Yelp Me Yelp You

Yelp Dataset Challenge, Round 9

Michael Barlow, Madeleine Crouch, Khoa Le, Dalton Morrow

Project Description

- Predict business success based on subset of attributes and trends
  - Eliminate redundant and non-decisive attributes
- Find “angry reviews” with sentiment analysis; find angriest times/locations or normalize reviews of angry users to provide more accurate ratings
  - Analyze “angry” keyword appearance
- Find recurring problems with restaurants by analyzing reviewer language and ratings over time

Prior Work

The dataset was first released in March 2013 and has been updated and released with additional features sporadically since then. Prominent topics of research include:

Prior Work

The dataset was first released in March 2013 and has been updated and released with additional features sporadically since then. Prominent topics of research include:

- Review classification
- Natural language processing
- Rating effect on business revenue
- Quantifying external influences to review/rating trends
- General business analysis
- Improved recommendation models

In short, a large body of research has taken place on this dataset since its release, so addressing novel yet useful questions may be difficult.

Dataset

- Data from Yelp Dataset Challenge, Round 9 (downloaded)
    - Businesses (144,000 businesses)
    - Check-In (aggregated from 125,000 businesses)
    - User Reviews (4,100,000 reviews)
    - User Tips (947,000 tips)
    - User Data (1,000,000 users)
    - Auxiliary photos (200,000)

    ~5 GB without photos; photos are another 6.5 GB

Proposed Work

- Data cleaning
  - Remove or substitute incomplete/optional attributes
- Data Reduction
  - Remove nonessential attributes
- Data Integration
  - Combine user, business, and review datasets

Tools

- SQL Database
- Amazon Redshift (AWS Warehouse)
- Plotly/D3
- Python SciPy; NumPy
- Weka
- Microsoft Sentiment Analysis

Results Evaluation

- Evaluate prediction accuracy for success based on attributes and trends
- Correlation analysis for location/time of angry reviews; topical and service analysis to determine if many people are angry about the same thing

Businesses can determine where the dissatisfied people are and decide if satisfaction could be improved through policy or service changes. If businesses have lots of bad reviews concerning the same problem, they can address the problem. If certain services are shown to increase customer satisfaction, businesses can integrate them.