Yelp Experience Enhancement Proposal

Gina Champion, Andrew Leonard, Hanh Cao, Phill Betts

Our Business Cases

Executive Summary: We are proposing two (2) solutions to existing Yelp Business Objectives that will assist with user interaction within the yelp site.

1. **Objective:** Increase exposure for listed businesses
   Solution: Incorporate a recommended business section post-review
   Detail: If a customer positively reviews a business within a particular category, we want to recommended other businesses they may like based off the experience at their reviewed business.

1. **Objective:** Improve Star-Rating system
   Solution: We would like to make it easier for customers to rate businesses.
   Detail: Often, we find that users will write their review, but then forget to give a star rating. We would like to automatically recommend a rating based off the written review given (and then allow the reviewer to edit if needed).

Source: https://www.yelp.com/dataset/documentation/main

Challenges: We limited the data scope to one of the most popular cities (Austin, TX) due to the cost of running BigQuery.

01

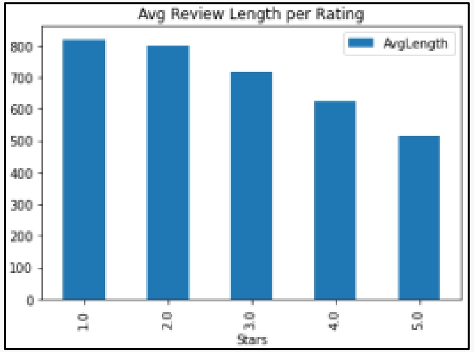
Exploratory Data Analysis

Reviews & Text Analysis

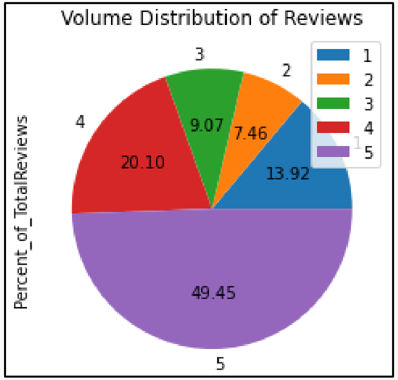
Notes:

- Reviews can be a maximum string length of 5,000
- Total of 1,425,227 Reviews and 22,416 Businesses in our Austin Data Set

The Average Length of a Review decreases the higher the business rating



Almost half of all Reviews are 5 Star Reviews



Based on the low average length and large volume of 5-star reviews, <u>we need to confirm they are not simple or useless to conduct Models on</u>

5-Star Review Analysis

Notes:

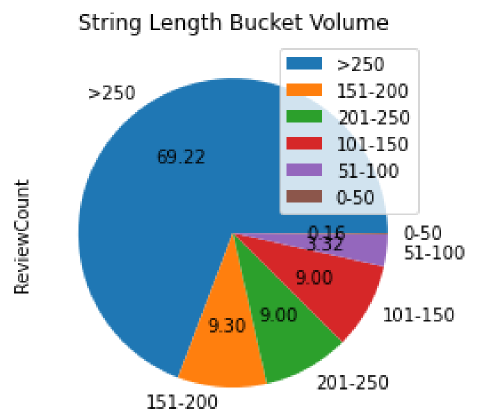
- We will estimate the average sentence string length between 75 and 100.

*https://strainindex.wordpress.com/2008/07/28/the-average-sentence-length/#:~:text=But%20sentences%20have%20three%20units,and%2075%2D100%20characters.%E2%80%9D

Review Length Volume per String Bucket

Length_Bucket	ReviewCount
>250	487924
151-200	65521
201-250	63451
101-150	63426
51-100	23393
0-50	1125

