Yelp Dataset Challenge

CS 591 Data Mining, Spring 2015

By Tyler Waltze

Where to find

http://www.yelp.com/dataset_challenge

Each file is composed of a single object type, one json-object per-line.

Yelp Dataset

- 10 cities across 4 countries
- 1.6 million reviews
- 61k businesses
- 61k checkins
- 366k users
- 500k tips

Processing Dataset

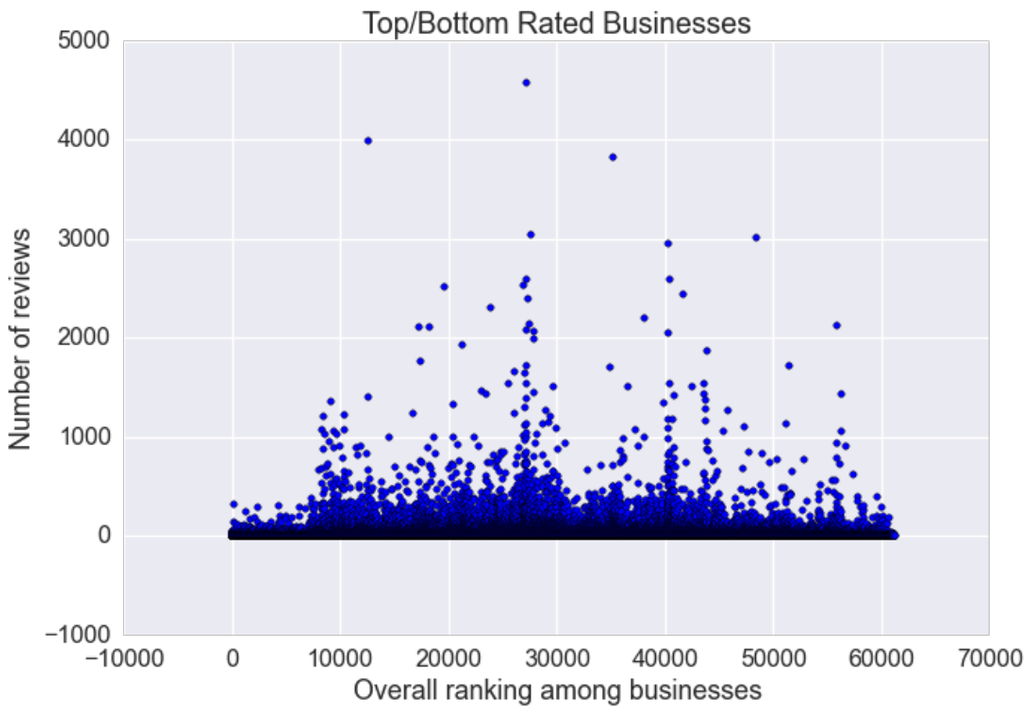
- Read in json file
- Load into Pandas data frame
- Save data frame as csv
- Separate out restaurants

