EMORY UNIVERSITY

GOIZUETA BUSINESS SCHOOL

Data for Predictive Analytics Projects

Predictive Analytics: Team Projects

Vilma Todri
Assistant Professor
Goizueta Business School
Emory University
vtodri@emory.edu

Data Sources

You will mine actual data for a problem of interest. These could be data from:

- a problem from your past / current / future job,
- something of interest to the school,
- data acquired from the web,
- ..

The following slides provide several examples of data sets

- *Be critical*; not all data sets are appropriate for a project

Publicly Available Data Sets: Repositories

- Kaggle Data Sets
- KDD Cup
  - http://www.kdd.org/kdd-cup
- Public Data Sets on AWS
- Stanford Large Network Dataset Collection
- NYC Open Data
  - https://data.cityofnewyork.us/
- Quandl.com
- Yahoo! Labs
  - http://webscope.sandbox.yahoo.com/catalog.php

Publicly Available Data Sets: Academic Repositories

- University of Edinburgh
  - http://www.inf.ed.ac.uk/teaching/courses/dme/html/datasets0405.html
- UC Irvine Machine Learning Repository and Knowledge Discovery in Databases Archive
- LIBSVM Data
  - http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/
- ..

Publicly Available Data Sets Examples

- Lending Clubs
  - https://www.lendingclub.com/info/download-data.action
  - Prosper.com, etc.
- Recommender Systems
  - MovieLens, Netflix, Last.fm, BookCrossing, BibSonomy, Delicious, LDOS-CoMoDa (context rich movie recommender dataset), Million Song Dataset, News Recommendations (http://www.clef-newsreel.org/), DBpedia linked data sets (http://sisinflab.poliba.it), online dating (http://www.occamslab.com/petricek/data/), https://gist.github.com/entaroadun/1653794, etc.
- Privacy
  - https://data.gov.au/dataset/2013-community-attitudes-to-privacy-survey-data
- Patients and Hospitals
  - https://sites.google.com/site/informsdataminingcontest/home

Publicly Available Data Sets Examples

- Expedia
  - https://www.kaggle.com/c/expedia-personalized-sort
- YouTube videos
  - http://netsg.cs.sfu.ca/youtubedata/
- Sentiment Analysis
  - http://ai.stanford.edu/~amaas/data/sentiment/index.html
  - http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#datasets
- SMS spam
  - http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/
- Microsoft Bing
  - http://research.microsoft.com/en-us/projects/mslr/

Publicly Available Data Sets Example

- USDA National Nutrient Database
  - http://www.ars.usda.gov/Services/docs.htm?docid=8964
- Stack Exchange data dump
- Wikipedia data dump
- Reddit data dumps
  - e.g., http://www.reddit.com/r/redditdev/comments/bubhl/csv_dump_of_reddit_voting_data/
- Yelp.com
  - https://www.yelp.com/dataset_challenge
- Github
  - e.g., https://github.com/caesar0301/awesome-public-datasets
- …

Collect your own Data

- APIs
  - NYT
  - Twitter
  - Zillow.com
  - Bitcoin prices
  - DonorChoose.org (http://data.donorschoose.org/open-data/overview/)
  - …

Thank you!

Questions?