Yelp Business Data Analysis: Feature Selection, Ratings Prediction, Sentiment Analysis

Nikhil Patil

Problem Description

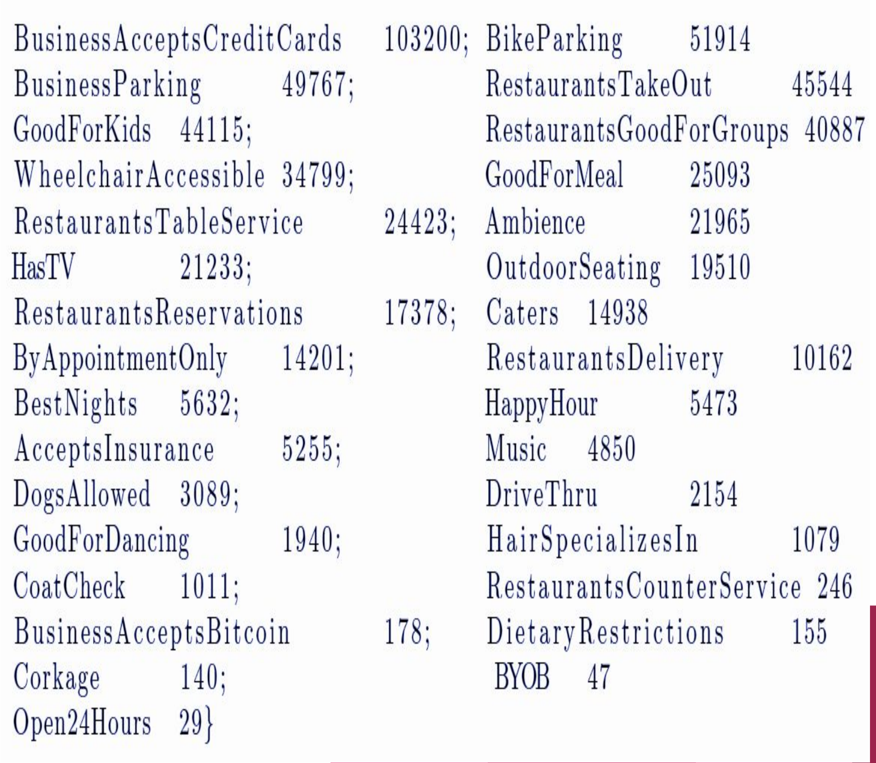
- Feature Selection
- Ratings Prediction
- Sentiment Analysis
- Geospatial Visualization of Ratings

Dataset Description:

- Sub-datasets: business, user review, check-in. Total size: 5 GB.
- Business: 144K businesses with stars, attributes, categories
- E.g. parking availability, happy hour, drive thru, restaurants table service etc.
- Review: 4 M user comments, useful votes
- Check-in: check-in counts
- Challenges:
  - Data extraction
  - Integration of sub-datasets
  - Categorical features: quantification of True or False