Review Length Volume Percentage per String Bucket



We can assume most 5-star reviews have substantive language to apply models on

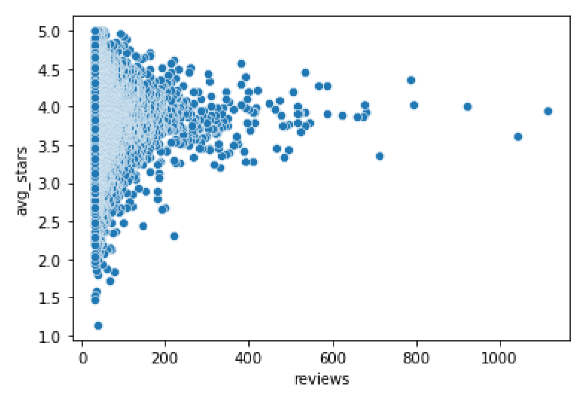
Review Behavior Analysis

We want to confirm that users are appropriately rating businesses

Number of Business &
Average Number of Reviews per Business

stars	Num_of_Businesses	Avg_Num_Reviews
1.0	198334	259
2.0	106323	429
5.0	704840	431
3.0	129322	526
4.0	286408	561

Average Star Rating per Number of User Reviews

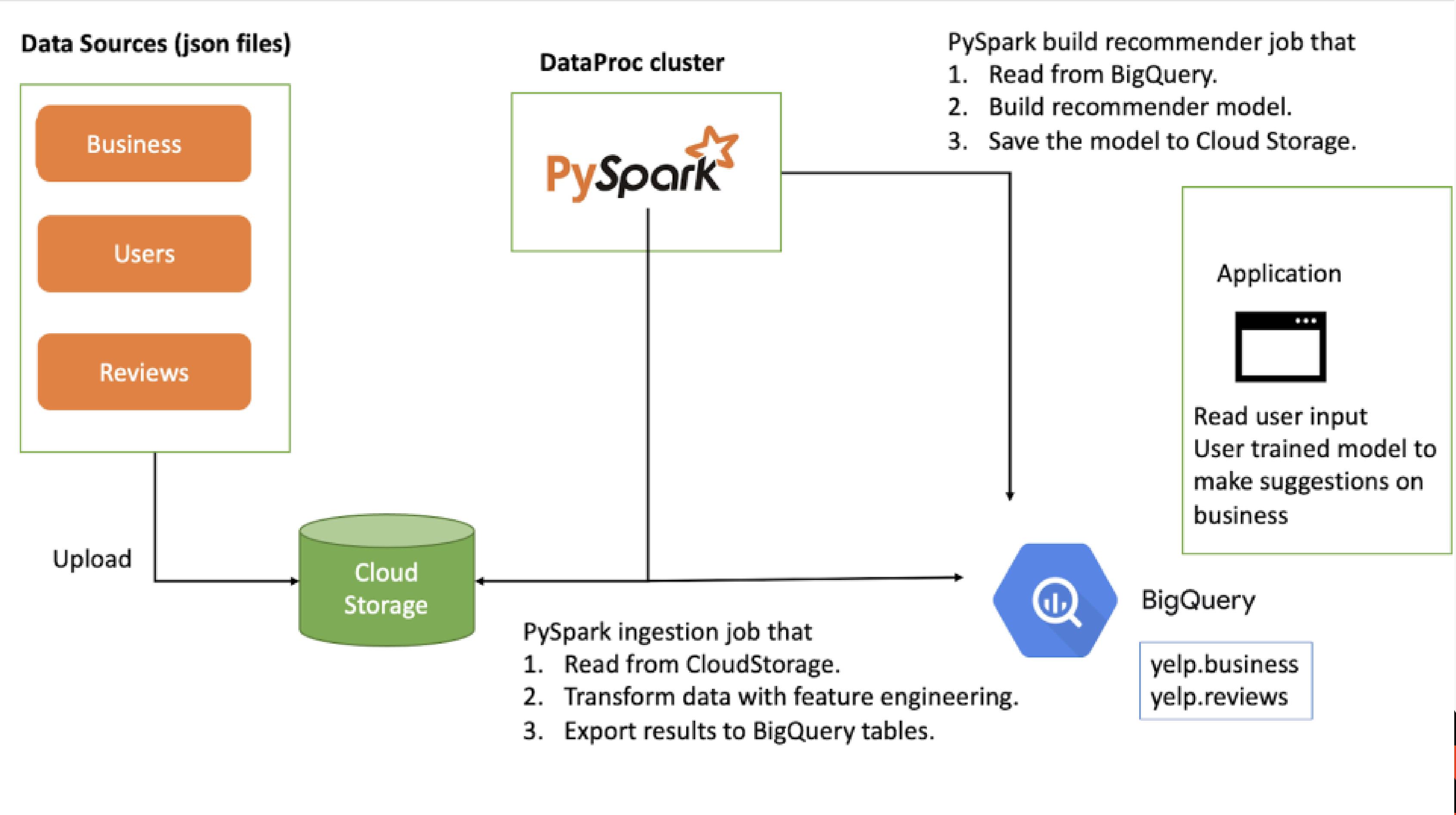


We can confirm there are no users who "spam" 5-star or 1-star reviews for any particular business.

02

Infrastructure & Feature Engineering

Google Cloud Platform



Ingestion job

- Job can be run with optional command arguments:
  - Save dataframes in Parquet compressed to Cloud Storage
  - Limit business data to particular city
  - Limit reviews starting from particular year

(Due to the cost of BigQuery, we ran the ingestion job with limit to city="Austin" and year>="2018")

Feature Engineering:

- Split "categories" column into 5 distinct features

Categories
Greek, Seafood, Gluten-Free

Cuisine	SpecialtyFood	IsFastFood	IsCafes	IsDiet
Greek	Seafood	No	No	Gluten-Free

- Extract "attributes" column and only keep those attributes with less than 30% null values. Transform those attributes, for example:
  - "attributes.Alcohol" (string) to "has Alcohol" (Boolean)

attributes.Ambience
{ divvy: False, romantic: True }

Ambience
romantic

Model Building job (content-based)

- Assemble vector for each business with one-hot encoding for categorical columns
- Calculate business similarity using cosine similarity between business vectors.
- Example of usage:
  - Given a list of business ids, the recommender returns 5 most similar businesses:
  - Given user_id, the recommender queries 5 business_id that user rated >= 4.0 before, then recommends the similar businesses

