Report
INFO-I590: SQL and NoSQL
Fall 2018
Machine Learning on Yelp Data Set Using NEO4J

Angad Beer Singh Dhillon (adhillon)
Pulkit Mathur (pulmath)
Prashanth Modak (prkumoda)

***Abstract*— Yelp Data Set is a data set where relations between different entities exist dynamically and hence a graph data set is the bet choice for such a data set. We chose NEO4J due to high performance and availability. We have done Exploratory Data Analysis and Machine Learning on this data set.**

I. INTRODUCTION

The motivation of this project came from the fact that we wanted to explore graph databases in more detail and execute machine learning algorithms on the data. After some research we found out more about Yelp Data Set which is an open source data set suited to be used in a graph databases. We loaded this data into the database after defining some constraints of unique properties.

II. ABOUT THE DATA SET

The Yelp dataset is a subset of our businesses, reviews, and user data for use in personal, educational, and academic purposes. Available as JSON files, use it to teach students about databases, to learn NLP, or for sample production data while you learn how to make mobile apps.

1,185,348 tips by 1,518,169 users Over 1.4 million business attributes like hours, parking, availability, and ambience Aggregated check-ins over time for each of the 188,593 businesses

- Business JSON contains business data including location data, attributes, and categories.
- Review JSON contains full review text data including the user id that wrote the review and the business id the review is written for.
- Review JSON contains user data including the user's friend mapping and all the metadata associated with the user.
- Checkin Data contains checkins on a business.
- Tips JSON contains tips written by a user on a business. Tips are shorter than reviews and tend to convey quick suggestions.
- Photo JSON contains photo data including the caption and classification (one of ”food”, ”drink”, ”menu”, ”inside” or ”outside”).

III. ABOUT NEO4J

Very simply, a graph database is a database designed to treat the relationships between data as equally important to the data itself. It is intended to hold data without constricting it to a pre-defined model. Instead, the data is stored like we first draw it out showing how each individual entity connects with or is related to others.

A. Nodes

- Nodes are the main data elements
- Nodes are connected to other nodes via relationships
- Nodes can have one or more properties (i.e., attributes stored as key/value pairs)
- Nodes have one or more labels that describes its role in the graph
- Example: Person nodes vs Car nodes

B. Relationships

- Relationships connect two nodes
- Relationships are directional
- Nodes can have multiple, even recursive relationships
- Relationships can have one or more properties (i.e., attributes stored as key/value pairs)

C. Properties

- Properties are named values where the name (or key) is a string
- Properties can be indexed and constrained
- Composite indexes can be created from multiple properties

D. Labels

- Labels are used to group nodes into sets
- A node may have multiple labels
- Labels are indexed to accelerate finding nodes in the graph
- Native label indexes are optimized for speed

IV. LOADING YELP DATA TO NEO4J

A. Schema

We used the following schema model to build our graph database

- Category of a business is a Node.
- Business is a Node.
- User is a Node.
- User is linked to the Review Node using the ”writes” relationship.
- The review Node contains all the review information from rating to comments.
- This review node is linked to the business for which it is reviewing using the ”Reviews” relationship.
- Business Node is linked to the Category Node using the ”IN Category” Relationship.

Fig. 1. Schema Model defines all nodes and relationships

B. Unique Properties

Having unique properties in a graph database helps to avoid duplicate nodes. Also querying becomes easier as the query is built upon a unique property. For instance we can query to find a busines named ”ABC” and then traverse the graph to find the reviews of that business.

The following are the unique properties in our database

- Category Name is unique, hence there can be only one node with one category name.
- User ID is unique. Since in JSON data users have unique ids and we have used these unique ids to build unique user nodes.
- Business ID is unique. Since in JSON data businesses have unique ids and we have used these unique ids to build unique business nodes.
- Review Nodes will automatically be unique since they are going from unique users to unique businesses.