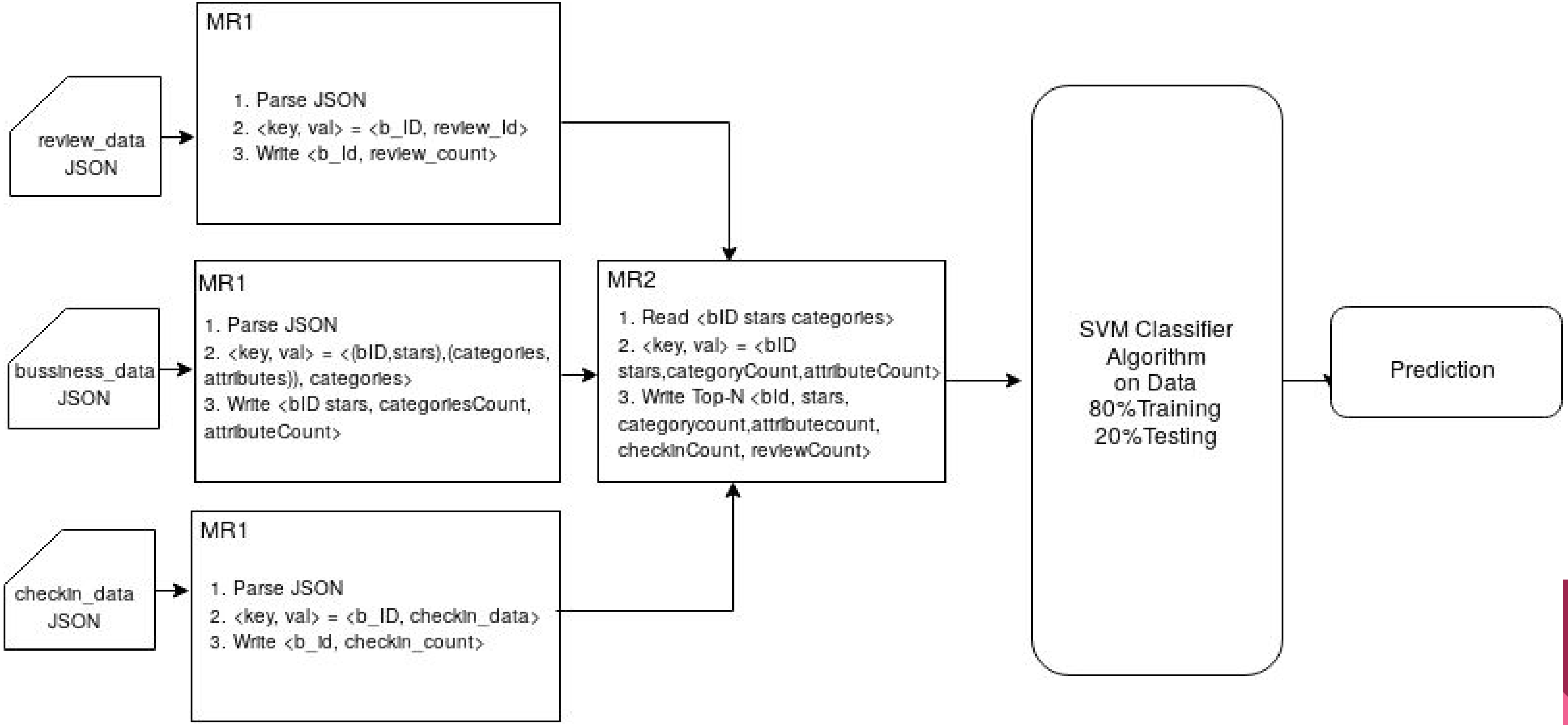
Feature Selection

- ❖ Map Reduce Jobs.
- ❖ Frequently occurred attributes, categories in high star rated businesses: Suggestion to new businesses.
- ❖ User review count, available attributes count, available categories count, total check in counts and stars for each business.
- ❖ Machine learning approach to predict stars, based on these features.



BusinessAcceptsCreditCards	103200;	BikeParking	51914
BusinessParking	49767;	RestaurantsTakeOut	45544
GoodForKids	44115;	RestaurantsGoodForGroups	40887
WheelchairAccessible	34799;	GoodForMeal	25093
RestaurantsTableService	24423;	Ambience	21965
HasTV	21233;	OutdoorSeating	19510
RestaurantsReservations	17378;	Caters	14938
ByAppointmentOnly	14201;	RestaurantsDelivery	10162
BestNights	5632;	HappyHour	5473
AcceptsInsurance	5255;	Music	4850
DogsAllowed	3089;	DriveThru	2154
GoodForDancing	1940;	HairSpecializesIn	1079
CoatCheck	1011;	RestaurantsCounterService	246
BusinessAcceptsBitcoin	178;	DietaryRestrictions	155
Corkage	140;	BYOB	47
Open24Hours	29}		

Architecture: Feature Selection & Ratings Prediction



review_data
JSON

MR1

1. Parse JSON
2. <key, val> = <b_ID, review_Id>
3. Write <b_Id, review_count>

bussiness_data
JSON

MR1

1. Parse JSON
2. <key, val> = <(bID,stars),(categories, attributes)), categories>
3. Write <bID stars, categoriesCount, attributeCount>

checkin_data
JSON

MR1

1. Parse JSON
2. <key, val> = <b_ID, checkin_data>
3. Write <b_Id, checkin_count>

MR2

1. Read <bID stars categories>
2. <key, val> = <bID stars,categoryCount,attributeCount>
3. Write Top-N <bId, stars, categorycount,attributecount, checkinCount, reviewCount>