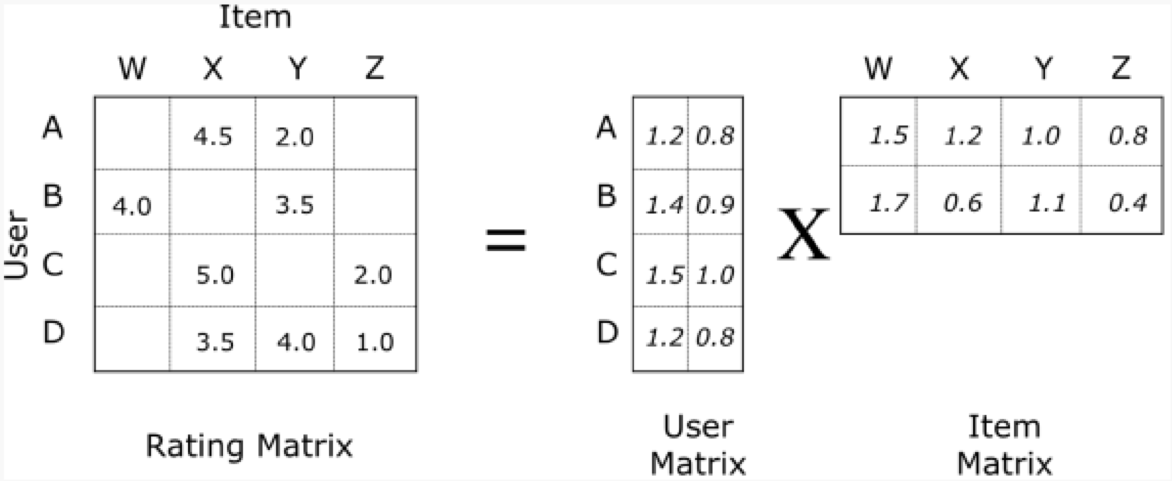
```
Input businesses:
              business_id                             name  stars  \
0  tqHZ-qFUH34Juvw_IQqWvA  Tortilleria El Taquito Marisquero    4.0

  cat_Cuisines cat_SpecialtyFood  cat_IsFastFood  cat_IsCafes cat_Diet  \
0      Mexican           Seafood               0            0       NA
```

```
         input_business_id     similar_business_id  cosine_similarity
0  tqHZ-qFUH34Juvw_IQqWvA  3jiQKBE8N2gkoOtv1wiScg           1.000000
1  tqHZ-qFUH34Juvw_IQqWvA  CcKKrDq-HdOAhDHAEDjRIQ           1.000000
2  tqHZ-qFUH34Juvw_IQqWvA  zM98ZSIJyuBQabyYornLpw           0.970143
3  tqHZ-qFUH34Juvw_IQqWvA  LHxDcsscgG-POCxFnxMrsg           0.970143
4  tqHZ-qFUH34Juvw_IQqWvA  4cQLu7PpGwRek_9q32Jp_A           0.937500

              business_id                                       name
0  3jiQKBE8N2gkoOtv1wiScg              La Catedral Del Marisco #2
1  CcKKrDq-HdOAhDHAEDjRIQ             La Feria Mexican Restaurant
2  zM98ZSIJyuBQabyYornLpw  Casa Chapala Mexican Cuisine & Tequila Bar
3  LHxDcsscgG-POCxFnxMrsg                             La Fantabulous
4  4cQLu7PpGwRek_9q32Jp_A                               Seafood Shack

  postal_code  stars cat_Cuisines cat_SpecialtyFood  cat_IsFastFood
0       78741    2.5      Mexican           Seafood               0
1       78729    3.5      Mexican           Seafood               0
2       78758    4.0      Mexican           Seafood               0
3       78735    3.5      Mexican           Seafood               0
4       78734    4.0           NA           Seafood               0
```

Recommender System 03

ALS Collaborative Filtering Model

- User business star rating (explicit feedback) to train ALS recommender model
- Hyperparameter tuning: max iterations (15), rank (10), regularization parameter (.45) to minimize RMSE
    - RMSE of 1.38 on test set
- Low number of reviews for businesses problematic



Example Predictions

stars	prediction
5	3.6761065
4	3.7837687
4	3.7837687
5	4.079739
4	3.6304028
5	4.1358953
5	3.3987837
5	4.591448
5	4.117466
5	3.9104998

