COM5507 Social Media Data Acquisition and Processing
Week 1b. Data Science and Media Data Analytics at a Glance

Lectured by: Dr. Xinzhi ZHANG

Research Assistant Professor, Department of Journalism

Hong Kong Baptist University

5 Sep 2018 @ CityU M5064

Agenda

- What
  - What do data scientists do?
  - What is social media data acquisition and processing?
- Why
  - Why shall I learn this course?
  - Why web/social data?
- How (to start and succeed)
  - Tools installation
  - Social scientists’ and humanities scholars’ domain knowledge
  - Practice and getting your hands dirty!

Who are they…?

- data science and data scientists
- computational social science
- computational communication research
- digital humanities
- data-driven journalism
- computational journalism
- programmer journalism
- social informatics
- business analytics
- big data analytics
- social media analytics
- …

Inter-(multi-)disciplinary Areas

http://chasingdeer.co.uk/analyst-venn-diagram.html

https://www.r-bloggers.com/data-science-in-businesscomputational-social-science-in-academia/

https://knightlab.northwestern.edu/2013/06/28/want-to-build-a-data-journalism-team-youll-need-these-three-people/

Data science in action

- Define the problem
- Scouting the data sources
- Accessing to and collecting the data
- (Pre-)processing and cleaning the data
- Exploring the data
- Analyzing the data
- Interpreting the results
- Offering insights and solutions
- …

Data science pipeline - a verbal explanation

- **The "OSEMN Pipeline"**
- **O —** Obtaining our data
- **S** — Scrubbing / Cleaning our data
- **E —** Exploring / Visualizing our data will allow us to find patterns and trends
- **M —** Modeling our data will give us our predictive power as a wizard
- **N —** Interpreting our data
- - Reference

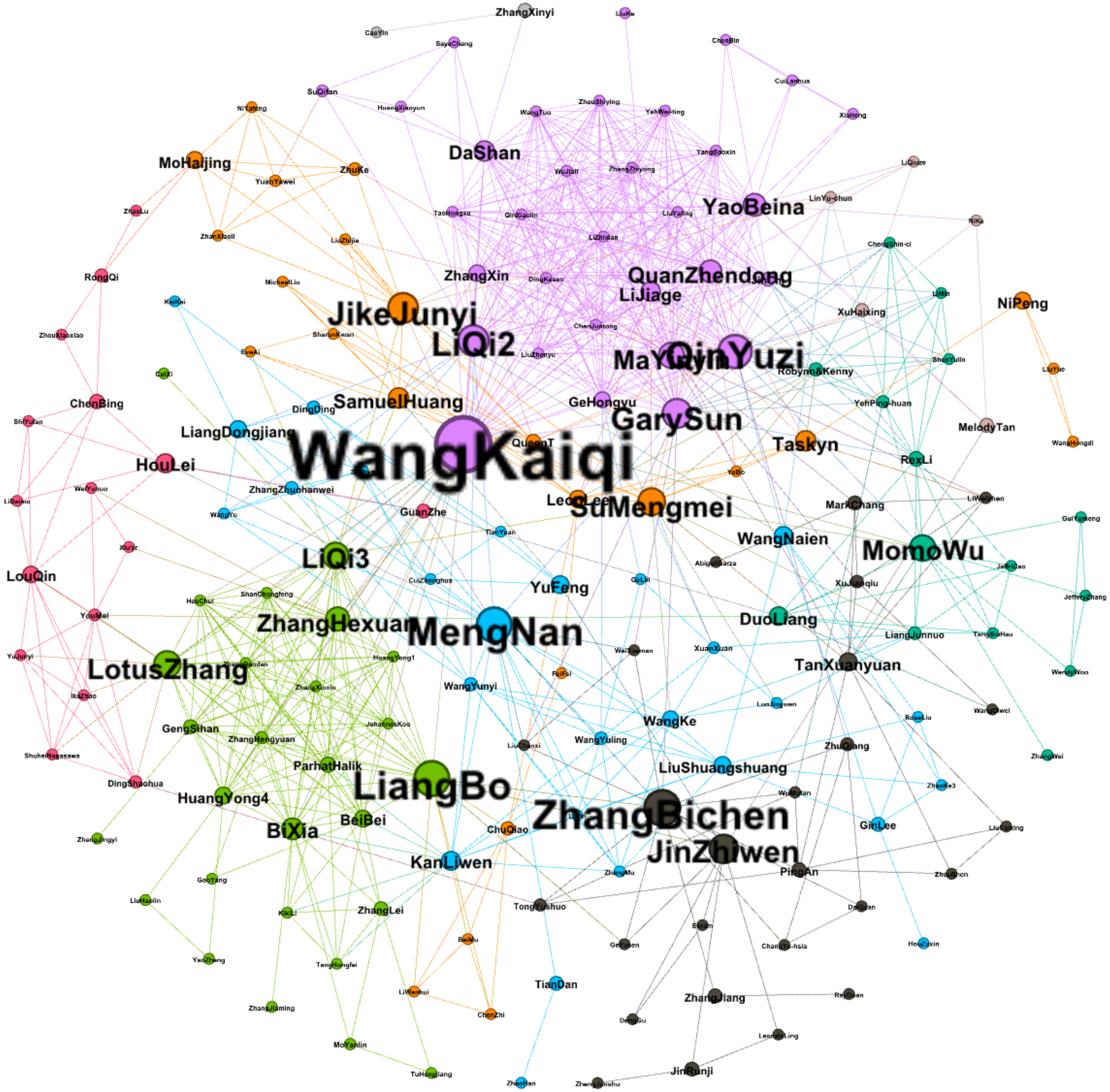
Data science pipeline - a graphical illustration