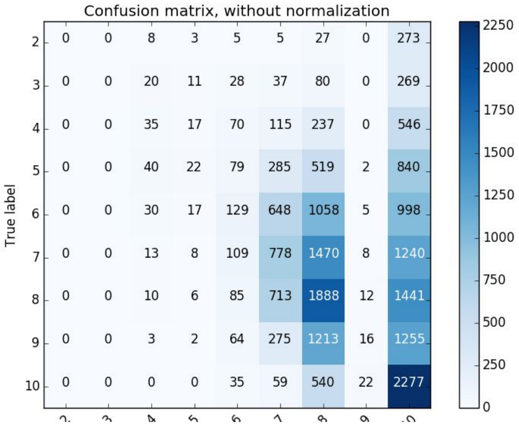
SVM Classifier
Algorithm
on Data
80%Training
20%Testing

Prediction

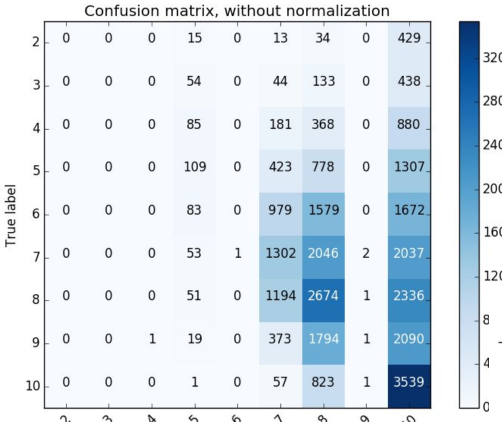
Ratings Prediction using Classification Algorithms

- Feature Spaces (X_train, test):
  - Count-based (4): [business_id, review count, attribute count, category count, checkin count]
  - Categorical (1 or 0) (31): [BusinessParking, HasTV, BYOB, etc.]
- Labels (y_train, test): Star ratings. [0,0.5,1,1.5,2,2.5,3,3.5,4,4.5,5] -> [0,1,2,3,4,5,6,7,8,9,10]
- Training Set: 75% (~110K); Test Set: 25% (~30K)
- Algorithms:
  - Regression: linear regression
  - Classification: SVM_linear, SVM_rbf, SGD

Evaluation of Analysis:

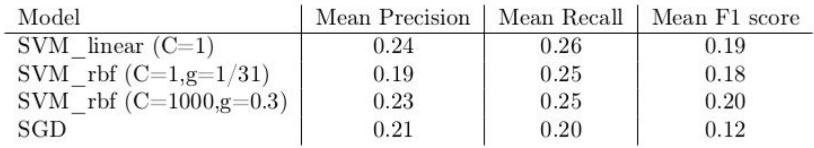


	2	3	4	5	6	7	8	9	10
2	0	0	8	3	5	5	27	0	273
3	0	0	20	11	28	37	80	0	269
4	0	0	35	17	70	115	237	0	546
5	0	0	40	22	79	285	519	2	840
6	0	0	30	17	129	648	1058	5	998
7	0	0	13	8	109	778	1470	8	1240
8	0	0	10	6	85	713	1888	12	1441
9	0	0	3	2	64	275	1213	16	1255
10	0	0	0	0	35	59	540	22	2277



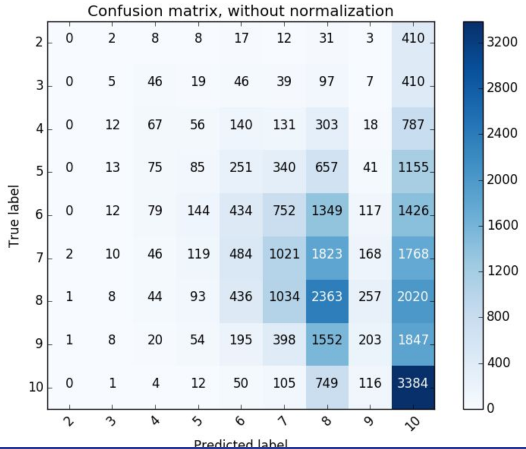
	2	3	4	5	6	7	8	9	10
2	0	0	0	15	0	13	34	0	429
3	0	0	0	54	0	44	133	0	438
4	0	0	0	85	0	181	368	0	880
5	0	0	0	109	0	423	778	0	1307
6	0	0	0	83	0	979	1579	0	1672
7	0	0	0	53	1	1302	2046	2	2037
8	0	0	0	51	0	1194	2674	1	2336
9	0	0	1	19	0	373	1794	1	2090
10	0	0	0	1	0	57	823	1	3539