Measure ALS Recommendation Relevance

- Use word2vec to create vectors on business categories column

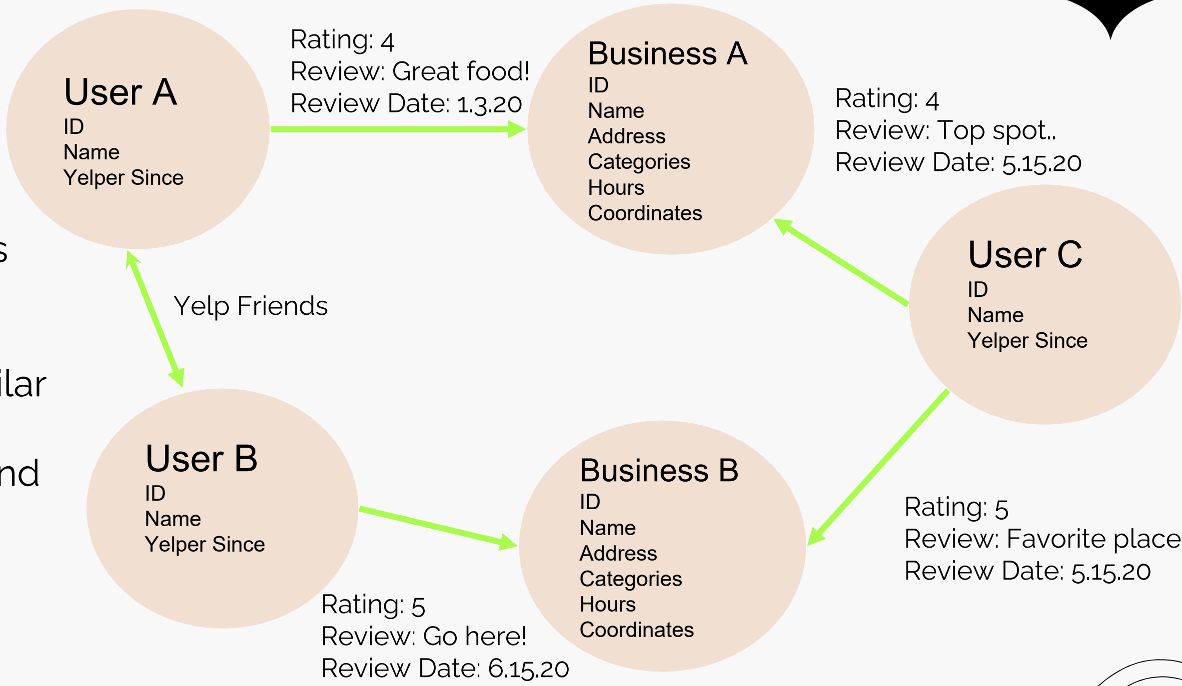
Business	Categories
1	'Breakfast & Brunch, Tacos, Mexican, Food Trucks, Food, Restaurants'
2	'Tacos, Food Stands, Hot Dogs, Food Trucks, Mexican, Yelp Events, Food, Local Flavor, Restaurants'
3	'Smog Check Stations, Oil Change Stations, Auto Repair, Auto Parts & Supplies, Automotive, Commercial Truck Repair, Transmission Repair'

Cosine similarity 1 to 2: .95
Cosine similarity 1 to 3: -.09

- Measure relevance of recommendations to user based on categories of previous businesses reviewed
- Cosine similarity of recommended business categories to user's history fell between .7 and .93 in test cases
- Recommendation results are highly dependent on data sparsity

Recommender - Graph Computing

- Nodes
  - Users
  - Businesses
- Edges
  - User reviews business
  - User friends with user
- Recommender Strategies
  - Connect users by similar business ratings
  - Connect users via friend edge, find new businesses through friend link



Graph Motif for Recommender

Motif

1. Takes an input user and finds all businesses reviewed
2. Returns all users who have also reviewed businesses of input user
3. Finds all businesses reviewed by the new users

Recommender Logic

1. Filter for same or similar business rating between an input user and other users
2. Filter for businesses input user has not reviewed and other users have rated highly

(input_user)-[e]->(input_user_business); (new_user)-[e2]->(input_user_business); (new_user)-[e3]->(new_user_business)

