Search GitHub | Pull requests | Issues | Marketplace | Explore

Anahita

anahitabahri

Follow

Block or report user

Boston, MA

Overview | Repositories 8 | Stars 8 | Followers 7 | Following 6

Pinned repositories

Million-Song-Project

Jupyter Notebook 1

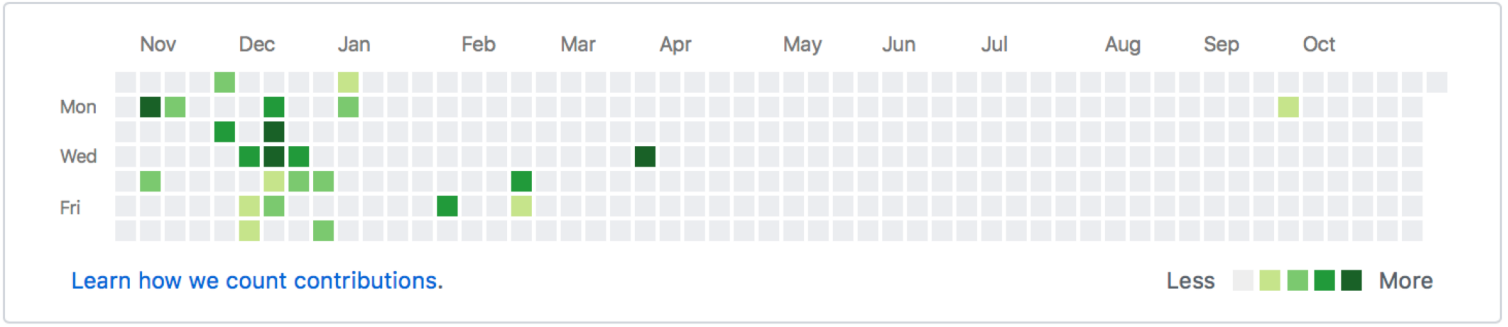
Rap-Twitter-Analysis

R

Yelp-Dataset-Challenge

Jupyter Notebook

87 contributions in the last year



Midterm Project

Midterm Project

- **Data Analysis Project**: Choose a dataset that is **relevant to your career goals** or your personal interest and propose an analysis that includes fitting at least a multilevel model.

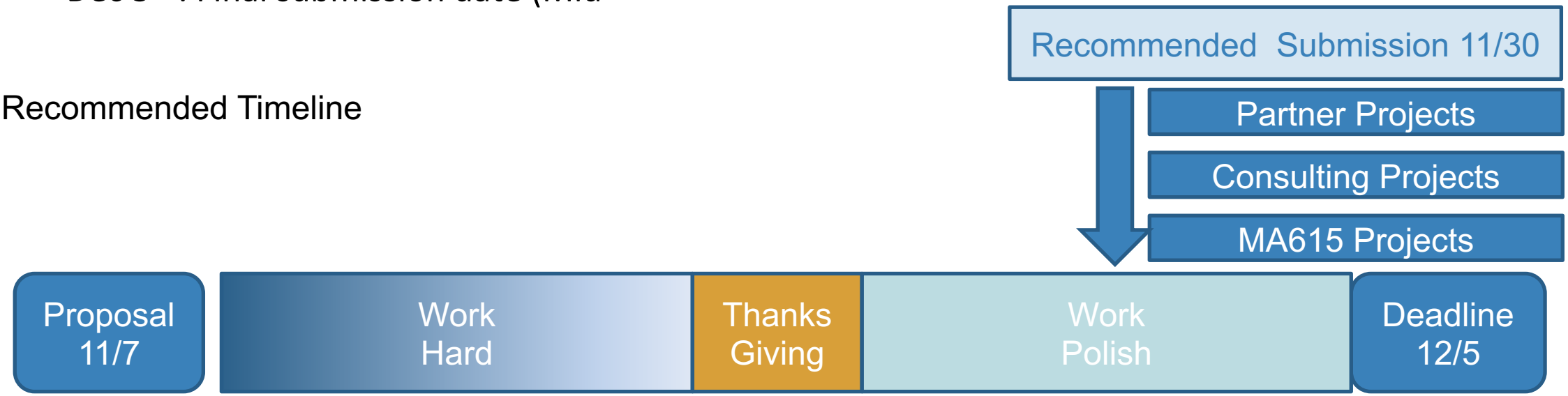
Data Analysis Project

- Example data:
  - Tech:
    - Yelp Data challenge: https://www.yelp.com/dataset/challenge
    - AirBnB: http://tomslee.net/airbnb-data-collection-get-the-data
  - Consumer:
    - Customer Revenue prediction: https://www.kaggle.com/c/ga-customer-revenue-prediction
  - Medical:
    - Medicare, CDC: https://data.medicare.gov/data/. https://wonder.cdc.gov
  - Financial:
    - IMF: http://data.imf.org/?sk=388DFA60-1D26-4ADE-B505-A05A558D9A42
    - Lending club: https://www.lendingclub.com/info/download-data.action
  - Music:
    - Million Songs: https://labrosa.ee.columbia.edu/millionsong/
  - etc. + extra bonus if you can combine different datasets.
- Criteria: A “LARGE” dataset with at least 10 groups that’s “Interesting”.