- ❖ Precision = TP/(TP+FP)
- ❖ Recall = TP/(TP+FN)
- ❖ F1 = 2*P*R/(P+R)



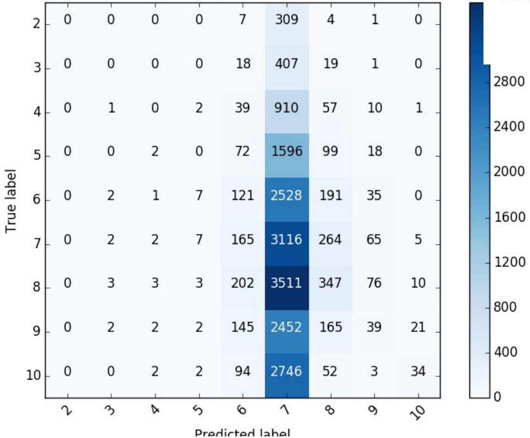
Model	Mean Precision	Mean Recall	Mean F1 score
SVM_linear (C=1)	0.24	0.26	0.19
SVM_rbf (C=1,g=1/31)	0.19	0.25	0.18
SVM_rbf (C=1000,g=0.3)	0.23	0.25	0.20
SGD	0.21	0.20	0.12

Table 3: Precision and Recall score for the models



	2	3	4	5	6	7	8	9	10
2	0	2	8	8	17	12	31	3	410
3	0	5	46	19	46	39	97	7	410
4	0	12	67	56	140	131	303	18	787
5	0	13	75	85	251	340	657	41	1155
6	0	12	79	144	434	752	1349	117	1426
7	2	10	46	119	484	1021	1823	168	1768
8	1	8	44	93	436	1034	2363	257	2020
9	1	8	20	54	195	398	1552	203	1847
10	0	1	4	12	50	105	749	116	3384

