**Computational Methods for Online Media Analytics**

*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 computational 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 (tutorial on R can be provided on request).*

**Session 1: Python Set-up**

***Week 1 Introduction & Installation***

1. Introduction of basic Python programming
   1. Workflow
   2. Typical RQs in Computational Communication Research
2. Installation
   1. Python Environment
   2. Interactive Editor: Jupyter Notebook
   3. Package 1 (Vector/Matrix + Basic Math): Numpy
   4. Package 2 (Data Frame): Pandas
   5. Package 3 (Web Crawling): Selenium
   6. Package 4 (Visualization): Plotly
   7. Package 5 (Machine Learning): Sklearn
3. Instructions on Assignment Submission [Optional]

***Week 2 Data Type & Function***

1. Data Type
   1. Number
   2. List
   3. String
   4. Dictionary
   5. Data Frame
2. Some useful in-built functions
3. Print
4. File I/O
5. Loop: For loop & While loop [optional]
6. Condition: If…Else…
7. Indexing
8. String Operation: split, strip, replace
9. Basic Math: + - \* / %, power, sum, average, length, count, min, max, sort
10. Create your own function: normalization, Euclidean distance
11. Practice: King’s Speech
    1. Word counts
    2. Sort & max/min
    3. Remove stop words
    4. Stemming with NLTK [optional]
    5. Dictionary-based Sentiment Analysis
12. Homework 1: Amazon Review data (sentence count, word counts, sentimental analysis)

**Session 2: Visualization**

***Week 3 Bar/Line Chart & Scatter Plot***

1. Plotly: Register, Token
2. Practice: Word Count of King’s Speech
3. Practice: Scatter plot of TED videos
4. Homework 2: Amazon Review data (scatter plot of products, sentiments and ratings)

***Week 4 Mapping***

1. Google API & Geocoding
2. Practice: Yelp Dataset
3. Homework 3: Mapping HK Legco election results

**Session 3: Web Crawling**

***Week 5 Knowing HTML & Web Crawling***

1. Intro to HTML
   1. Two types of web sites: Static & Dynamic
   2. Tag: name, value
   3. Content
2. Selenium
3. Practice: crawl headlines from New York Times
4. Homework 4: Mingpao

**Session 4: NLP**

***Week 6 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 app rates from text (Amazon Review Data)
7. N-gram [optional]
8. Homework 5: create a word embedding from Mingpao reports

**Session 5: Machine Learning**

***Week 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. Homework 6: Mining the trends of topics in communication research

*\* Applying LDA topic modeling in communication research: Toward a valid and reliable methodology*

***Week 8 Supervised Machine Learning (concept)***

1. Intro to Supervised Machine Learning:
2. General procedure is called Expectation-Maximization (EM).
3. 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 propagation.
4. Back propagation: derivative of cost function + chain rule
5. Overfitting vs Underfitting
6. Techniques including K-fold Cross-validation and regularization are adopted to avoid overfitting problem.
7. Cross-validation: dataset is usually split into three parts: training set, test set and validation set in a 6-2-2 way.
8. L2 Regularization: punish model when excessive features are adopted
9. Model Evaluation
10. Classification: Accuracy, F1 score, and Cross-entropy Loss
11. Regression: Mean Squared Error and Mean Absolute Error
12. Learning curve
13. Homework 7: Identify the optimal/sufficient size of training data.

***Week 9 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. Homework 8: KNN for political affiliation detection

***Week 10 Supervised Machine Learning (continue)***

1. Background: Logistic Regression
2. Support Vector Machine
   * Practice: spam detection/Stanford sentiment dataset
3. Neural Network
4. Homework 9: SVM for political affiliation detection