- Zacharias Voulgaris (2017): The Data Science Pipeline - Data Science: Mindset, Methodologies, and Misconceptions

The figure is removed here.
To view the figure, please check the above reference.

Examples

- Fivethirtyeight: Lionel Messi Is Impossible [Link]
- SCMP: Hong Kong Budget Proposal [Link]
- Obama’s 2013 Budget Proposal – by New York Times [Link]
- Gun Ownership [Link]
- “*Unfounded*” - The Global and Mail, 2017 [Data Journalism Award, 2017, Investigation of the Year] [Link]



Zhang (2017).
Figure 2. A snapshot of the contestant network of the Voice of China, illustrating the co-cover process (note: only the connected section of the graph is illustrated. The figure was plotted by Gephi).

Data acquisition and processing

- What is social media (web) data acquisition?
  - web data collection, data acquisition, screen scraping, data mining, web data harvesting, web crawlers…
  - “social media” – web data, cultural and social artifacts
- Web scraping
  - the practice of gathering data through any means other than a program interacting with an API (or, obviously, through a human using a web browser);
  - It is accomplished by writing an automated program that queries a web server, requests data (usually in the form of HTML and other files that compose web pages), and then parses that data to extract needed information (Mitchell, 2018).
- Automated data collection via social media or institutions
  - The usage of APIs (Twitter, Fb, NYT)
- It is the process of taking unstructured information from the web (webpages, sites, social media services, institutions) and turning it in to structured information that can be used in a subsequent stage of analysis.

https://www.quora.com/In-Python-how-can-I-save-data-from-a-website-to-CSV-using-BeautifulSoup

