# *COM5507 Social Media Data Acquisition and Processing*
# Week 8. More topics in web data collection

Lectured by: Dr. Xinzhi ZHANG

Research Assistant Professor, Department of Journalism

Hong Kong Baptist University

31 Oct 2018 @ CityU M4003

# Agenda

- Structuring the codes
  - Using self-defined functions
- Advanced automated data collection
  - The case of Selenium
  - The case of xpath

# Structuring the codes

- Rather than asking "What exists," try to ask: "**What do I need?”**
- Then find ways to seek the information that you need from there (Ryan Mitchell, 2018)
- Issues to be considered:
  - Harvesting multiple pages
  - Using functions

# Check-list before web data collection *

- Is this information helpful in my projects?
- Is it something “nice to have” to is it something “have to have”?
- How “variable” the information is?
  - is it a constant or it is subject to change?
- Is this data sparse or dense?
- How large is the data?
  - how many pages in total? how many items per page?...

# Advanced automated data collection

- Selenium
- Selenium Python bindings provides a simple API to write functional/acceptance tests using Selenium WebDriver.
- Through Selenium Python API you can access all functionalities of Selenium WebDriver in an intuitive way.

https://selenium-python.readthedocs.io/installation.html

# Selenium

- In-class demonstration.
- Download Python bindings for Selenium
  - pip install selenium
- Download the driver
  - depending on which browser you prefer to use
- Windows users may pay attention to section 1.4

# Case demonstration

- Workout 1. Load the page and perform a search
- Workout 2. Load the site and navigate around the website to locate the proper information
- Workout 3. *“Bags cure all”*