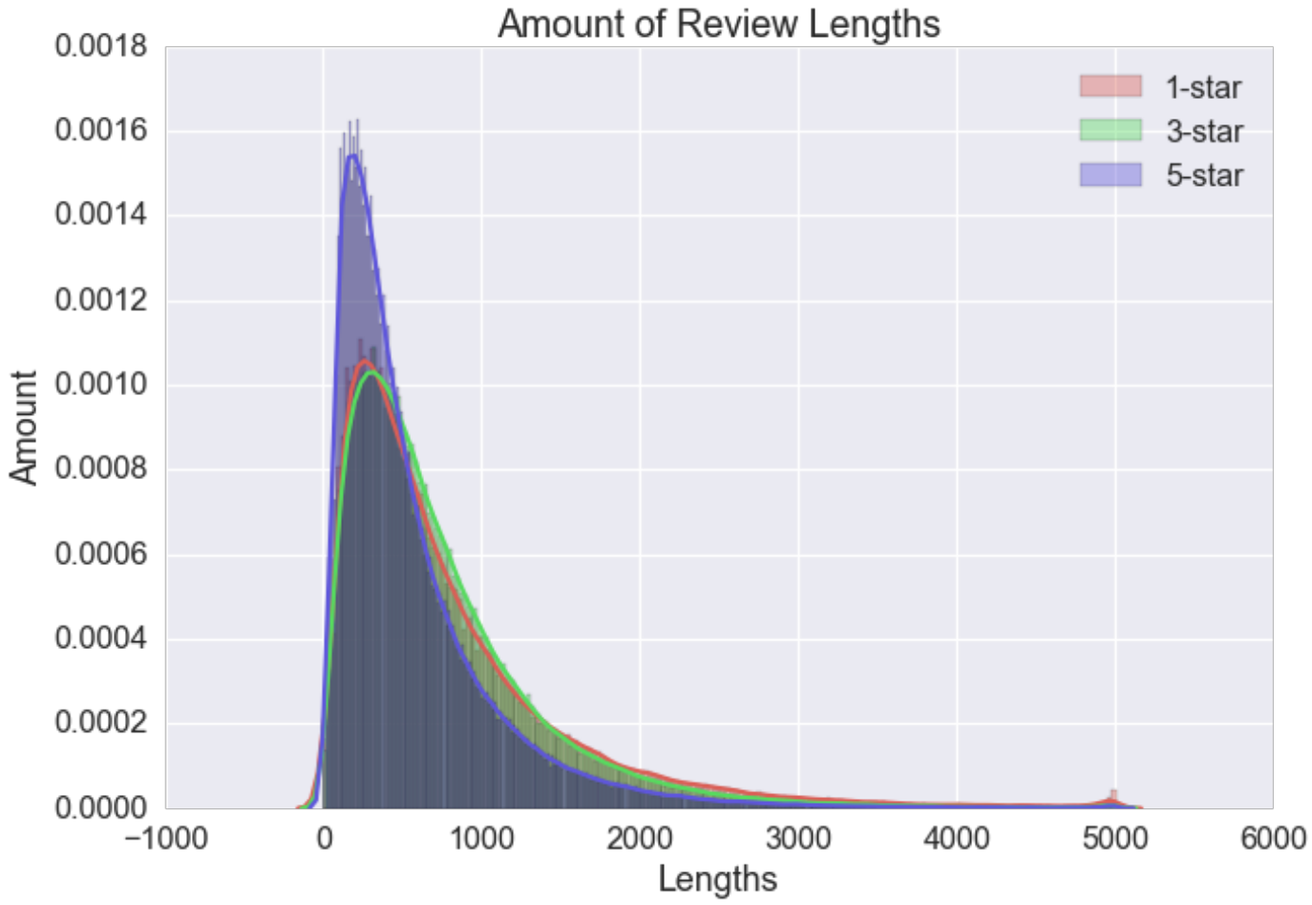
```
def convert(x):
    ''' Convert a json string to a flat python dictionary
    which can be passed into Pandas. '''
    ob = json.loads(x)
    for k, v in ob.items():
        if isinstance(v, list):
            ob[k] = ','.join(str(v))
        elif isinstance(v, dict):
            for kk, vv in v.items():
                ob['%s_%s' % (k, kk)] = vv
            del ob[k]
    return ob

for json_filename in glob(data + '*.json'):
    csv_filename = '%s.csv' % json_filename[:-5]

    print 'Converting %s to %s' % (json_filename, csv_filename)

    df = pd.DataFrame([convert(line) for line in file(json_filename)])
    df.to_csv(csv_filename, encoding='utf-8', index=False)
```

Hypotheses

1. Businesses can be broken down into consistent subcategories based on reviews
    1. Service
    2. Food
    3. etc
2. Location greatly affects a restaurant's popularity and correlates with "good" restaurants
3. Highly rated restaurants are in popular locations