C. Queries using Python Client

Please find below few queries which were run after entire yelp dataset was uploaded to neo4j:

V. Exploratory Data Analysis

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns,to spot anomalies,to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

```
# get number of business nodes in the graph
query = """
    MATCH (b:Business)
    RETURN COUNT(b) as num_business_nodes
    """

with driver.session() as session:
    business_nodes = session.run(query)

business_nodes_df = pd.DataFrame([dict(record) for record in business_nodes])
display(business_nodes_df)
```

	num_business_nodes
0	180009

Fig. 2. Query 1

 # get distribution of num_reviews for businesses

```
query = """
        MATCH (b:Business)
        RETURN b.name as name, b.num_reviews as num_reviews
        ORDER BY num_reviews DESCENDING
        """

with driver.session() as session:
    business_reviews = session.run(query)

business_reviews_df = pd.DataFrame([dict(record) for record in business_reviews])
display(business_reviews_df.head(10))

display(business_reviews_df.describe(percentiles=[0.1,0.15,0.2,0.25,0.5,0.6,0.75,0.8,0.9]))
```

	name	num_reviews
0	Pennzoil Tire Lube Express	None
1	Midtown Digital	None
2	Las Vegas Red Rock Tours	None
3	Precious Paws Pet Sitting	None
4	Cost Cutters	None
5	Bikers Bay	None
6	Phillip Berry - American Family Insurance	None
7	Ciao Pizzeria Cerino	None
8	O'Reilly Auto Parts	None
9	Nielsen's Frozen Custard	None

Fig. 3. Query 2

Exploratory Data Analysis is a philosophical and an artistical approach to guage every nuance from the data at early encounter.

Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to:

- maximize insight into a data set
- uncover underlying structure
- extract important variables
- detect outliers and anomalies
- test underlying assumptions
- develop parsimonious models
- determine optimal factor settings

EDA is not identical to statistical graphics although the two terms are used almost interchangeably. EDA is an approach to data analysis that postpones the usual assumptions about what kind of model the data follow with the more direct approach of allowing the data itself to reveal its underlying structure and model. EDA is not a mere collection of techniques. EDA is a philosophy as to how we dissect a data set; what we look for; how we look; and how we interpret. It is true that EDA heavily uses the collection of techniques that we call ”statistical graphics”, but it is not

```
query = """
        MATCH (c:Category)
        RETURN c.name as name, size(()-[:IN_CATEGORY]->(c)) AS num_business_in_cat
        ORDER BY c, num_business_in_cat DESC
        """

with driver.session() as session:
    business_in_category_count = session.run(query)

business_in_category_count_df = pd.DataFrame([dict(record) for record in business_in_category_count])
display(business_in_category_count_df.sort_values('num_business_in_cat', ascending=False).head(10))

plt.hist(business_in_category_count_df['num_business_in_cat'])
plt.show()
display(business_in_category_count_df.describe(percentiles=[0.1,0.15,0.2,0.25,0.5,0.6,0.75,0.8,0.9]))
```

	name	num_business_in_cat
55	Restaurants, Pizza	1049
27	Pizza, Restaurants	1002
5	Coffee & Tea, Food	991
85	Nail Salons, Beauty & Spas	964
90	Beauty & Spas, Nail Salons	944
281	Food, Coffee & Tea	925
433	Restaurants, Mexican	888
8	Mexican, Restaurants	867
175	Beauty & Spas, Hair Salons	860
195	Restaurants, Chinese	854

Fig. 4. Query 3

identical to statistical graphics per se. The primary goal of EDA is to maximize the analyst's insight into a data set and into the underlying structure of a data set, while providing all of the specific items that an analyst would want to extract from a data set.

Exploratory data analysis is generally cross-classified in two ways. First, each method is either non-graphical or graphical. And second, each method is either univariate or multivariate (usually just bivariate).The four types of EDA are univariate non-graphical, multivariate non-graphical, univariate graphical, and multivariate graphical.

