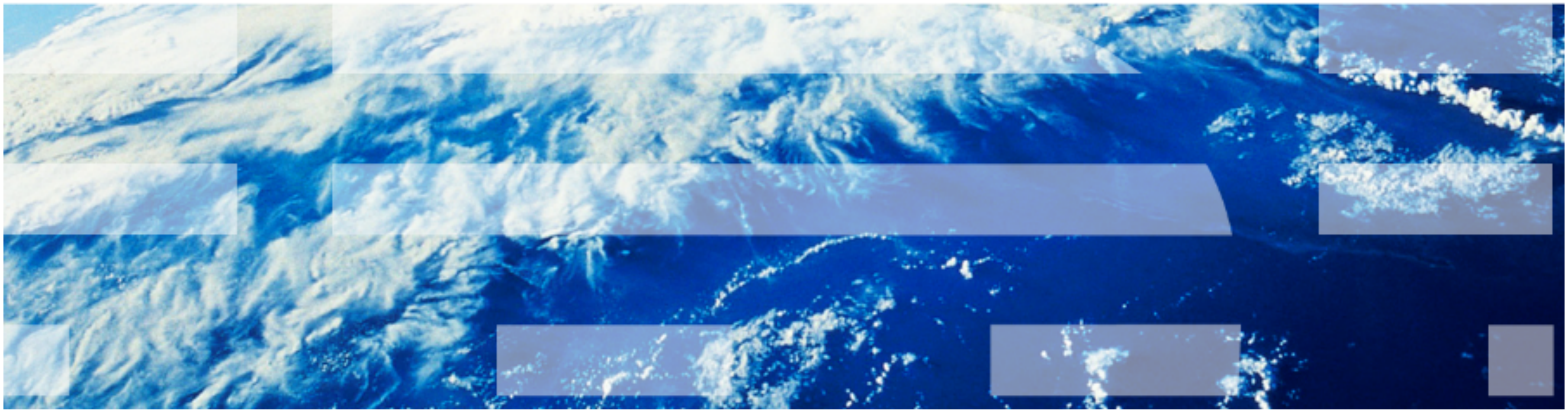
E6893 Big Data Analytics:

yelp *Reviews Exploration & Visualization*

Team Members:

Tongyun Wu tw2568
Zhi Zheng zz2406
Ming Zhou mz2591



December 15, 2016

Outline

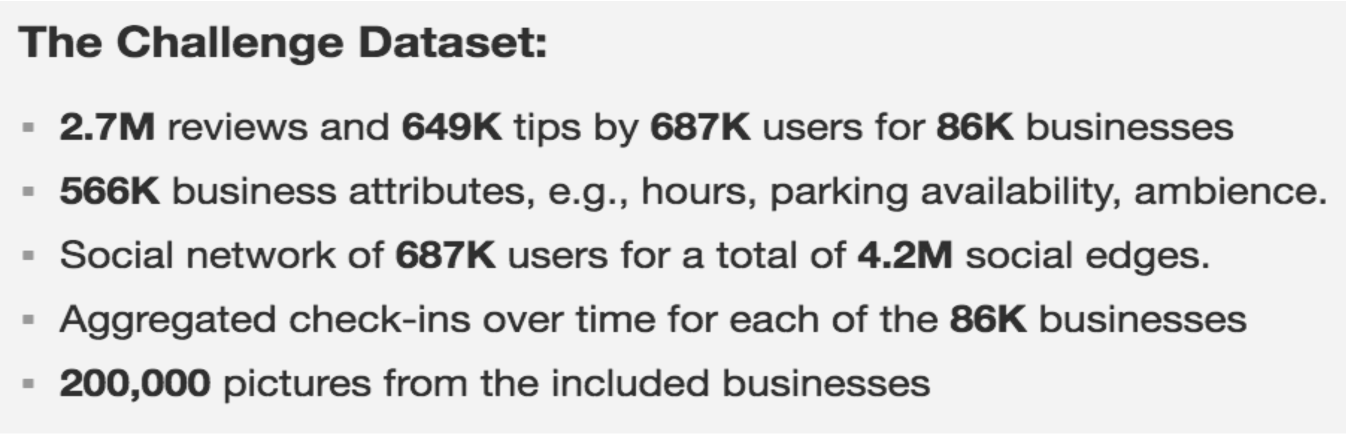
- Technologies Used
- Datasets Overview
- Challenges faced
- Data explore
- Review Stars Prediction
- Demo
- Conclusion
- Next Steps