Hypothesis #1 Figures



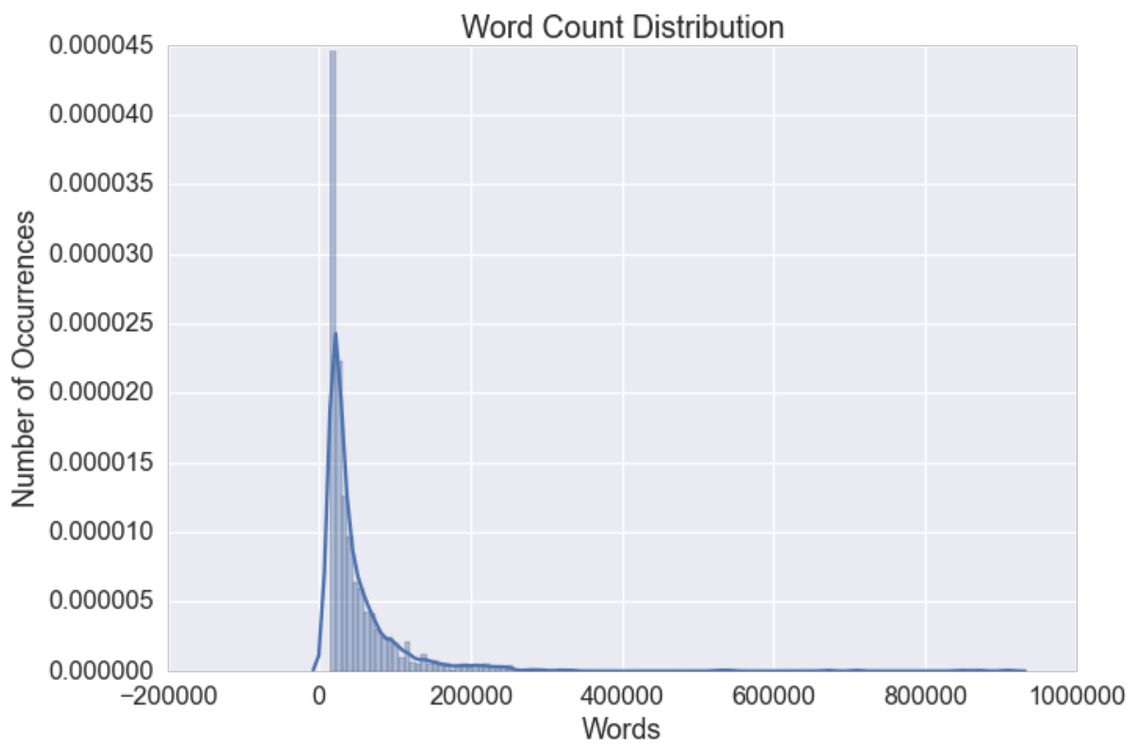


Reviews' Word Statistics

Most common words

- Food
- Service
- Place
- Good
- Great
- Nice
- Best

- 80,618,734 words
- 341,183 unique words
- 170,663 occur only once
- 318,729 occur <= 100
- 1,391 occur >= 10,000



Hypothesis #1 Result

- Difficulty attaching descriptive sentiments to specific words that fall into “consistent subcategories”

- Instead of describing a business with a few subcategories, a business can be defined by its most relevant and useful reviews

- Topic Modeling - LDA

    - Cycle through a restaurant’s reviews and find the most important words in that review.

