Restaurant Health Predictions via Yelp

Sean Mozarmi

Jason Mach

Jonathan Nakhla

Problem Statement

- Predict SF restaurant health scores based on yelp user reviews.

Transforming Data for Labeling

- **Ground truth**: SF food inspection board's health scores for restaurants
- We factored out time from our data:

Restaurant	Yelp Reviews (per period)	Health Score
Larry's Burgers 0	Very nice burgers but the fires ain't that gre…	80
Larry's Burgers 1	I love fries but yesterday I saw a fly….	90
Larry's Burgers 2	I prefer Bob's Burgers over on Victoria, but it…	95

Data extraction/featurization

- Scrape Yelp restaurant websites
- Featurization methods:
  - Bag of words – TFIDF
  - **NLP – Dependency parsing – noun phrases**
    - **Included bigrams and single words**

Regression

- Use regularized regression
  - Dependent variable: health scores
  - Independent variable: yelp user reviews

- We also regressed health scores on
  - Restaurant price category
  - Restaurant type (e.g. Chinese, Italian)
  - Restaurant overall user rating

Tools

- For scraping and data handling
  - Beautiful soup to parse yelp sites
  - EC2 to distribute scrape jobs (*ongoing*)
  - Pandas to manipulate data
- For regression
  - Scikit-learn for regression packages/featurization
  - Malt parser to featurize word phrases
  - TextBlob (scrapped)

Unexpected Challenges

- **Data extraction being the bottleneck**
  - Yelp rate limiting our scraping attempts or IP blocking us
- Labeling data
  - Solved by factoring out time from our data
  - Treat each inspection period as an independent data point

Preliminary Results

- Using sample of ~45 restaurants/300 data points/26,000 yelp reviews:
  - Residual sum of squares ~= 60
  - Variance ~= 0.25
  - negative features: “fried”, “pork”, “take out”
- There **is** a correlation between yelp user reviews and health scores

Preliminary Results

- No correlation between overall user restaurant rating and health score
- No correlation between # of reviews and score
- There *is strong* correlation between health score and price category/restaurant category
  - RSS ~= 50
  - Variance ~= .35

Lessons Learned / Mitigations

- Don’t underestimate the time required in data extraction phase