Timeline

- Nov 7th: Proposal Due
  - You are to chose from the two options below and propose a project that will help you showcase your skills to your future employee.
- Nov 30th: **Recommended** submission date
- Dec 5th : Final submission date (Mid

Recommended Timeline

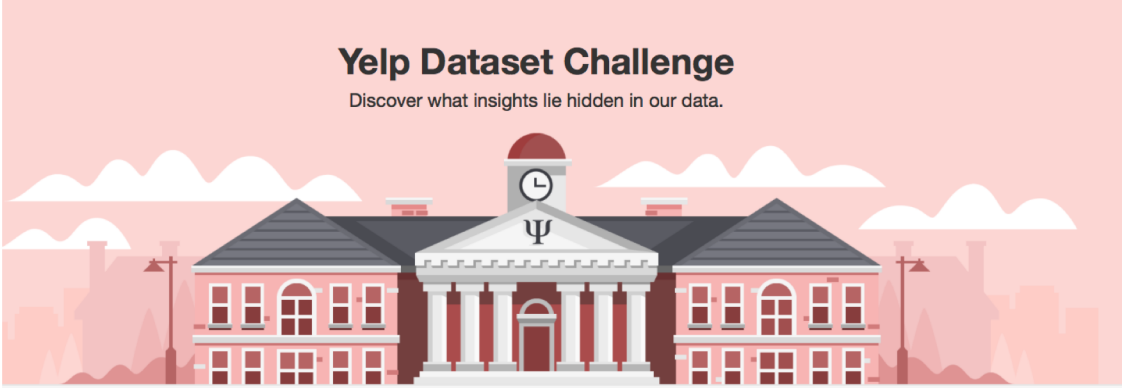


Grading Rubric for the project

- (10) Proposal
- (10) Overall Format: Can you confidently show it to a recruiter?
  - Does it look professional? Is it written in proper language?
- (10) Novelty: New in some ways and interesting?
- (30) Accuracy: Model choice reasonable? Interpretations correct?
- (20) Validation: Detailed model checking to justify the result?
- (20) Discussions: Assessment of the result. Limitations? Future directions?
- (+10) Technical:
  - How did you deal with the big data challenge?
  - Did you integrate data from multiple sources?
- (∞) Passion

Yelp Data Challenge

- The 13th installment
  - https://www.yelp.com/dataset/challenge



What is the dataset challenge?

The challenge is a chance for students to conduct research or analysis on our data and share their discoveries with us. Whether you're trying to figure out how food trends start or identify the impact of different connections from the local graph, you'll have a chance to win cash prizes for your work! See some of the past winners and hundreds of academic papers written using the dataset.

The Challenge

We challenge students to use our data in innovative ways and break ground in research. Here are some examples of topics we find interesting, but remember these are only to get you thinking and we welcome novel approaches!

Photo Classification

Maybe you've heard of our ability to identify hot dogs (and other foods) in photos. Or how we can tell you if your photo will be beautiful or not. Can you do better?

Natural Language Processing & Sentiment Analysis

What's in a review? Is it positive or negative? Our reviews contain a lot of metadata that can be mined and used to infer meaning, business attributes, and sentiment.

Graph Mining

We recently launched our Local Graph but can you take the graph further? How do user's relationships define their usage patterns? Where are the trend setters eating before it becomes popular?

- **Public Datasets**
  - https://github.com/awesomedata/awesome-public-datasets
  - https://github.com/apiad/datasets-list
  - https://github.com/datasets/openml-datasets/tree/master/data the same data as this list https://www.openml.org/search?type=data
  - https://archive.ics.uci.edu/ml/datasets.html
  - https://www.data.gov/
  - https://www.kaggle.com/datasets
  - http://datamob.org/
  - https://sites.google.com/a/drwren.com/wmd/details
  - https://data.cityofnewyork.us/data
  - http://snap.stanford.edu/data/index.html
  - http://aws.amazon.com/datasets/

- **Event Data**
  • GeoLife GPS Trajectories http://research.microsoft.com/en-us/downloads/b16d359d-d164-469e-9fd4-daa38f2b2e13/
  • Activity Recognition in the Home Setting Using Simple and Ubiquitous Sensors (MIT Project 2004) http://courses.media.mit.edu/2004fall/mas622j/04.projects/home/
  • https://sites.google.com/a/drwren.com/wmd/home
  Frequent Itemset Mining Dataset Repository
  • http://fimi.ua.ac.be/data/
  Airline On-time Performance
  • http://www.eecs.wsu.edu/~yyao/StreamingGraphs.html
  • http://openflights.org/data.html
  Collection and Streaming of Graph Datasets
  • http://www.eecs.w su.edu/~yyao/StreamingGraphs.html
  Data Streams
  • http://www.quora.com/Where-can-I-find-public-or-free-real-time-or-streaming-data-sources
  • 3 hourly weather forecast and observational data - UK locations http://data.gov.uk/dataset/metoffice_uklocs3hr_fc
  • There is also an IRC chan with the live log of wikipedia edits on the #en.wikipedia channel of

- **New York Taxi Datasets**
  - https://data.ny.gov/
  - TLC Trip Record Data

  http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml
  http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml
  - Another set published by chris whong http://chriswhong.com/open-data/foil_nyc_taxi/

  this data set is used for the DEBS 2015 challenge http://www.debs2015.org/call-grand-challenge.html
  - Another Description of this dataset http://publish.illinois.edu/dbwork/open-data/

  OpenStreet Map
  - http://wiki.openstreetmap.org/wiki/Planet.osm

  Dublin Bus GPS sample data from Dublin City Council (Insight Project)
  - https://data.gov.ie/dataset/dublin-bus-gps-sample-data-from-dublin-city-council-insight-project

  Further data set at https://data.gov.ie

  Wikipedia Dump
  - Wikpedia Dump https://dumps.wikimedia.org/

  Amazon Review Data Downloader
  - https://github.c om/aesuli/Amazon-downloader