    - Bag-of-words of all “important words” from all reviews

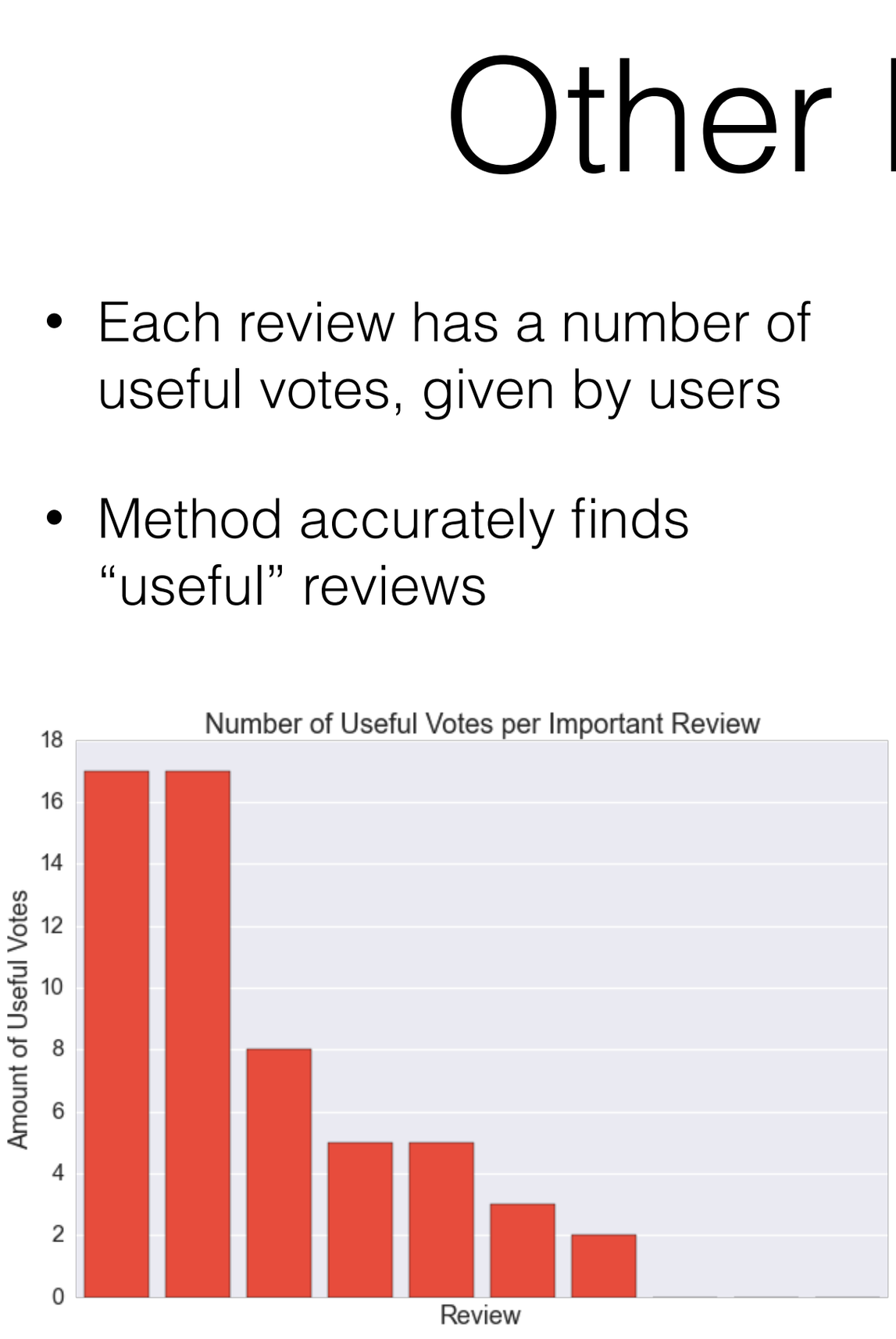
    - Important reviews are those that have the highest frequency of these “important words”