Technologies

- **Language:** Python
- **Platform:** Apache Spark
- **Library:**
    - **ML :** NLTK, ScitKitLearn, Keras, TensorFlow
    - **Visualization:** Seaborn

DataSet OverView

- Dataset Provided by Yelp

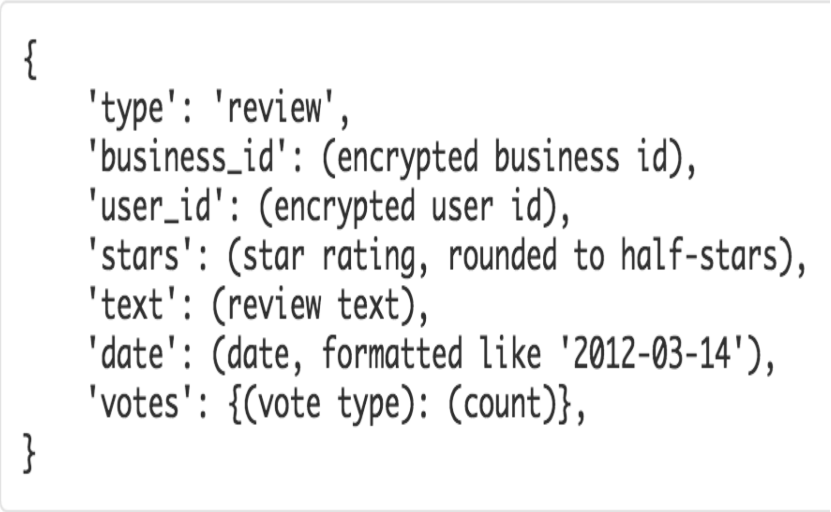


The Challenge Dataset:

- **2.7M** reviews and **649K** tips by **687K** users for **86K** businesses
- **566K** business attributes, e.g., hours, parking availability, ambience.
- Social network of **687K** users for a total of **4.2M** social edges.
- Aggregated check-ins over time for each of the **86K** businesses
- **200,000** pictures from the included businesses

- Included Five JSON File (Total 2.8G) :
  - User Information
  - Business Information
  - Tips (text)
  - Reviews
  - Check-in Details

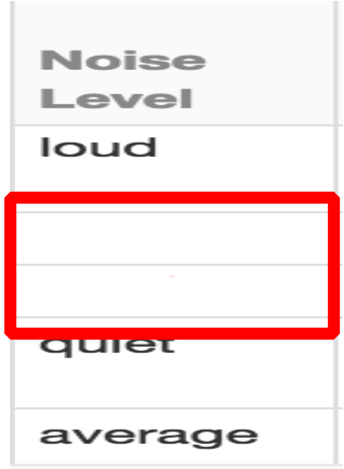
review



```
{
    'type': 'review',
    'business_id': (encrypted business id),
    'user_id': (encrypted user id),
    'stars': (star rating, rounded to half-stars),
    'text': (review text),
    'date': (date, formatted like '2012-03-14'),
    'votes': {(vote type): (count)},
}
```

Challenges / Struggles

- DataSets are not very organized:
  - Some Fields are empty: eg. Price Range, Noise Level…