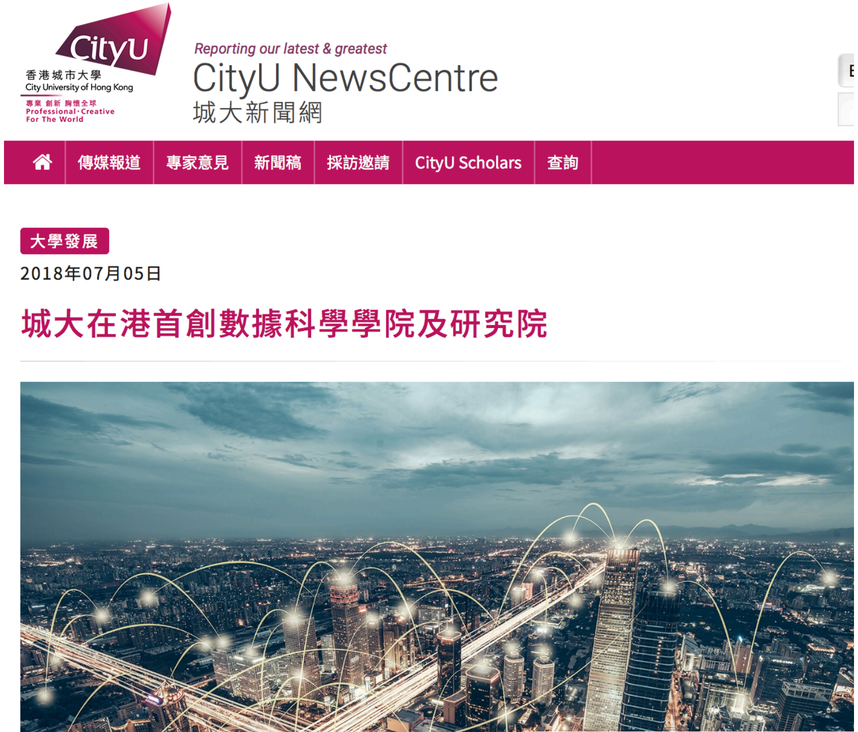
http://pbpython.com/web-scraping-mn-budget.html

Data acquisition and processing

- What is data processing?
- (pre-)processing, processing, data cleaning…
- The purpose of data processing is to transfer the collected data (i.e., the survey responses in the survey study, or the scraped webpages in the digital age) into machine-readable form, so that the data can be ready for further statistical analysis and interpretation.
- Reproducible
- Transparent
- Automated

Why taking this class?

- Increasing market demands
- The availability of digital footprints (digital traces)
- Internet as a rich database
- The price and difficulty of collecting and storing massive online data has dramatically reduced.
- The field itself is becoming more interdisciplinary.