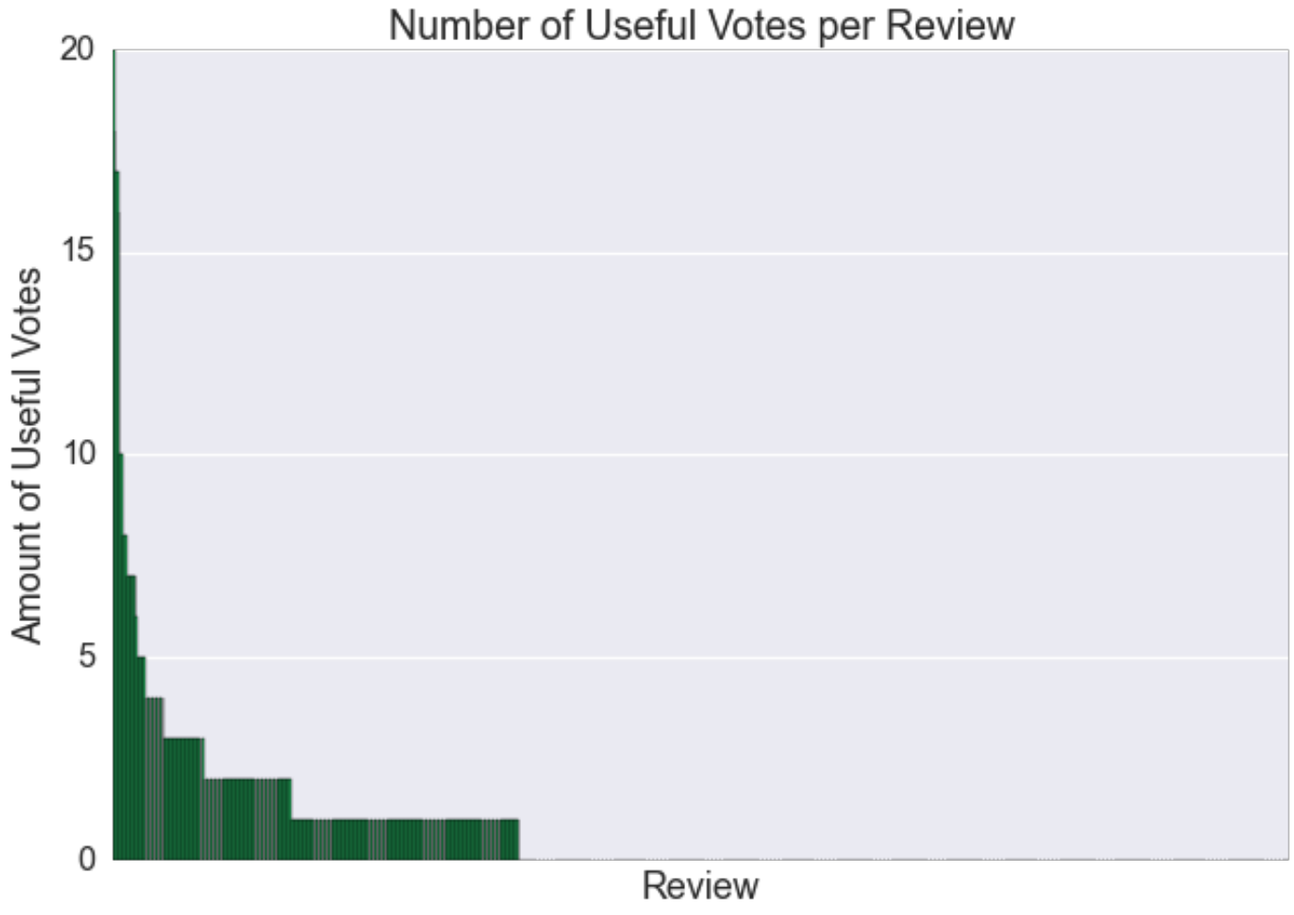
    - Predict a restaurant’s rating based on the average of these important reviews

Prediction Results

Other Findings

- Each review has a number of useful votes, given by users
- Method accurately finds “useful” reviews



