Web Crawler Practice

Introduction to Web Crawler

Dr. Chun-Hsiang Chan

Department of Geography,
National Taiwan Normal University

Outline

- About CCH
- Course Introduction
- Grading Policy
- Why do you need to take this course?
- What will you learn from this course?
- Textbook

About CCH

現職:

國立臺灣師範大學地理系 助理教授

主要經歷:

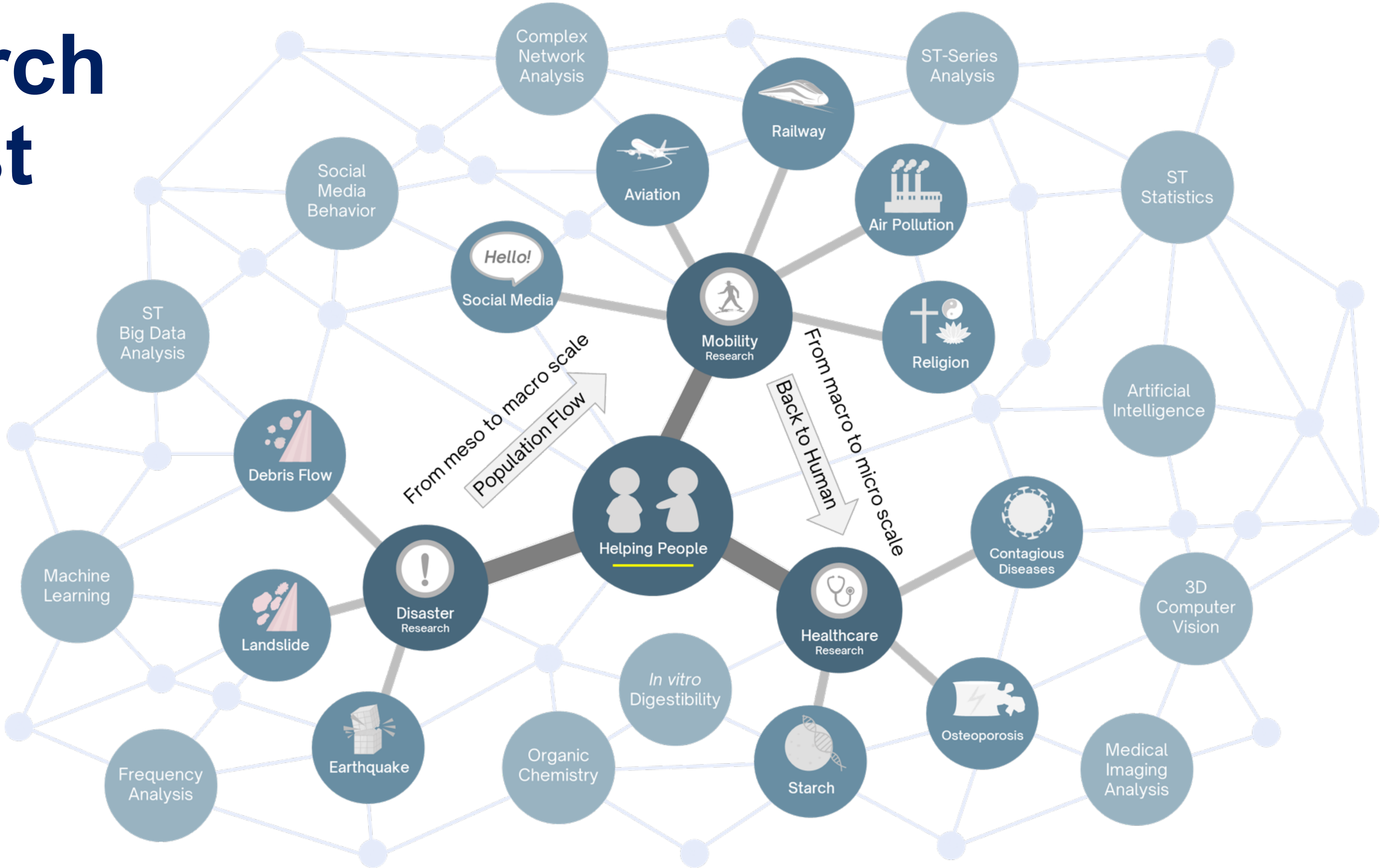
中原大學智慧運算與大數據學士班/碩士學位學程 助理教授
台灣資安鑄造股份有限公司 人工智慧分析顧問
臺北醫學大學醫學系放射線學科 博士後研究員
臺北市立萬芳醫院影像醫學部 博士後研究員
中央研究院社會學研究所 兼任資料分析師
資訊工業策進會資安科技研究所 工程師
國家災害防救科技中心坡地組 實習生
國立臺灣大學化學系 專題生

學歷:

