Semester 2, 2022-2023, Hong Kong Baptist University

COMM7780 Big Data Analytics for Media and Communication

Tuesday, CVA 506

Instructor: Dr. Yuner ZHU 朱蘊兒
Office: CVA 925 | Telephone: 852-3411-6553 (Office)
Email: yunerzhu@hkbu.edu.hk
Office hours: 3-6pm, Thursday (By appointment)

Tutor: Randy Lin (22482016@life.hkbu.edu.hk)
Tutorials: By appointment

Goal

In this era of social media, everyone and everything is online. People receive information, connect with others and express their opinions on social media. However, with limited training in computational skills, social scientists may lack a capacity to systematically process and interpret online observational data, which is fundamentally different with experimental data in terms of structure, size and variability.

By drawing inspiration from computer sciences, this course introduces the basic background on natural language processing, web crawling, sentiment analysis, data visualization, and machine learning. More importantly, aside from the theoretical underpinnings of computational methods, you will learn critical know-how about social media analytics and gain practical implementation experiences in the class. We will use Python coherently and exclusively throughout the course.

Format

We will have a three-hour meeting every time, involving lectures and hands-on exercises. Your participation is critical. Feel free to ask questions and give comments during and after class.

Assessment Methods (AMs)

Quiz	8 * 1%
Weekly Assignments	12 * 6%
Hackathon Team Project	
- Report	8%
- Code	8%
- Peer Evaluation	4%
Total	100%

***Academic dishonesty in the form of cheating and/or plagiarism in all its forms will result in a grade of "F" for the assignment and exams. For details, see:**
https://ar.hkbu.edu.hk/quality-assurance/university-policy-and-guidelines/academic-integrity

Session 1: Python Set-up

Week 1 (9 Jan) Introduction & Installation

1. Introduction of basic Python programming
    a. Workflow
    b. Typical RQs in Computational Communication Research
2. Installation
    a. Python Environment
    b. Interactive Editor: Jupyter Notebook
    c. Package 1 (Vector/Matrix + Basic Math): Numpy
    d. Package 2 (Data Frame): Pandas
    e. Package 3 (Web Crawling): Selenium
    f. Package 4 (Visualization): Plotly
    g. Package 5 (Machine Learning): Sklearn
3. Warm-up Practice:
    a. Print() Function
    b. Basic Math: + - * / %, power
4. Instructions on Assignment Submission
5. Weekly Assignment: Basic Math (2 marks) + Quiz (4 marks)

Week 2 (16 Jan) Data Type

1. Data Type
    a. Number
    b. List
    c. String
        i. String Operation: split, strip, replace
    d. Dictionary
    e. Data Frame
    f. Data Type conversion
2. Weekly Assignment (6 marks)

Week 3 (23 Jan) Function

1. File I/O
2. For Loop and While Loop
3. List Comprehension
4. If/Else statement

Week 4 (30 Jan) Function (Continued)

1. Function Recap
2. Customer Function: Create your own function
3. Practice: Presidential Inauguration Speech
    a. Sentence count
    b. Word count
    c. Readability Test
4. Weekly Assignment: Lexicon-based Text Analysis

Session 2: Visualization

Week 5 (6 Feb) Bar/Line Chart & Scatter Plot & Map

1. Plotly: Register, Token
2. Mapbox: Register, Token
3. Weekly Assignment: Testing the relationship between Freedom indexes and GDP values of countries

Week 6 (13 Feb) No Class. Lunar New Year Holidays.

Session 3: Web Crawling

Week 7 (20 Feb) Knowing HTML

1. Intro to HTML
    a. Two types of web sites: Static & Dynamic
    b. Tag: name, value
    c. Content
2. Regular Expressions
3. Create your own website
4. Weekly Assignment: Website design

Week 8 (27 Feb) Web Crawling

1. Selenium
2. Programs that surf the Internet: Selenium & Beautiful Soup
3. Practice: crawl Google Search Results
4. Weekly Assignment: IMDB

Week 9 (5 Mar) Web Crawling (Continued)

1. Collect posts and comments from Red Book & Weibo
2. Practice: Douban
3. Weekly Assignment

Session 4: NLP

Week 10 (12 Mar) Natural Language Processing

1. English tokenization: Regular expressions
2. Chinese tokenization: Jieba
3. Co-occurrence analysis
4. Weekly Assignment

Week 11 (19 Mar) Word2Vec

1. Word2Vec: Representing texts with numbers
2. Practice: predicting ratings from text (Amazon Review Data)

3. Algorithmic auditing: assessing gender bias by WEAT
4. Weekly Assignment: Word embedding of IMDB reviews

* Readings:
1. Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356, 183–186. https://doi.org/10.1145/3306618.3314267
2. Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. 30th Conference on Neural Information Processing Systems, 1–9.

Session 5: Machine Learning

Week 12 (26 Mar) Unsupervised Machine Learning

1. ML in a nutshell: mechanism & convention
2. Basics of Linear Algebra
3. Matrix transformation
4. Feature Selection: by variance, by information redundancy, by discriminative ability
5. K-means
    - Practice: clustering Amazon users into groups
    - Parameter tuning: How to choose K?
6. Weekly Assignment

Week 13 (2 Apr) No Class. Ching Ming Festival.

Week 14 (9 Apr) Topic Modeling & Course Summary

1. Concepts in topic modeling
2. LDA
3. Mining the trends of topics in People’s Daily news
4. Course summary

Week 15 (16 Apr) Hackathon

References:

O’Reilly Python Handbook Series: https://github.com/Jianhua-Wang/oreilly-animal-books-for-Python

Tutorials: https://realpython.com/