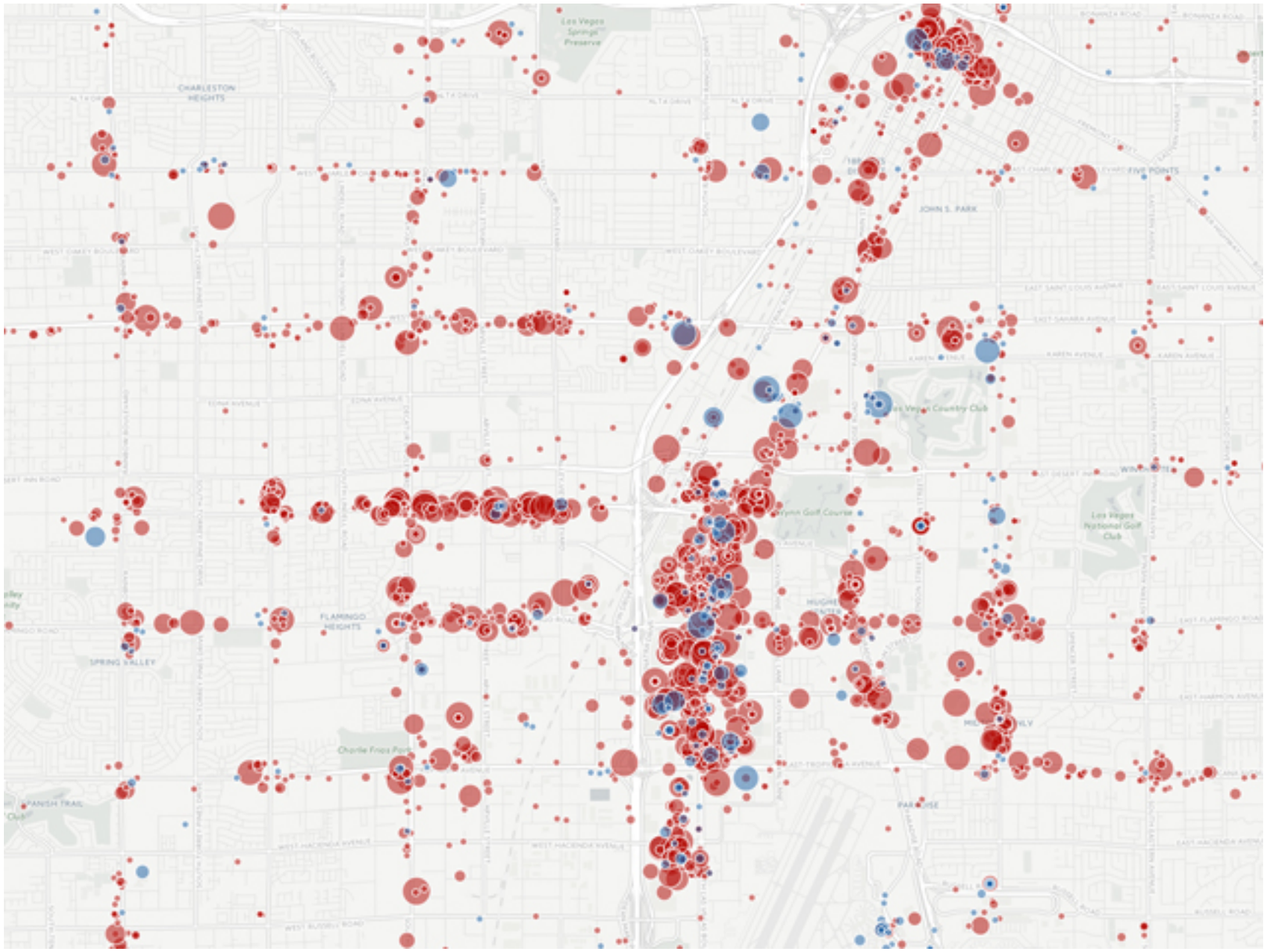


Hypotheses # 2 & 3

- Based on largely visual evidence, they are both mostly false

- There is little geographic distinction between highly and lowly rated businesses