Reporting our latest & greatest

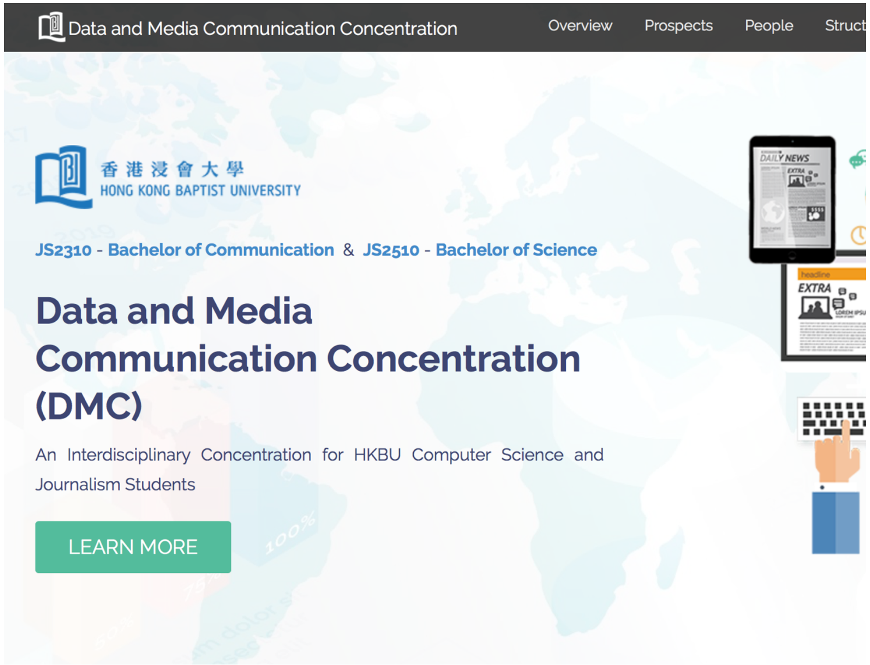
CityU NewsCentre

城大新聞網

大學發展

2018年07月05日

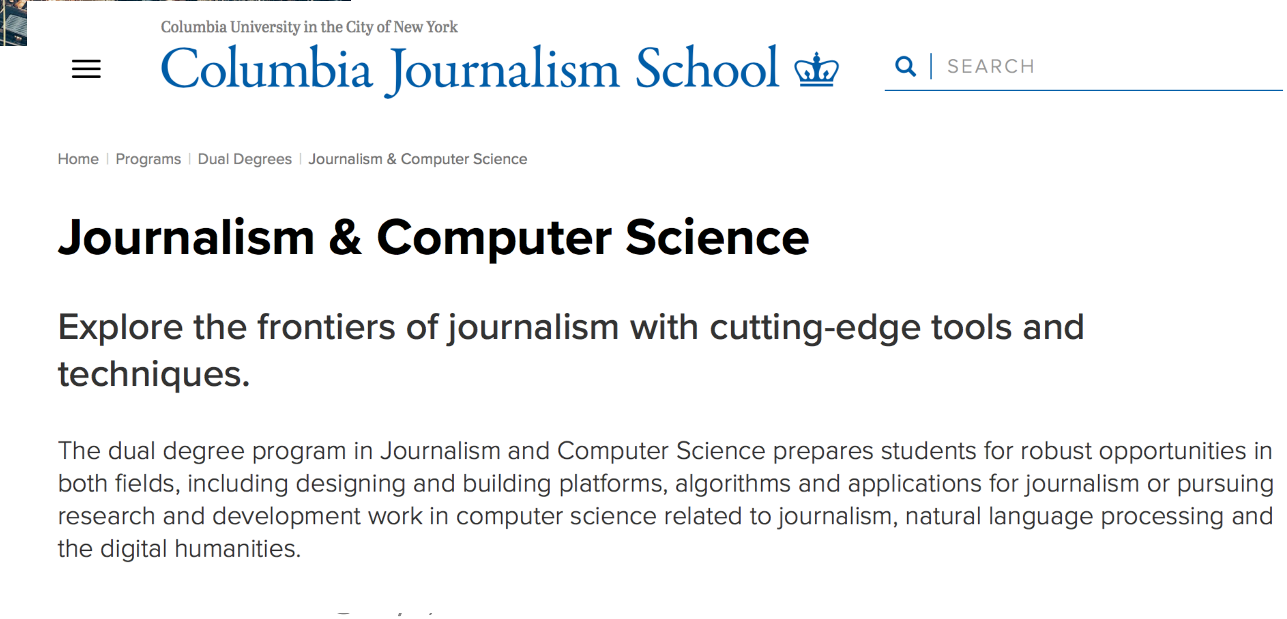
城大在港首創數據科學學院及研究院


Data and Media Communication Concentration
Overview
Prospects
People
Struct
香港浸會大學
HONG KONG BAPTIST UNIVERSITY
LEARN MORE

JS2310 - Bachelor of Communication & JS2510 - Bachelor of Science

Data and Media Communication Concentration (DMC)

An Interdisciplinary Concentration for HKBU Computer Science and Journalism Students


Columbia University in the City of New York
Columbia Journalism School
SEARCH
Home | Programs | Dual Degrees | Journalism & Computer Science

Journalism & Computer Science

Explore the frontiers of journalism with cutting-edge tools and techniques.

The dual degree program in Journalism and Computer Science prepares students for robust opportunities in both fields, including designing and building platforms, algorithms and applications for journalism or pursuing research and development work in computer science related to journalism, natural language processing and the digital humanities.

5 Sep 2018

Getting started: Tools

1. Talking to your computer: Command line interface (CLI)
2. Text Editor: Atom, or Sublime Text, or others
3. Platform for publishing and socializing: Git and GitHub, and Markdown language
4. The tool: Python 3.x and Jupyter Notebook

Getting started: Domain knowledge + motivations

- Domain knowledge
  - Theories from discipline areas: Journalism, advertising, marketing, management, sociology, political science, arts,…
  - Issues and topics: sports, fashion, popular culture, folk music, movies, cartoons, cuisines,…
- Motivations
  - ”When there is a gap between the ideal situation and the reality, there is an investigation.”
  - “Trend? correlation? outliers?”

Tips: How to ask for help?