Noise Level
loud
quiet
average

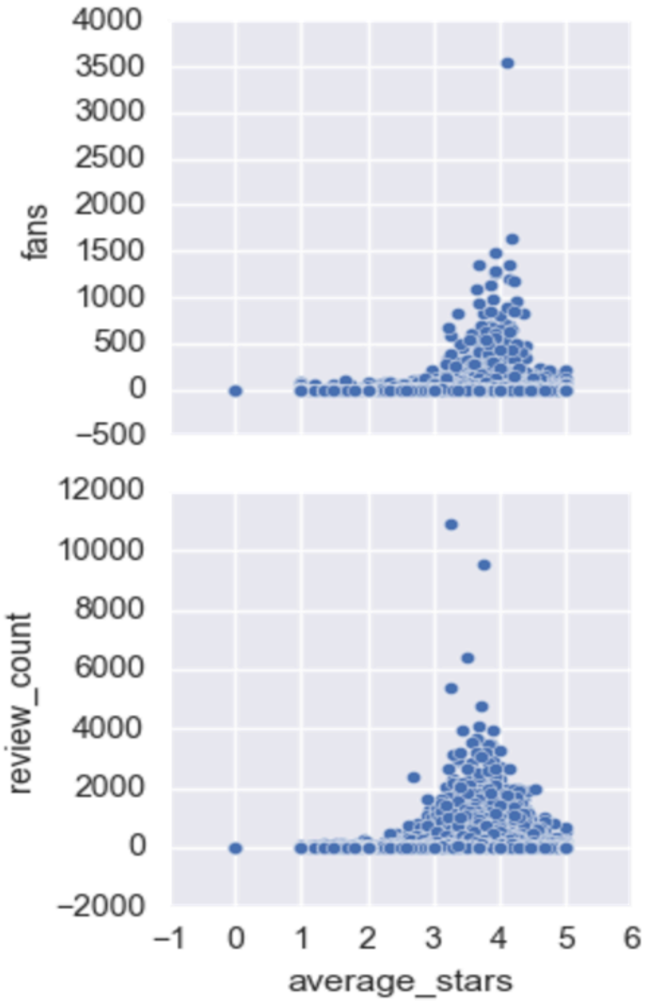
  - Some Fields are not easy to normalize: eg. WIFI: {no, free, paid, yes, …}
- Datasets Categories are not well balanced (bootstrap resampling)

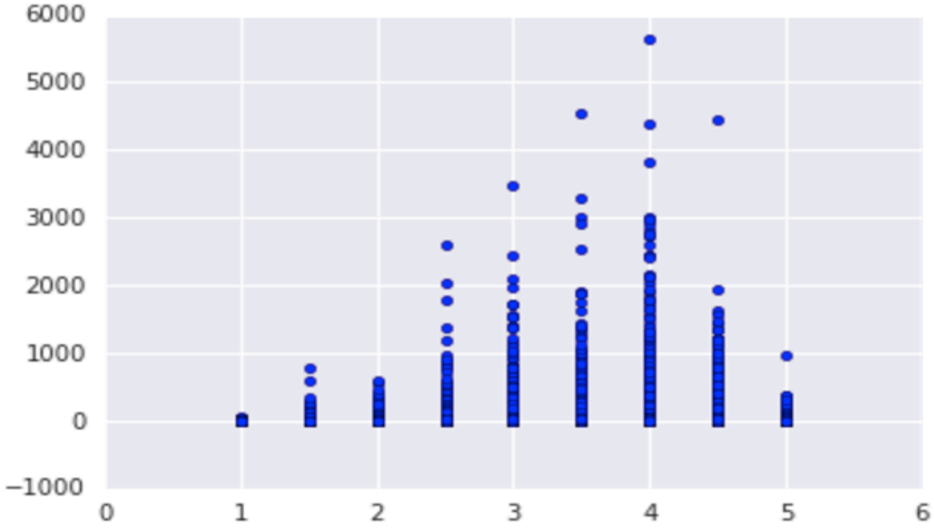
Analysis and Modeling Approach

- **Natural Language Processing**
- **Sentiment Analysis**
- **Visual Analysis**
- **Feature Importance**

Findings

- Sentiments are more obvious in the ‘tips’ text as opposed to the ‘reviews’ text.
- Most users average ratings are between 3.5 - 4.5 stars.
- Most businesses average ratings are between 3.5 - 4.5 stars.