The graphical EDA on the entire Yelp Data Set gave the following results:

```
yelp_review.head()
```

	review_id	date	stars	useful
0	x7mDIiDB3jEiPGPHOmDzyw	2011-02-25	2	0
1	dDl8zu1vWPdKGihJrwQbpw	2012-11-13	5	0
2	LZp4UX5zK3e-c5ZGSeo3kA	2014-10-23	1	3
3	Er4NBWCmCD4nM8_p1GRdow	2011-02-25	2	2
4	jsDu6QEJHbwP2Blom1PLCA	2014-09-05	5	0

Fig. 5. Yelp Review Data

VI. Machine Learning

Machine learning (ML) is a category of algorithm that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. The basic premise of machine learning is to build algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available.

```
yelp_business.head()
```

	business_id	name	neighborhood	address	city	state	postal_code	latitude	longitude	stars	review_count	is_
0	AjEbIBw6ZFfln7ePHha9PA	CK'S BBQ & Catering	NaN	NaN	Henderson	NV	89002	35.960734	-114.939821	4.5	3	0
1	O8S5hYJ1SMc8fA4QBtVujA	La Bastringue	Rosemont-La Petite-Patrie	1335 rue Beaubien E	Montréal	QC	H2G 1K7	45.540503	-73.599300	4.0	5	0
2	bFzdJJ3wp3PZssNEsyU23g	Geico Insurance	NaN	211 W Monroe St	Phoenix	AZ	85003	33.449999	-112.076979	1.5	8	1

Fig. 6. Yelp Business Data

Top 50 most reviewed businesses

```
yelp_business[['name', 'review_count', 'city', 'stars']].sort_values(ascending=False, by="review_count")[0:50]
```

	name	review_count	city	stars
137634	Mon Ami Gabi	7968	Las Vegas	4.0
185166	Bacchanal Buffet	7866	Las Vegas	4.0
62722	Wicked Spoon	6446	Las Vegas	3.5
188308	Gordon Ramsay BurGR	5472	Las Vegas	4.0
170128	Hash House A Go Go	5382	Las Vegas	4.0
177572	Earl of Sandwich	4981	Las Vegas	4.5
181523	The Buffet	4240	Las Vegas	3.5
116243	The Cosmopolitan of Las Vegas	4097	Las Vegas	4.0
135007	The Buffet at Bellagio	4091	Las Vegas	3.5
180794	Secret Pizza	4078	Las Vegas	4.0
174881	ARIA Resort & Casino	4041	Las Vegas	3.5

Fig. 7. Top 50 Most Reviewed Businesses

A. Supervised Learning

- Machine learning algorithms are often categorized as supervised or unsupervised.
- Supervised algorithms have both input and desired output, in addition to furnishing feedback about the accuracy of predictions during algorithm training. One needs to determine which variables, or features, the model should analyze and use to develop predictions
- Once training is complete, the algorithm will apply what was learned to new data. During training for supervised learning, systems are exposed to large amounts of labelled data, for example images of handwritten figures annotated to indicate which number they correspond to.
- Given sufficient examples, a supervised-learning system would learn to recognize the clusters of pixels and shapes associated with each number and eventually be able to recognize handwritten numbers, able to reliably distinguish between the numbers 9 and 4 or 6 and 8.

B. Unsupervised Learning

- Unsupervised algorithms do not need to be trained with desired outcome data. Instead, they use an iterative approach called deep learning to review data and arrive at conclusions.
- Unsupervised learning algorithms – also called neural networks – are used for more complex processing tasks than supervised learning systems, including image recognition, speech-to-text and natural language generation.
- These neural networks work by combing through millions of examples of training data and automatically identifying often subtle correlations between many vari-