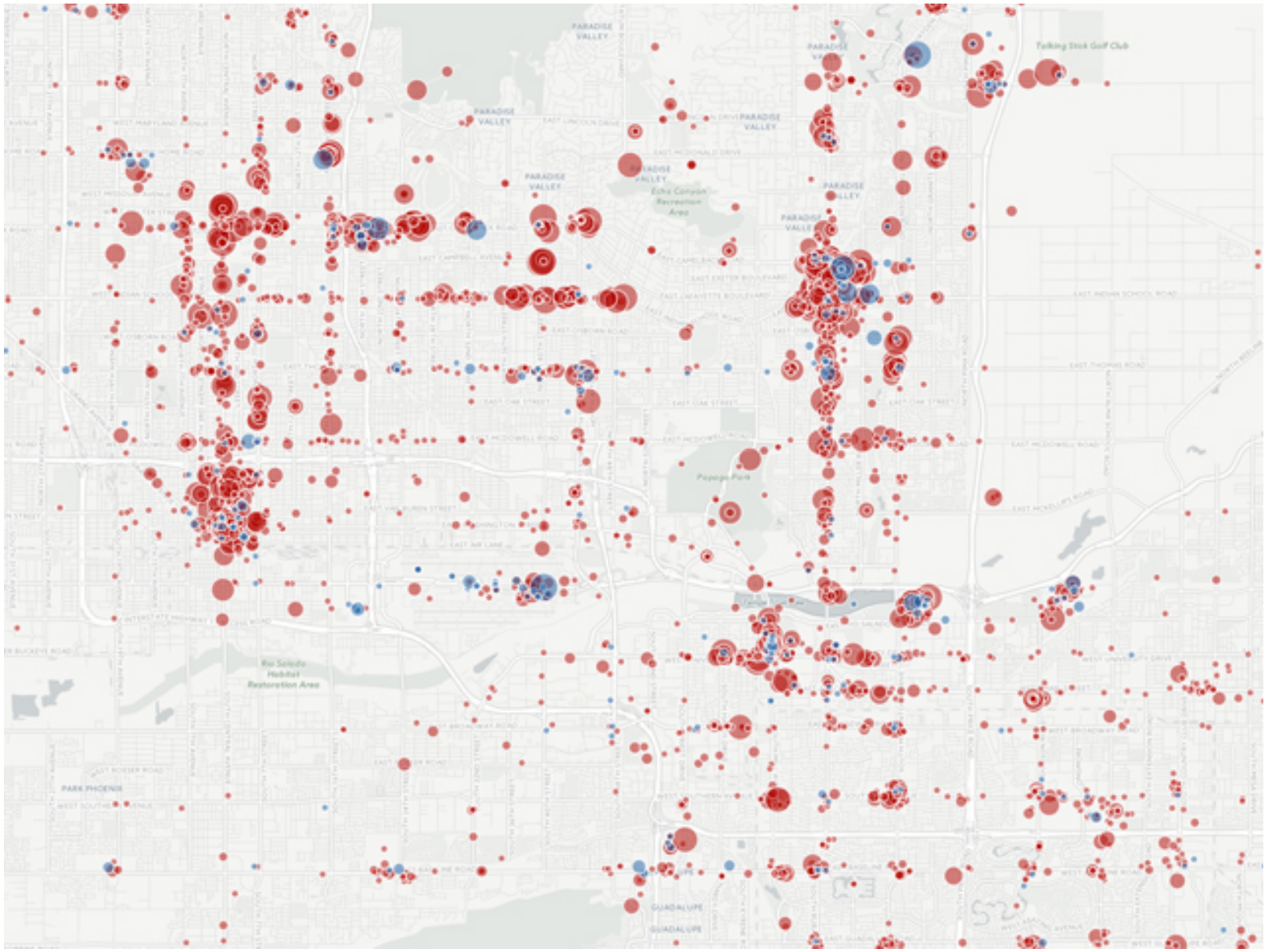
- Popular places though do tend to be clustered with other popular places



