Semester 2, 2022-2023, Hong Kong Baptist University

COMM7780 Big Data Analytics for Media and Communication

Tuesday, CVA 517

Instructor: Dr. Yuner ZHU 朱蘊兒
Office: CVA 925 | Telephone: 852-34116553 (Office)
Email: yunerzhu@hkbu.edu.hk
Office hours: By appointment
Anonymous QA Channel: https://www.tapeclub.net/uu/Y1CE6E/SF816JCB

Tutor: Luo Yifeng (21482845@life.hkbu.edu.hk)
QA Sessions: X – X, CVA

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.**

Session 1: Python Set-up

Week 1 (Jan 10) 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

Week 2 (Jan 17) Data Type & Function

1. Data Type
    a. Number
    b. List
    c. String
    d. Dictionary
    e. Data Frame
2. Some useful built-in functions
    a. Print
    b. File I/O
    c. Loop: For loop & While loop [optional]
    d. Condition: If…Else…
    e. Indexing
    f. String Operation: split, strip, replace, match
3. Create your own function
4. Practice: Presidential Inauguration Speech
    a. Sentence count
    b. Word count
    c. Readability Test
5. Weekly Assignment: Lexicon-based Keyword Searching

Week 3 (Jan 24) No Class. Lunar New Year Holidays.

Week 4 (Jan 31) Function (continued)

1. Recap: File I/O, For loop
2. If/Else statement
3. Customer Function
4. Practice: Readability Test of Inauguration Speech
5. Weekly Assignment: Lexicon-based Text Analysis

Session 2: Visualization

Week 5 (Feb 7) 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

Session 3: Web Crawling

Week 6 (Feb 14) 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 7 (Feb 21) Web Crawling

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

Session 4: NLP

Week 8 (Feb 28) Chinese Tokenization + Word2Vec

1. Dictionary-based + HMM: Jieba
2. Discriminatory index: TF-IDF
3. Co-occurrence analysis
4. Word2Vec
5. Practice: create a word embedding from Chinese Wikipedia
6. Practice: predicting ratings from text (Amazon Review Data)
7. Algorithmic auditing: assessing gender bias by WEAT
8. 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 9 (Mar 7) Unsupervised Machine Learning

1. ML in a nutshell: mechanism & convention
2. Feature Selection: by variance, by information redundancy, by discriminative ability
3. K-means
    - Practice: clustering Amazon users into groups
    - Parameter tuning: How to choose K?
4. LDA (Topic Modeling)
5. Weekly Assignment: Mining the trends of topics in communication research
   ** Applying LDA topic modeling in communication research: Toward a valid and reliable methodology*

Week 10 (Mar 14) Supervised Machine Learning (concept)

1. Intro to Supervised Machine Learning:
    a. General procedure of Expectation-Maximization (EM)
    b. Expectation is given based on current setting of weights. Maximization is to maximize the value of likelihood between expectation and real value, or conversely to minimize the difference (cost function) through back propagatio0 n.
    c. Back propagation: derivative of cost function + chain rule
2. Overfitting vs Underfitting
    a. Techniques including K-fold Cross-validation and regularization are adopted to avoid overfitting problem.
    b. Cross-validation: dataset is usually split into three parts: training set, test set and validation set in a 6-2-2 way.
    c. L2 Regularization: punish model when excessive features are adopted
3. Model Evaluation
    a. Classification: Accuracy, F1 score, and Cross-entropy Loss
    b. Regression: Mean Squared Error and Mean Absolute Error
4. Learning curve
5. Weekly Assignment: Identify the optimal/sufficient size of training data.

Week 11 (Mar 21) Supervised Machine Learning (practice)

1. K Nearest Neighbors (k-NN): Euclidean distance + Data normalization vs Cosine Similarity
    - Practice: handwriting digit recognition
2. Naïve Bayes:
    - Practice: spam detection/Stanford sentiment dataset
3. Bayesian Network
4. Weekly Assignment 9: KNN for political affiliation detection

Week 12 (Mar 28) Supervised Machine Learning (continue)

1. Background: Logistic Regression
2. Support Vector Machine
    - Practice: spam detection/Stanford sentiment dataset
3. Neural Network

4. Weekly Assignment: SVM for political affiliation detection

Week 13 (Apr 4) Using APIs

1. Knowing APIs (Application Interface)
2. Authentication: HTTP, Basic, Bearer, Keys, OAuth
3. Reading documentations: Rate Limit
4. Practices: Reddits, Instagram, Google Translate, Bloomberg Stock Market and Financial News API, Amazon
5. Weekly Assignment

Week 14 (Apr 11) No Class. Easter Holidays.

Week 15 (Apr 18) Hackathon

References:

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

Tutorials: https://realpython.com/