input_user	stars	name	stars	new_user	stars	name
[Lu4-NKrpJbSBpUcZ...	5.0	Texas Roadhouse	4.0	[3sI5kFZp81KWohkH...	4.0	Vespaio Ristorante
[Lu4-NKrpJbSBpUcZ...	4.0	P. Terry's Burger...	3.0	[1ya2z81pqWVGD3u4...	5.0	Perry's Steakhous...
[Lu4-NKrpJbSBpUcZ...	5.0	Rudy's "Country S...	4.0	[I2AM0Xh5clFA3iyF...	1.0	Ramen Tatsu-Ya
[Lu4-NKrpJbSBpUcZ...	1.0	Eurasia Sushi Bar...	5.0	[T6K1U65wS7NtR1QX...	5.0	Ramen Tatsu-Ya
[Lu4-NKrpJbSBpUcZ...	3.0	Eurasia Sushi Bar...	5.0	[Kj_MYdysEwQORXOG...	1.0	Sandy's Hamburgers
[Lu4-NKrpJbSBpUcZ...	1.0	Sonic Drive-In	3.0	[q3cxC9tv3bmPE74i...	3.0	Bert's BBQ
[Lu4-NKrpJbSBpUcZ...	3.0	Alamo Drafthouse ...	5.0	[t6eNIzThY2QCarVZ...	5.0	PostalAnnex+
[Lu4-NKrpJbSBpUcZ...	5.0	Target	4.0	[q3cxC9tv3bmPE74i...	5.0	St Andrew's Episc...
[Lu4-NKrpJbSBpUcZ...	3.0	Pinthouse Pizza	4.0	[hBRPfyanAA-0xxlv...	4.0	Cooper's Old Time...
[Lu4-NKrpJbSBpUcZ...	4.0	Texas Roadhouse	4.0	[GLjWC3oPZJ1BUYUw...	3.0	Home Slice Pizza
[Lu4-NKrpJbSBpUcZ...	3.0	Pinthouse Pizza	5.0	[JaqcCU3nxReTW2cB...	4.0	Cosmic Coffee + B...
[Lu4-NKrpJbSBpUcZ...	1.0	Eurasia Sushi Bar...	5.0	[iK3rXDUZCdc7BJ5m...	5.0	BookPeople
[Lu4-NKrpJbSBpUcZ...	4.0	Pieous	4.0	[_tm5XdVoIlfH5PCe...	4.0	China's Family Re...
[Lu4-NKrpJbSBpUcZ...	5.0	Maudie's Hacienda	5.0	[IA6q H9QlY yXsT9...	5.0	The Belmont
[Lu4-NKrpJbSBpUcZ...	3.0	Taco Ranch	3.0	[s9jbQyCn2p_SDc7o...	5.0	Moonshine Patio B...
[Lu4-NKrpJbSBpUcZ...	3.0	Pinthouse Pizza	5.0	[jGRAfOXCqGPny0U2...	4.0	The Grove Wine Ba...
[Lu4-NKrpJbSBpUcZ...	3.0	Alamo Drafthouse ...	5.0	[oRO3H4BW-IvEi9GS...	5.0	Starbucks
[Lu4-NKrpJbSBpUcZ...	3.0	Taco Ranch	3.0	[_L0vlwBOSdNHs3vh...	4.0	Starbucks
[Lu4-NKrpJbSBpUcZ...	5.0	Me Con Bistro	3.0	[HMUnp55Q8_vxIEPl...	1.0	Burger King
[Lu4-NKrpJbSBpUcZ...	4.0	Me Con Bistro	3.0	[HMUnp55Q8_vxIEPl...	4.0	Otherside Deli an...

only showing top 20 rows

Graph Recommendation Relevance

Order recommendations by relevance:

1. User/business location
   a. Use users geolocation to sort recommendations by distance to user
2. Business category relevance
   a. Use word2vec to create word embeddings on business categories. Sort recommendations based on category similarity
   b. Easily repeatable on a user's n last reviews to generate highly relevant recommendations

Example - Last reviewed business categories:
Thai, Restaurants, Food, Food Trucks

Top 5 recommendations:

	business_id	business_name	categories	similarity
1246	XZb-K_pP8Roz8WIG2hPFEg	Tuk Tuk Thai Cafe	Thai, Restaurants, Food	0.937396
2960	vuOfLg269Rr4-moMAidLqg	Veracruz All Natural	Food, Restaurants, Food Trucks, Mexican	0.842530
2582	tHv6_4DKOV8sZnlvTrCN9Q	Al Pastor	Food Trucks, Restaurants, Mexican, Food	0.842530
353	btqvmsmX5Phgr1A0jH6j0w	LUV Thai Cuisine	Restaurants, Thai	0.842305
362	xFgliLmJVCKqKX8Ra_ZNQQ	Chi'Lantro	Korean, Restaurants, Food, Asian Fusion, Barbe...	0.810751