Las Vegas

http://cdb.io/1OCeGRG

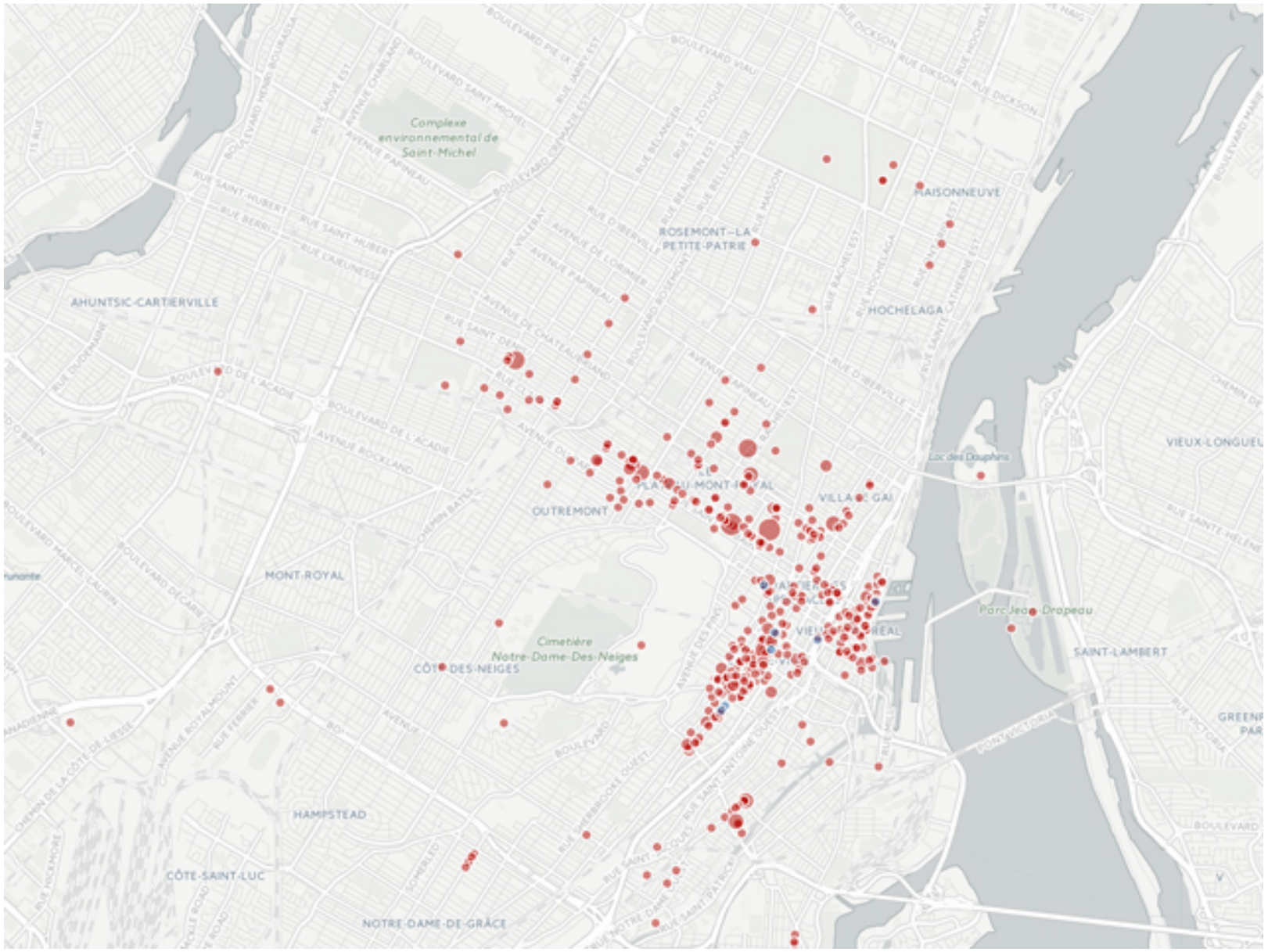
Good Bad Popularity



Scottsdale

http://cdb.io/1OCeGRG

Good Bad Popularity



Montreal

http://cdb.io/1OCeGRG

Good Bad Popularity