may occur. Furthermore, a well-commented code for each website is needed to get the points.

Estimated completion time: 35 hours + 25 hours each for bonus.

Suggestive points:

Deadline:

Bonus points:

1. Work completion one day before the deadline will generate an extra 20% bonus.
2. Powerpoint slides for each code will give 50 points/effective page.

Penalty:

1. Late delivery/missing deadline will incur -10% of finally determined points of the whole project per missing day.
2. Serious mistakes that made the submitted results not useful will incur -120% of the whole project.
3. No mark penalty if the website changes their layout after submission of the code.