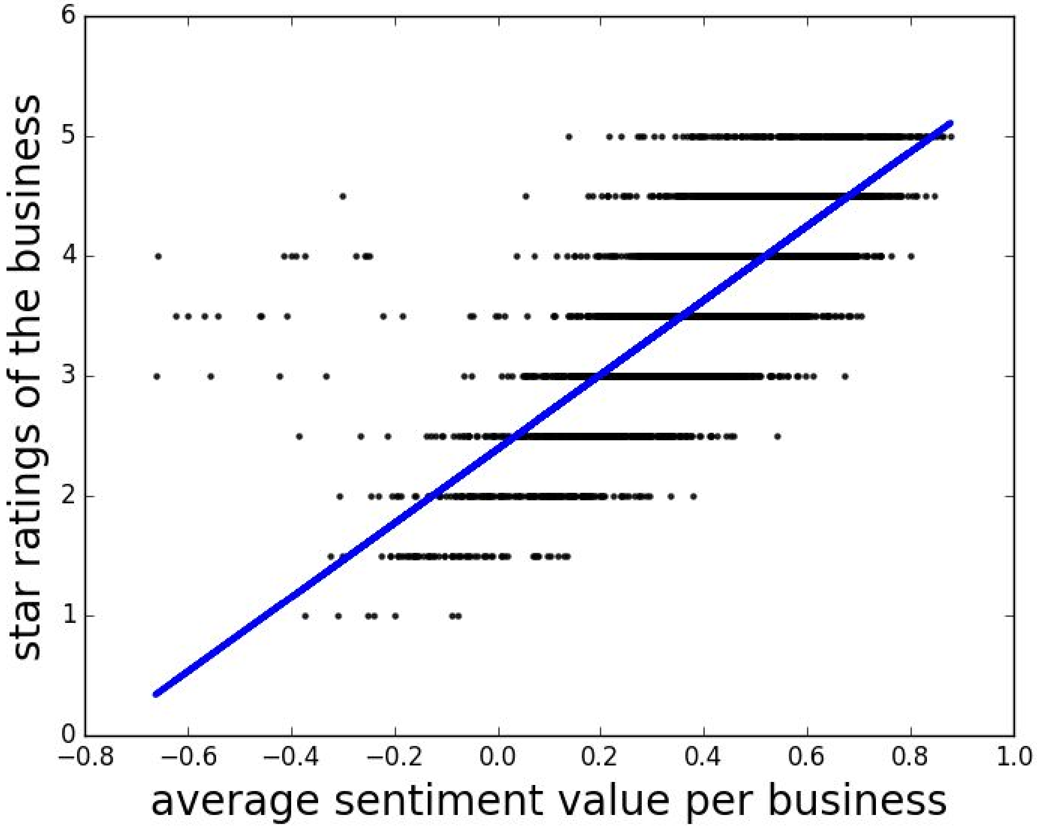
Confusion matrix, without normalization

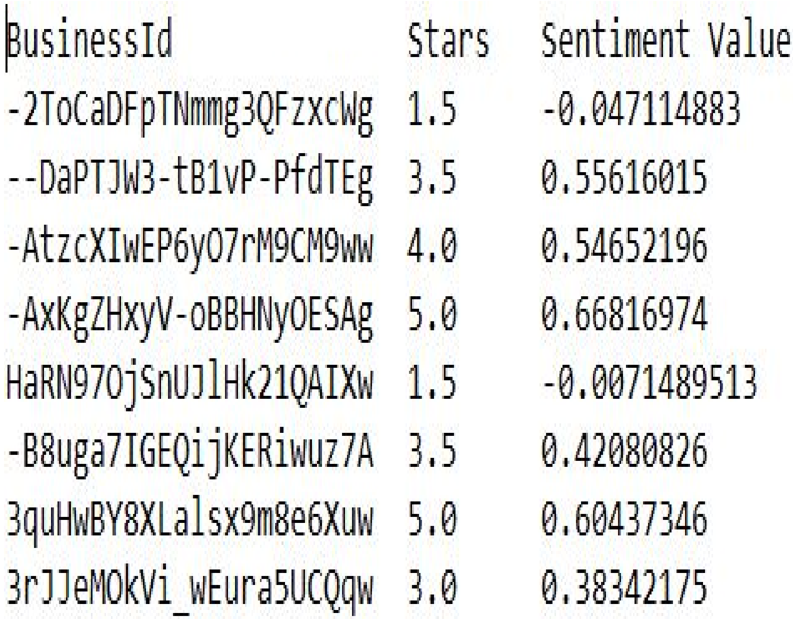


	2	3	4	5	6	7	8	9	10
2	0	0	0	0	7	309	4	1	0
3	0	0	0	0	18	407	19	1	0
4	0	1	0	2	39	910	57	10	1
5	0	0	2	0	72	1596	99	18	0
6	0	2	1	7	121	2528	191	35	0
7	0	2	2	7	165	3116	264	65	5
8	0	3	3	3	202	3511	347	76	10
9	0	2	2	2	145	2452	165	39	21
10	0	0	2	2	94	2746	52	3	34

Sentiment Analysis of Reviews on Business

- Sentiment score for each review
- Sentiment score = (n_Positive - n_Negative)/(n_Positive + n_Negative)
- Priority based sorting of users - stars, sentiment score, useful votes
- Top-N user suggestion.
- Average sentiment score for each business.
- Linear regression based star prediction using average sentiment score of business.
- Mean Square Error = 0.23