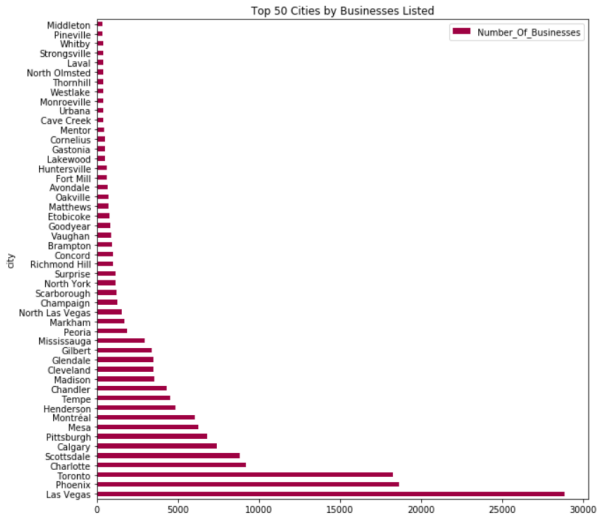

Fig. 8. Top 50 Cities by Businesses

Cities with most reviews and best ratings for their businesses

```
city_business_reviews = yelp_business[['city', 'review_count', 'stars']].groupby(['city']).\
agg({'review_count': 'sum', 'stars': 'mean'}).sort_values(by='review_count', ascending=False)
city_business_reviews.head(10)
```

city	review_count	stars
Las Vegas	1825401	3.731024
Phoenix	660975	3.686739
Toronto	474780	3.490073
Scottsdale	351139	3.969225
Charlotte	273831	3.583442
Pittsburgh	201624	3.640212
Henderson	194803	3.798546
Tempe	182638	3.740205
Mesa	154284	3.657557
Chandler	141092	3.774110

Fig. 9. Cities with Most Reviews and Best Ratings

ables. Once trained, the algorithm can use its bank of associations to interpret new data.

- These algorithms have only become feasible in the age of big data, as they require massive amounts of training data.Unsupervised learning tasks algorithms with identifying patterns in data, trying to spot similarities that split that data into categories.

C. *Semi-supervised Learning*

- Semi supervised learning mixes supervised and unsupervised learning. The technique relies upon using a small amount of labelled data and a large amount of unlabelled data to train systems.
- The labelled data is used to partially train a machine-learning model, and then that partially trained model is used to label the unlabelled data, a process called pseudo-labelling.
- The model is then trained on the resulting mix of the labelled and pseudo-labelled data.

Our Machine learning involved predicting rating of restaurants. The rating lied in the range of 1 to 5 so it was a multi-class classification problem. We used business and checking data and combined the data using business ID as a common feature and developed a new file more suitable for rating prediction.

Firstly, we dropped the unnecessary features. Dropped highly correlated features. The following correlation matrix was generated:

Next step to handle missing values. All categorical features were replaced by mode of that feature and all numeric features were replaced by the mean of that feature. Then we converted categorical columns to numeric and performed mean normalization of the data.

```
city_business_reviews['review_count'][0:50].plot(kind='barh', stacked=False, figsize=[10,10], \
                                                 colormap='seismic')
plt.title('Top 50 cities by reviews')
```

```
Text(0.5,1,'Top 50 cities by reviews')
```