04 Natural Language Processing

Spark NLP:
Aspect Based Sentiment Analysis for Restaurant Reviews

- Automatically detects positive, negative and neutral aspects about restaurants from user reviews
- Helps identify which exact phrases relate to the sentiment identified in the review

```
-RECORD 0-----------------------------------------------------------------------------------------
 ner_chunk | [{chunk, 7, 10, food, {entity -> POS, sentence -> 0, chunk -> 0, confidence -> 0.9998}, []},
 stars     | 4.0
-RECORD 1-----------------------------------------------------------------------------------------
 ner_chunk | [{chunk, 0, 5, Drinks, {entity -> NEG, sentence -> 0, chunk -> 0, confidence -> 0.9975}, []}
 stars     | 1.0
-RECORD 2-----------------------------------------------------------------------------------------
 ner_chunk | [{chunk, 13, 19, service, {entity -> POS, sentence -> 0, chunk -> 0, confidence -> 1.0}, []}
 stars     | 5.0
-RECORD 3-----------------------------------------------------------------------------------------
 ner_chunk | [{chunk, 4, 7, food, {entity -> POS, sentence -> 0, chunk -> 0, confidence -> 0.9994}, []},
 stars     | 5.0
-RECORD 4-----------------------------------------------------------------------------------------
 ner_chunk | [{chunk, 0, 3, Wing, {entity -> POS, sentence -> 0, chunk -> 0, confidence -> 0.5789}, []},
 stars     | 4.0
-RECORD 5-----------------------------------------------------------------------------------------
 ner_chunk | [{chunk, 38, 42, Wings, {entity -> NEG, sentence -> 1, chunk -> 0, confidence -> 0.9557}, []
 stars     | 1.0
-RECORD 6-----------------------------------------------------------------------------------------
 ner_chunk | [{chunk, 247, 252, server, {entity -> NEG, sentence -> 4, chunk -> 0, confidence -> 0.6179},
 stars     | 4.0
```

```
+-----------------+---------+
|chunk            |ner_label|
+-----------------+---------+
|food             |POS      |
|service          |POS      |
|waitress         |POS      |
|haha ladies      |POS      |
|Drinks           |NEG      |
|ribs             |NEG      |
|mcdonalds        |NEG      |
|wings            |NEG      |
|tables           |NEG      |
|service          |POS      |
|food             |POS      |
|bartender        |POS      |
|Gigi             |POS      |
|food             |POS      |
|portions         |POS      |
|server           |POS      |
|Wing             |POS      |
|tables           |NEG      |
|chairs           |NEG      |
|Wings wings wings|NEG      |
+-----------------+---------+
```

NLP ML Process

Text Preparation

- Tokenize
- Remove Stop Words
- Indexer

NLP of Reviews

- TF-IDF
- Word2Vec
- CountVectorizer

Split Data into Train/Test

- 80/20 split

Train Predictive Model

- Logistic Regression
- Naive Bayes
- Linear SVM

Evaluate Model on Test Data

- Accuracy of Prediction of Star Rating

Accuracy of Results

	Logistic Regression	Naive Bayes	Linear SVM
TF-IDF	61%	60%	53%
Word2Vec	36%	--	37%
CountVectorizer	30%	30%	2%

05 Wrap Up

Recommendations for Future Work

Updates based on Project Experience

- **Issue**: It's time consuming to run content-based recommender giving user_id, which is not efficient to run in production.
    - **Solution:** Run recommender beforehand and save top n results per user in databases.
- **Issue**: Number of times the business was reviewed had an impact on ALS results
    - **Solution**: Reduce impact of Number of Reviews
- **Issue**: Graph network recommender will not perform well in real time
    - **Solution**: Additional dataset preprocessing
- **Issue**: Lengthy and complicated review text impacted accuracy scores for prediction
    - **Solution**: Additional text cleaning to optimize accuracy of results
- **Issue:** Less than ideal accuracy scores for predicting star rating
    - **Solution:** Try using different Spark NLP pre-trained models to see if different NLP models would yield better accuracy of predictions

Q & A