Yelp Review Stars Prediction – CNN

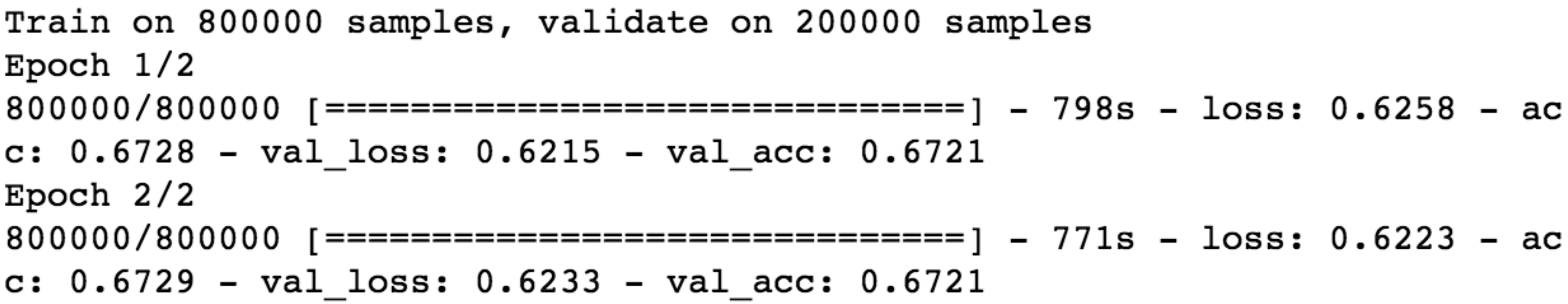
- Use word2vec as the word embedding layer

- Transform each review into a fixed length of words with each word represented by its word2vec vector
  - max number of words for each review : 50
  - max number of word features in the word2vec model : 5000

- The architecture of this model:

  Embedding layer - Dropout - Convolution1D - MaxPooling1D - Full Connected layer - Dropout - Relu activation - Sigmoid (with binary cross entropy loss)

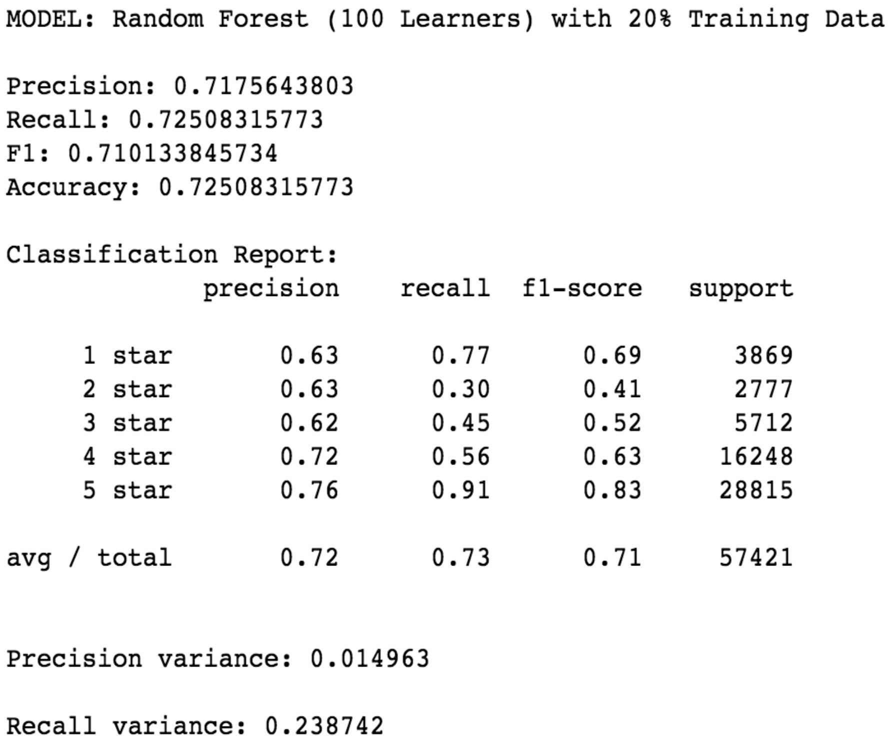
- Accuracy: training accuracy: 67.28%, validation accuracy: 67.21%