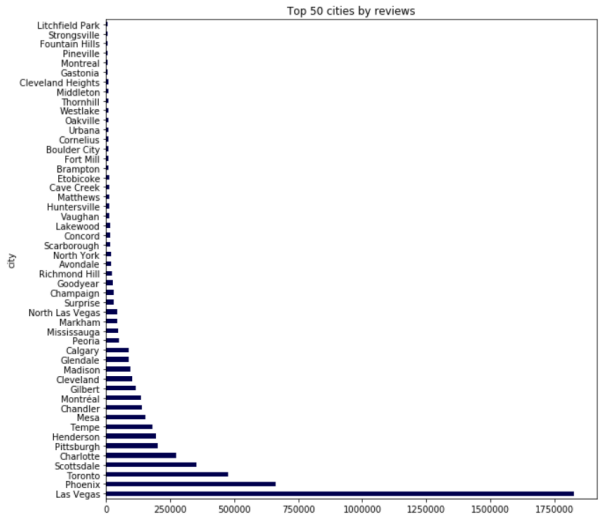

Fig. 10. Top 50 Cities by Review

```
city_business_reviews[city_business_reviews.review_count > 50000]['stars'].sort_values()\
.plot(kind='barh', stacked=False, figsize=[10,10], colormap='gist_heat')
plt.title('Cities with greater than 50k reviews ranked by average stars')
```

```
Text(0.5,1,'Cities with greater than 50k reviews ranked by average stars')
```

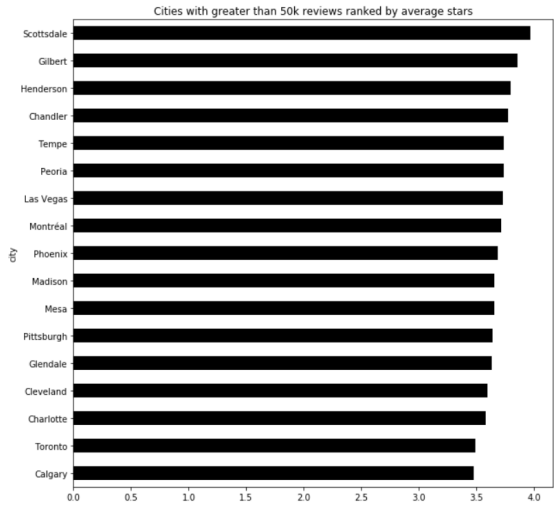

Fig. 11. Cities with more that 50k Reviews sorted by stars

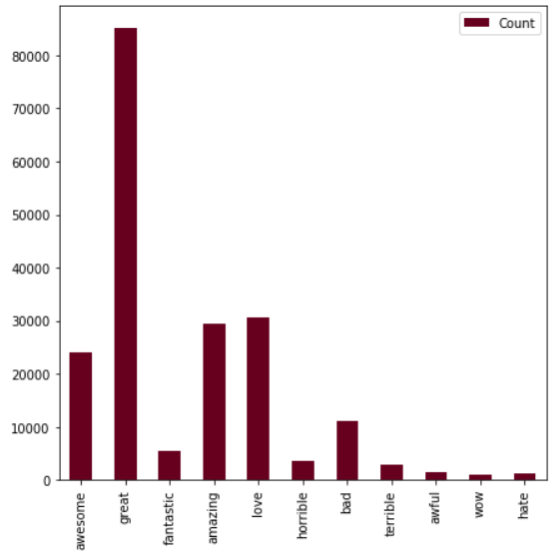

Fig. 12. Count of Most Common Words - Used tf-idf

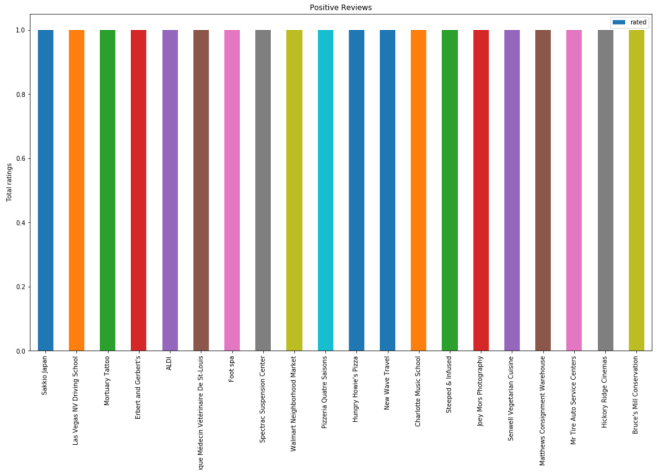

Fig. 13. Positive Reviews v/s Ratings

We trained the following models and results obtained were as follows:

- Decision Tree Classifier
  - Default depth
  - One with Max depth as 5
  - One with Max depth as 5 and probability True
  - One with Max depth as 10 and probability True
  - One with Max depth as 15 and probability True
- Ada Boost Ensemble Technique
- Support Vector Machine
- Linear SVC
- SVM with Linear Kernel
- SVM with RBF Kernel
- SVC with Various Values of Gamma and C
- Random Forest
  - Number of Estimators: 30 Maximum Depth: 10
  - Number of Estimators: 40 Maximum Depth: 5
  - Number of Estimators: 60 Maximum Depth: 8

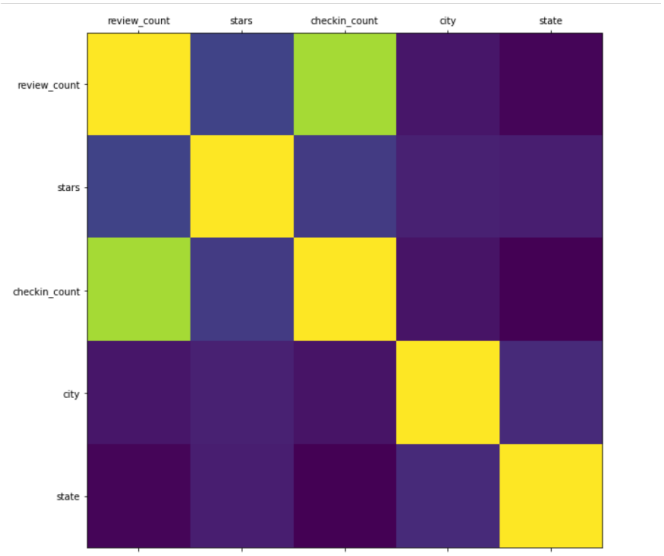

Fig. 14. Correlation Matrix

- Logistic Regression
- k-Nearest Neighbor Classifier with odd values of k in the range [3,21]
- Naive Bayes
- Neural Network
  - Hidden Layers: 3; Number of Neurons: 100,100,100; Activation Function: logistic
  - Hidden Layers: 2; Number of Neurons: 80,160; Activation Function: tanh
  - Hidden Layers: 4; Number of Neurons: 50,100,150,200; Activation Function: relu

VII. Conclusion

The decision tree classifier gave a test accuracy of about 61.

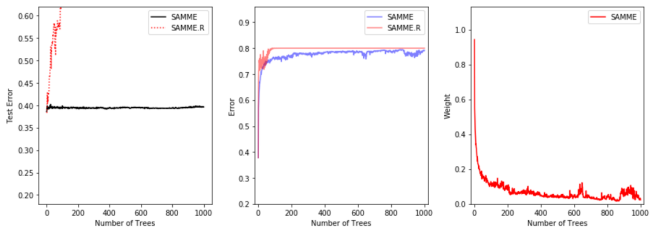

Fig. 15. Decision Tree Classifier

The Support Vector Machines were able to give a test accuracy of about 61 and a precision of 0.56.

Similarly all the models worked fairly well and were able to get highest test accuracy with a Linear SVC of 69.

References

[1] Neo4j: https://neo4j.com/graphacademy/
[2] Supervised Learning: http://www.compciv.org/topics/data/yelp-and-sentiment-analysis/
[3] Predicting Yelp Star Reviews Based on Network Structure with Deep Learning: https://arxiv.org/abs/1712.04350

[4] Machine Learning: https://medium.com/machine-learning-for-humans/supervised-learning-740383a2feab
[5] Random Forest Algorithm: https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd
[6] Support Vector Machine Active Learning with Applications to Text Classification: http://www.jmlr.org/papers/v2/tong01a.html
[7] Text categorization with Support Vector Machines: Learning with many relevant features: https://link.springer.com/chapter/10.1007
[8] SVM: https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72
[9] Neural Networks and Deep Learning: http://neuralnetworksanddeeplearning.com/