- "Hacking skills:" Asking questions and finding answers
- Key characteristics of hackers
  - willing to find answer on their own
  - knowledgeable about where to find answers on their own
  - unintimidated by new data types or packages
  - not being afraid of saying that they don't know the answer
  - polite but relentless
- Reference: Eric Steven Raymond: *How To Ask Questions The Smart Way* [*Must read, URL here*]

Tips: How to ask for help?

- Asking reproducible questions (other can understand your question and rework it on their own machines)
- What is the question you are going to answer?
- What steps did you use to find out the answer?
- What is the expected output?
- What do you see instead?
- What version and operating system are you using?
- What are the data analytical tools/functions you are using?
- What other solutions have your thought about?

Tips: How to ask for help?

- Be polite: others do not have the obligations to help you
- Be explicit: Try to be as specific and detailed as you can! Don't ask too general questions
- Following up and post solutions - helping others, knowledge increments

Tips: Where to look for help?

- Ready-made:
  - Software’s manuals and helping documents
  - Official tutorials
- Online sources:
  - Stack overflow
  - GitHub pages
  - Google and Google scholar
  - Course forums
  - WeChat or Twitter public accounts
  - Online courses
- Offline sources
  - A skilled friend?
  - Workshops, seminars, hackathons, meetups

Tips: Practice, practice, practice!

- Get your hands dirty!

Summary

- Course position and mutual expectations:
  - a course for “humanities and social science scholars working in data science and data scientists to know about humanities and social science scholars”
  - hacker’s characteristics: this course is only the start, not the end; asking questions and solve the problems
- My expectation: questions, approaches (“research design”) are more important than methods and coding, on top of that, it is your curiosity and motivation that make you go further and further; your expectations: give your imaginations and thoughts a wing of codes and a lens of telescope
- Course materials: technical issue (not data science course), coding tutorials (numerous tutorials and online short video lectures online, plus multiple technical and vocational training colleges), actual cases (value added part, journalism and business), academic papers (social science’s mindset)

Assignments

- Getting familiar with the tools
- Preview: Python
  1. What are the basic commands in Python?
  2. What are the common data structure (data types) in Python?
  3. What are the flow control in Python?

References: The basics

- Earl Babbie: The practice of social research (any version will be fine)
- Matthew J. Salganik: Bit by bit: social research in the digital age
- Ryan Mitchell: Web scraping with Python (2017, the 2nd edition)
- Katharin Jarmul & Richard Lawson: Python web scraping (2017, the 2nd edition)

References: Academic journals

- *Communication Research*
- *Journal of Computer-Mediated Communication*
- *Computers in Human Behaviors*
- *Cyberpsychology, Behavior, and Social Networking*
- *New Media & Society*
- *Information, Communication, & Society*
- *Digital Journalism*
- …

References: Public Twitter accounts and blog platforms

- High-quality Twitter accounts to follow (must read):
  - https://twitter.com/FiveThirtyEight
  - https://twitter.com/paulbradshaw
  - https://twitter.com/ProPublica
  - https://twitter.com/pewjournalism
  - https://twitter.com/GuardianData
  - https://twitter.com/ftdata
  - https://twitter.com/WSJGraphics
  - https://twitter.com/PostGraphics
  - https://twitter.com/BBC_News_Labs
  - https://twitter.com/BBGVisualData
  - https://twitter.com/ReutersGraphics
  - https://twitter.com/LATimesGraphics
  - https://twitter.com/UpshotNYT
  - https://twitter.com/GlobeSpotlight

- Professional organizations and blog communities
  - https://medium.com
  - Global Investigative Journalism Network
  - ICA Computation Methods Interest Group
  - DT Data
  - The Data and News Society @ HKBU

References: A batch of GitHub “Repos”

- # “Repos” on general data science
  - Data-X@Berkerly
  - Computational Sociology @ Duke
  - SICSS [2018]
- # data science based on python
  - WhirlwindTourofPython
  - PythonDataScienceHandbook
- # similar courses offered by other institutions
  - webscraping tutorial
  - JOUR7080 @HKBU
  - Big data course @NJU