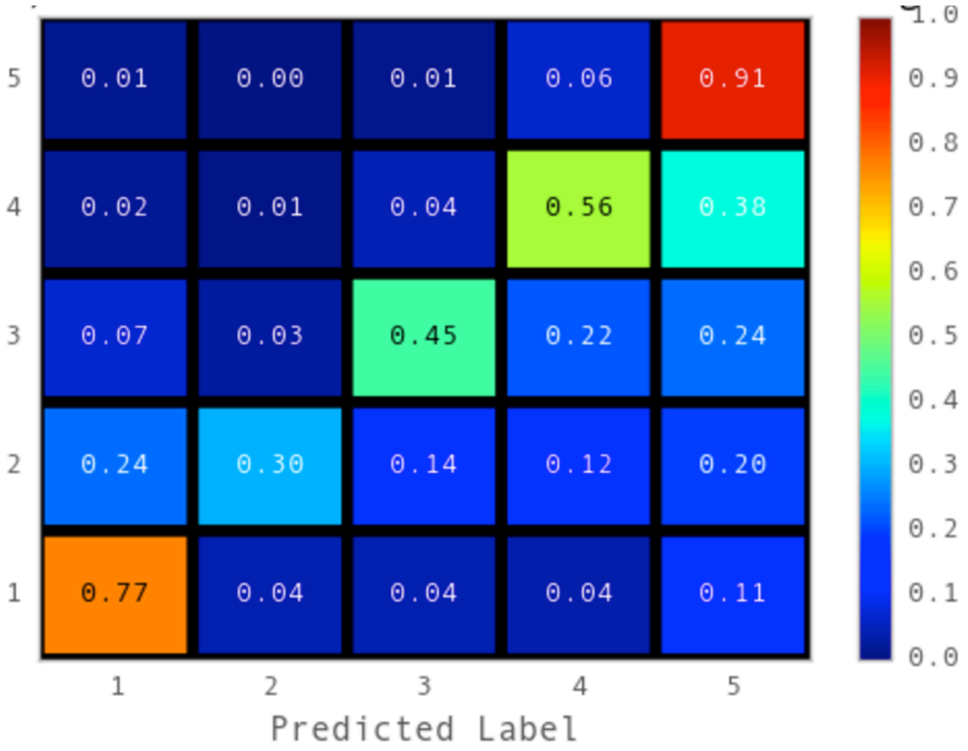
```
Train on 800000 samples, validate on 200000 samples
Epoch 1/2
800000/800000 [==============================] - 798s - loss: 0.6258 - ac
c: 0.6728 - val_loss: 0.6215 - val_acc: 0.6721
Epoch 2/2
800000/800000 [==============================] - 771s - loss: 0.6223 - ac
c: 0.6729 - val_loss: 0.6233 - val_acc: 0.6721
```

Yelp Review Stars Prediction – Traditional ML Algorithm

- Text prepocessing: remove stopwords, using bigrams model
- Ultilize sentiment analysis result into stars prediction
- Bigram Multinomial Bayes Classifier



```
MODEL: Random Forest (100 Learners) with 20% Training Data

Precision: 0.7175643803
Recall: 0.72508315773
F1: 0.710133845734
Accuracy: 0.72508315773

Classification Report:
             precision    recall  f1-score   support

     1 star       0.63      0.77      0.69      3869
     2 star       0.63      0.30      0.41      2777
     3 star       0.62      0.45      0.52      5712
     4 star       0.72      0.56      0.63     16248
     5 star       0.76      0.91      0.83     28815

avg / total       0.72      0.73      0.71     57421


Precision variance: 0.014963

Recall variance: 0.238742
```



Demos

Conclusion

- Use visualization technique to explore the data, get the overview of the structure of the data, generate wordcloud, identify important features.

- Use machine learning algorithms to perform sentiment analysis and predict the reviews stars

- Build web pages to let user explore restaurant reviews interactively.

Next Steps

- **There are various interesting questions can be explored on the yelp datasets !!!**

- **Graph Mining:** Figure out who the trend setters are? How much influence does people's social circle have on their business choices and their ratings?

- **Seasonal Trends:** Are there more reviews for sports bars on major game days and if so, could you predict that?

- **Location Mining and Urban Planning:** How much of a business' success is really just location, location, location? Do you see reviewers' behavior change when they travel?