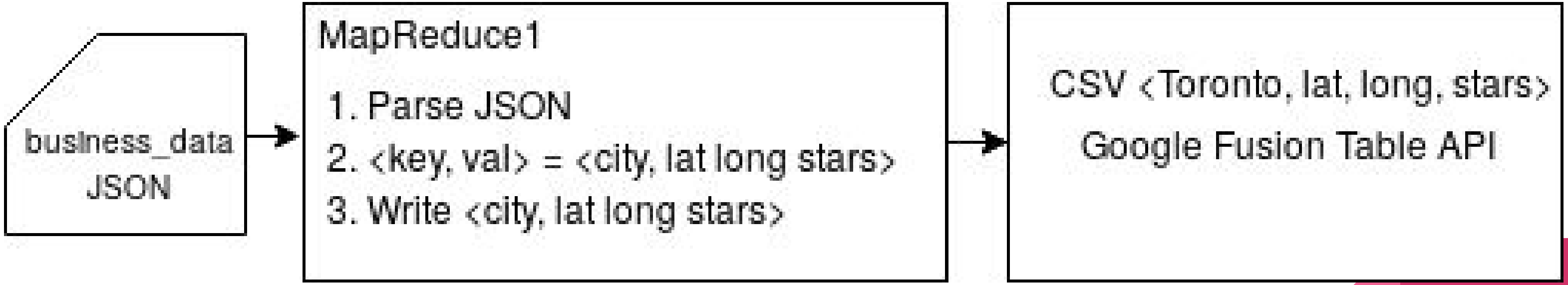




BusinessId	Stars	Sentiment Value
-2ToCaDFpTNmmg3QFzxcWg	1.5	-0.047114883
--DaPTJW3-tB1vP-PfdTEg	3.5	0.55616015
-AtzcXIwEP6yO7rM9CM9ww	4.0	0.54652196
-AxKgZHxyV-oBBHNyOESAg	5.0	0.66816974
HaRN97OjSnUJlHk21QAIXw	1.5	-0.0071489513
-B8uga7IGEQijKERiwuz7A	3.5	0.42080826
3quHwBY8XLalsx9m8e6Xuw	5.0	0.60437346
3rJJeMOkVi_wEura5UCQqw	3.0	0.38342175

Geospatial Visualization of Ratings

- Visualization to reveal trends in star ratings within a city
- Parse business_data in MapReduce with ‘city’ as key
- Get all Business_id, Latitude, Longitude, and Stars
- Use Google fusion table API for heatmap

