CS105 - Data Analysis Methods

UC Riverside

Project Proposal

In the final project for this course, you will apply the techniques learned in this class to analyze a data sets of interest to you. Your goal should be to create an original project that you would be proud to show off to a potential employer. The project will require you to collect and obtain several data-sets that can be correlated in someways, clean, integrate and apply EDA on the data, perform analysis using ML techniques, and then present your findings by building a dashboard for the intended audience.

Sample Project: Yelp is a great website for users and business owners alike, however, there are still several features that will be interesting to add. Lets assume will focus on businesses in Riverside only, we can build a web-crawler to crawl Yelp's website for businesses in Riverside. We can crawl or locate a dataset of health ratings for restaurants' in the same area. We can crawl or locate a dataset for a business's BBB rating, and so forth. Once all the data is obtained, we will clean the data, integrate the various datasets, and apply EDA. When applying EDA, one item we may be interested to answer is number of businesses per category, number of reviews per category type, avg rating per category type, etc. Finally, we must outline some data analysis tasks. For example: 1) train an ML algorithm to predict the category' based on reviews (possibly do this only for restaurants). 2) predict if a review is 'funny', 'useful', 'interesting', etc. 3) extract the main entities and sentiment based on reviews. 4) analyze whether health ratings are correlated with location or another feature.

The project can be done in teams of 2. It is critical that each person on the team contribute to the project's codebase, because your grade will be determined based on the Github commit history. If you are pair programming (working on the notebook together), then be sure to take turns coding so that the commits reflects the pair programming effort.

The project is divided into several phases. In the first phase, we ask you to collect a dataset through web-scraping. You are asked on build a simple web-crawler in Python to collect data of interest. You are also asked to obtain more than one dataset, the additional datasets can either be obtained via web-crawling, APIs, downloading online, etc. The datasets used must be sufficiently large and non-trivial. You can consult with me if you have questions regarding this point.

If you have a completely different idea for a data science project that does not fit neatly with the requirements outlines, then please come talk to me.

0.1 Proposal

The first turn-in is the proposal, which should describe the objective of your project. This should include:

- Which datasets (minimum is 2) you plan to use and how you will obtain them (crawling, API, download). Include links to each dataset you plan to use. Note, APIs often require payment but they sometimes allow limited academic use if you email them. You should verify that the items identified is doable.

- Description of how the datasets are correlated, what information they provide, and the type of analysis you plan to perform.

0.2 Useful links for datasets

Bike Sharing `https://www.inferentialthinking.com/chapters/08/5/Bike_Sharing_in_the_Bay_Area.html`

Network/Graph Data `https://snap.stanford.edu/data/`

Census Data `https://www.census.gov/data/datasets.html`

Data.gov `https://www.data.gov/`

County of Riverside `https://data.countyofriverside.us/`

CDC API `https://open.cdc.gov/apis.html`

DCInbox (congress newsletters) `https://www.dcinbox.com/`

Twitter API `https://developer.twitter.com/en/docs`

NBA API `http://nbasense.com/nba-api/`

Have you found additional interesting sources for datasets? Please share so we can add them to the list.

0.3 Submission

Submit your proposal to iLearn. Note, your file should be named $teamname_{p}roj_{p}roposal$. Submissions will not be accepted via email; you must turn your assignment via iLearn. Submit only one proposal per team.