**Project 5.18 Hong Kong Tutoring Websites**

In order to cover different expenses spent in the university, most of the university students earn their income by working as a private tutor for secondary school students. We mainly find the students through various agency websites. However, due to business competition, the agency website will not share the information over each other, we have to visit each website one by one to find the suitable student, which is extremely troublesome. Therefore, my suggestion is to build a web crawler to obtain the essential information in different tuition agency website and output all of them into one single spreadsheet.

**Suggested website:**

**Main**

1. Kits Tutor

* [http://www.kits-tutor.com/2015/index.php#](http://www.kits-tutor.com/2015/index.php)

1. TutorOK

* <http://www.tutorok.com/>

1. Sky Tutor

* <http://www.sky-tutor.com/>

1. Uni Tutor

* <http://uni-tutor.com/index.php?action=view_find>

1. Kitkentutor

* <http://www.kitkentutor.com/best%20tutor/tutor%20square.asp>

1. HKTutor

* <http://www.hktutor.hk/?gclid=CjwKCAiAxuTQBRBmEiwAAkFF1kAWiEIbmxHReCKJnei_QMT4L0HqmByVIIcxVTlT_5tM2Jh2GRneaxoCc6gQAvD_BwE>

**Bonus**

1. Professional Tutor

* <http://www.professionaltutor.hk/%E8%A3%9C%E7%BF%92%E5%80%8B%E6%A1%88>

1. HKTA

* <https://www.hkta.edu.hk/?gclid=CjwKCAiAxuTQBRBmEiwAAkFF1sV6eD6F1i_u9D_Ab5fq6yfOBYNV88l8jgIfdg92nJ40K0ltquJaLxoCx6kQAvD_BwE>

1. Cedars – Private Tuition

* <http://www.cedars.hku.hk/careers/home>

1. Tutor Circle

* <http://www.tutorcircle.hk/case-list.php>

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

**Main – Data manipulation**

1. Integrate all the data into one file
2. Data cleansing
   1. Drop duplicated cases
   2. Organize all information from different website into unified format
   * By subject
   * By district
   * By students’ year of study
   * By date
   * Etc

Requirement: submit a code for each website (6+5 codes) and an additional code to combine the data collected from each website. The contractor also needs to submit the csv files of each website’s data and a combined dataset.

Estimated completion time: 90 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 penalty if the website changes their layout after submission of the code.