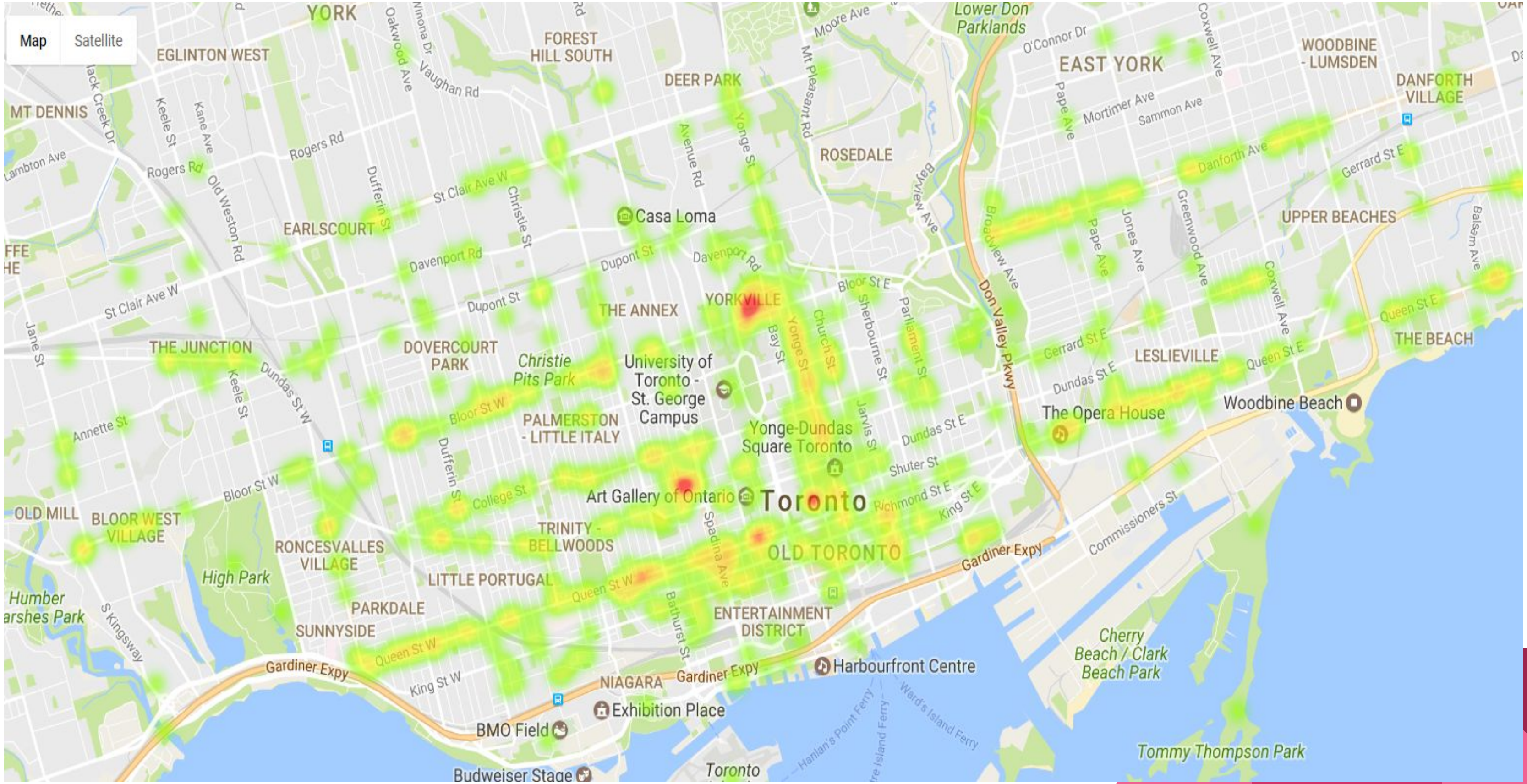


business_data
JSON

MapReduce1

1. Parse JSON
2. <key, val> = <city, lat long stars>
3. Write <city, lat long stars>

CSV <Toronto, lat, long, stars>
Google Fusion Table API


Map
Satellite
YORK
EGLINTON WEST
FOREST HILL SOUTH
DEER PARK
Moore Ave
Lower Don Parklands
O'Connor Dr
EAST YORK
WOODBINE - LUMSDEN
DANFORTH VILLAGE
MT DENNIS
Oakwood Ave
Vaughan Rd
Rogers Rd
Keele St
Kane Ave
Lambton Ave
Old Weston Rd
Dufferin St
St Clair Ave W
Christie St
Avenue Rd
Yonge St
Mt Pleasant Rd
ROSEDALE
Bayview Ave
Pape Ave
Mortimer Ave
Sammon Ave
Coxwell Ave
Danforth Ave
Gerrard St E
EARLSCOURT
Casa Loma
Davenport Rd
Dupont St
Broadview Ave
Jones Ave
Greenwood Ave
UPPER BEACHES
Balsam Ave
St Clair Ave W
Dupont St
THE ANNEX
YORKVILLE
Bloor St E
Sherbourne St
Parliament St
Don Valley Pkwy
Queen St E
THE JUNCTION
DOVERCOURT PARK
Christie Pits Park
University of Toronto - St. George Campus
Bay St
Yonge St
Church St
Gerrard St E
LESLIEVILLE
THE BEACH
Jane St
Dundas St W
Bloor St W
PALMERSTON - LITTLE ITALY
Yonge-Dundas Square Toronto
Jarvis St
Dundas St E
Dundas St E
The Opera House
Queen St E
Woodbine Beach
Annette St
Shuter St
Bloor St W
College St
Art Gallery of Ontario
Toronto
Richmond St E
King St E
Commissioners St
OLD MILL
BLOOR WEST VILLAGE
RONCESVALLES VILLAGE
TRINITY - BELLWOODS
Spadina Ave
OLD TORONTO
Gardiner Expy
High Park
LITTLE PORTUGAL
Queen St W
Humber Marshes Park
S Kingsway
PARKDALE
Bathurst St
ENTERTAINMENT DISTRICT
SUNNYSIDE
Queen St W
Gardiner Expy
King St W
NIAGARA
Gardiner Expy
Harbourfront Centre
Cherry Beach / Clark Beach Park
Exhibition Place
BMO Field
Hanlan's Point Ferry
Ward's Island Ferry
Toronto
Budweiser Stage
Tommy Thompson Park

Conclusion

- We did frequency based feature selection for attributes of business.
- Ratings prediction formulated as a 11-classification problem.
- To improve predictions:
    - Need better features
    - Need to optimize parameters for classification algorithms
- Sentiment Analysis: Unigrams-based sentiment results in positive correlation between sentiment value and the star ratings.
- Visualization helps to decide what areas, streets in a city have higher density of 4.5 or 5 stars.

Questions?

Thank You.