國立臺灣大學地理環境資源學系 博士
國立臺灣大學地理環境資源學系 碩士
實踐大學食品營養與保健生技學系 碩士
國立臺北教育大學社會與區域發展學系 學士

SCAN ME

Research Interest



Previous Projects

Global

Global Airline Alliance Airport Network

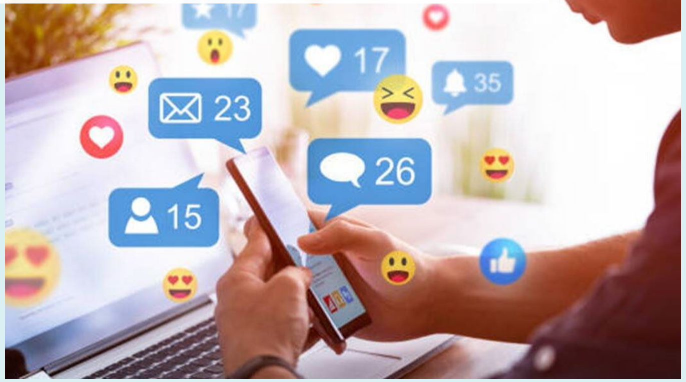
Timely Exposure Risk Estimation

Urban

COVID-19 Disease Transmission

Spatiotemporal Religious Dissemination

Social Media Marketing



Other Interests

Русский | Español | 日本語

當代英雄 從台灣出發

台大學生的俄羅斯
與後蘇聯行旅

金威澄、李正哲、李冠儒、詹竣翔、
江承頤、曾毓媛、吳連哲、王立宏、
廖德融、劉柏賢、沈祥茵
著

莫斯科台北經濟文化協調委員會
駐台北代表處代表　白樂賢、
台灣大學外文系副教授　索奧加、
台灣大學外文系副教授　熊宗慧
序

Герои нашего времени
Из Тайваня в Россию
и постсоветское пространство

My first Russian Book |
Published in Nov. 2021

Exhibition Staff |
Moscow, 2015

2015 TAIWAN TRADE MEETING IN ST. PETERSBURG
Bos' International Co. Ltd.

Exhibition Staff | St. Petersburg, 2015

Host | NTU Russian Night, 2017

Exhibition Staff | República Dominicana, 2015

Exhibition Staff | Colombia, 2015

MASIVA CONCURRENCIA DE VISITANTES A LA EXPO ESTE DOMINGO

Un joven con rasgos orientales realizaba una videollamada en inmediaciones del Pabellón Chino. CELSO RÍOS, ABC COLOR

ABC news | Paraguay, 2015
Exhibition Staff | Paraguay, 2015

Course Introduction

- Many datasets are available online and often consist of a vast amount of data. Downloading the entire dataset quickly can be a challenge. However, manually selecting data to download can result in missing important information.
- This course aims to teach you the most widely used web crawler package in Python for both static and dynamic websites, enabling you to efficiently download the desired datasets.

Course Introduction

- Due to time limitations, we will not help you review Python programming; therefore, if you are not familiar with programming or Python.
- You may watch and practice with my Python programming tutorials:
- Detailed version: https://toodou.github.io/UrbanGIS/index.html
- Simplified version: https://toodou.github.io/BigData/index.html
- To be honest, we should be familiar with data preprocessing and cleaning, but if you are a hardworking learner. I believe that you will understand and be familiar with all of them.
- So, … trust yourself … you can make it!

Course Introduction

Week	Date	Content
1	Feb. 19	Introduction to Web Crawler
2	Feb. 26	Web Crawler Ethics & Web Design (I) – HTML
3	Mar. 4	Web Design (II) – CSS (I)
4	Mar. 11	Web Design (III) – CSS (II)
5	Mar. 18	Web Design (IV) – JavaScript Basic
6	Mar. 25	Web Design (V) – Java Script Advance with XML and JSON
7	Apr. 1	Static Website Crawler – 批踢踢
8	**Apr. 8**	**Midterm Proposal Pitch**
9	Apr. 15	**[Video]** Solving the CAPTCHA
10	Apr. 22	**[Video]** Dynamic Website Crawler (I) – Selenium Introduction
11	Apr. 29	Dynamic Website Crawler (II) – Dcard and 小紅書
12	May 6	Dynamic Website Crawler (III) – Treads and Instagram
13	May 13	Social Media API – Facebook & X API
14	May 20	Social Media Crawler Practice – Facebook
15	**May 27**	**Final Report Presentation**
16	Jun 3	Final Exam Week (no class)

Grading Policy

All you have to do is study hard and feel free to ask question when you do not understand.

I believe that if you fulfill all required items, and then you will pass this course / get a high GPA.

Do not worry about the grade! The most important things is what you learn from this course.

Attendence	10%	Midterm Report	20%
Assignment	40%	Final Report	30%

Midterm Report

- As a programming course, you must practice as much as possible. In general, using a use case could enhance the motivation to learn and search for related information to achieve your final goals.
- In the midterm report, you only have 3 minutes to present which website or online information you attempted to download. To clearly demonstrate the data you obtained, you need to leverage a website as a dashboard to perform your discoveries.
- The recommended report architecture is as follows:
- (1) Background (why?); (2) Target Website (brief intro.); (3) Target Info. (Detailed); (4) Expected Dashboard (fake demo)

Final Report

- In the final presentation, each student has 5 – 6 minutes to present your online dashboard (website). The recommended report format is depicted as follows.

a) Background – Brief intro.

b) Target website – Screenshot with a brief intro.

c) Target Info. – Table, figure, and description

d) Introduce your dashboard – Charts with sufficient decription

Why do you need to take this course?

- As a member of the era of information explosion, most crucial information and data are available online; however, it is hard to manually obtain all data within a short time.
- In this course, we will teach you how to design your own (including get ting your website to show up online) and how to get website information from static, dynamic, and social media.

What will you learn from this course?

- According to the scheduled syllabus, you will learn …
  a) Website architecture (***HTML, CSS, and JS***)
  b) Front-end design
  c) GitHub
  d) Animation
  e) Visualization
  f) Static website crawler
  g) Dynamic website crawler

Textbook

- We have no fixed textbook or online tutorials.
- You must continually search for the necessary information to accomplish your final project and all the problems.

- **Suggested sources:**
  - Ryan Mitchell (2024) Web Scraping with Python (3rd). O'Reilly Media, Inc. ISBN: 9781098145354
  - Katharine Jarmul and Richard Lawson (2017) Python Web Scraping (2nd). Packet. ISBN: 978-1-78646-258-9

The End

Thank you for your attention!

Email: chchan@ntnu.edu.tw
Web: toodou.github.io

SCAN ME