Course Code
Web Scraping and Text Mining
Module 1, 2020-2021

Course Information

Instructor: Yangjuan Hu
Office: PHBS Building, Room
Phone: 86-186-1712-5001
Email: huyangjuan@phbs.pku.edu.cn
Office Hour: Mon & Thur 14:00-15:00

Teaching Assistant:
Phone:
Email:

Classes:
Lectures: Mon & Thur 10:30-12:20
Venue: PHBS Building, Room

Course Website:
If any.

1. Course Description

1.1 Context

Course overview: We are living in a data saturated world. Modern web is full of valuable data about people's behaviour. This course aims to help students learn Python programming language so that they can use it to collect data from the web, clean the data, and make use of the data to solve business problems or to do social science research.

Prerequisites: No prior knowledge is required.

1.2 Textbooks and Reading Materials

Mitchell, R. (2015). *Web Scraping with Python*. O'Reilly Media.
Bird, S., Klein, E. & Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media.

2. Learning Outcomes

2.1 Intended Learning Outcomes

Learning Goals	**Objectives**	**Assessment (YES with details or NO)**
1. Our graduates will be effective communicators.	1.1. Our students will produce quality business and research-oriented documents.	YES
	1.2. Students are able to professionally present their ideas and also logically explain	YES

	and defend their argument.	
2. Our graduates will be skilled in team work and leadership.	2.1. Students will be able to lead and participate in group for projects, discussion, and presentation.	YES
	2.2. Students will be able to apply leadership theories and related skills.	YES
3. Our graduates will be trained in ethics.	3.1. In a case setting, students will use appropriate techniques to analyze business problems and identify the ethical aspects, provide a solution and defend it.	YES
	3.2. Our students will practice ethics in the duration of the program.	YES
4. Our graduates will have a global perspective.	4.1. Students will have an international exposure.	YES
5. Our graduates will be skilled in problem-solving and critical thinking.	5.1. Our students will have a good understanding of fundamental theories in their fields.	YES
	5.2. Our students will be prepared to face problems in various business settings and find solutions.	YES
	5.3. Our students will demonstrate competency in critical thinking.	YES

2.2 Course specific objectives

This course aims to teach students how to use Python to:

- Parse complicated HTML pages
- Crawl through APIs
- Crawl through forms and logins
- Learn methods to store data they scrape
- Clean and normalize badly formatted data
- Process natural languages (including English and Chinese)
- Visualize data

2.3 Assessment/Grading Details

2.4 Academic Honesty and Plagiarism

It is important for a student’s effort and credit to be recognized through class assessment. Credits earned for a student work due to efforts done by others are clearly unfair. Deliberate dishonesty is considered academic misconducts, which include plagiarism; cheating on assignments or examinations; engaging in unauthorized collaboration on academic work; taking, acquiring, or using test materials without faculty permission; submitting false or incomplete records of academic achievement; acting alone or in cooperation with another to falsify records or to obtain dishonestly grades, honors, awards, or professional endorsement; or altering, forging, or misusing a University academic record; or fabricating or falsifying of data, research procedures, or data analysis.

All assessments are subject to academic misconduct check. Misconduct check may include reproducing the assessment, providing a copy to another member of faculty, and/or communicate a copy of this assignment to the PHBS Discipline Committee. A suspected plagiarized document/assignment submitted to a plagiarism checking service may be kept in its database for future reference purpose.

Where violation is suspected, penalties will be implemented. The penalties for academic misconduct may include: deduction of honour points, a mark of zero on the assessment, a fail grade for the whole course, and reference of the matter to the Peking University Registrar.

For more information of plagiarism, please refer to *PHBS Student Handbook*.

3. Topics, Teaching and Assessment Schedule

Session	Topic
1	Introduction to Python I: Data Types in Python
2	Introduction to Python II: Method, Function and Statement
3	Web scraping with Beautiful Soup I
4	Web scraping with Beautiful Soup II
5	Web scraping with Selenium I
6	Web scraping with Selenium II
7	Case study: Getting data from Weibo
8	Case study: Getting data from Zhihu
9	Data cleaning with Pandas I
10	Data cleaning with Pandas II
11	Data analysis with Pandas
12	Data visualization with Python
13	Case study: Movies and genres
14	NLP with Python I: NLTK
15	NLP with Python II: Text Classification
16	NLP with Python III: Sentiment Analysis
17	NLP with Python IV: How to handle Chinese
18	Modelling and machine learning with Python

4. Miscellaneous