International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 21 (2018) pp. 15035-15039

Exploratory Data Analysis and Latent Dirichlet Allocation on Yelp Database

Sairam Kaleru[1], Srinivasa Rao Dhanikonda [2]

[1]BPK Tech Services India Private LTD, Hyderabad, India.

[2] GITAM (Deemed to be University) , Hyderabad, India.

Abstract

Exploratory data analysis plays an important role in understanding the underlying content in the data and explores statistical analysis in the form of descriptive and inferential analysis. In this paper we are taking yelp dataset and we applied exploratory data analysis to understand the features and to understand the underlying content in the dataset, from this result again we applied deep dive individual analysis on reviews given by the users on an individual entity, by using Topic modeling (Latent Dirichlet Allocation) we were able to get the frequent topics and create word cloud visualization on negative and positive words.

Keywords : Exploratory data analysis, Latent Dirichlet Allocation, Topic modeling, Word cloud, Visualization, Yelp dataset

INTRODUCTION

Exploratory data analysis (EDA) in statistics helps data analyst to understand the main characteristics of the data mostly in visual methods, which will further lead to formulating hypothesis, and conduction new experiments. Exploratory data analysis is divided in two ways. First, method is either non-graphical or graphical. And second, method is either univariate or multivariate. This paper uses EDA to understand the yelp dataset which is an interesting schema data or relational data. Yelp dataset (https://www.yelp.com/dataset) has around 5,200,000 user reviews and information about 174,000 businesses from 11 metropolitan areas, with this rich data and relational between the data and we have taken a sample of 20000 rows. After doing EDA, Word cloud visualization was used to understand the negative and positive words on a single entity and Latent Dirichlet Allocation gives a convenient way to analyze unclassified text. LDA contains a cluster of words that frequently occurs together. We used this topic modeling to create topics on the text for one single entity and programming language used was Python and R

We also did the analysis of customer feedback search from google search engine with a search string "customer feedback" this was done in R

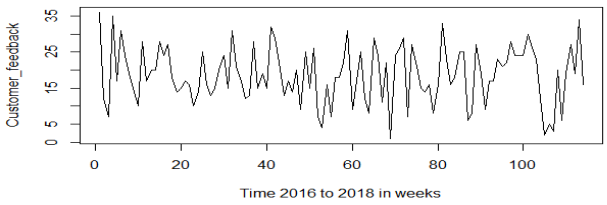

Figure 1. Goolge Trends(www.google.com/trends) time vs customer feedback

LITERATURE SURVEY

Exploratory data analysis by Roger D. Peng 2012[1] this book is having extensive analysis on EDA. Applied Text Analysis with Python**:** Enabling Language Aware Data Products with Machine Learning by Benjamin Bengfort 2017[2]:- Chapters 3 text preprocessing and wrangling and Chapter 6 Clustering for text similarity has been used in this paper. Think Stats: Exploratory Data Analysis by Allen B. Downey 2014 [3] this book discusses about entire process of data analytics from collection data to generating statistical results. Yelp Dataset Challenge: Review Rating Prediction by Nabiha Asghar 2016[4] wrote about review ratings predictions. Good, I. J. "The Philosophy of Exploratory Data Analysis 1983[5]:- paper tries to explain EDA in terms of philosophy. Topic modeling: - Topic modeling presents a way to analyze unclassified the text, A survey of topic modeling in text mining (2015) [6] paper, authors present, what are the different methods in topic modeling and how are they been used.

Text Similarity Computing Based on LDA Topic Model and Word Co-occurrence Minglai Shao and Liangxi Qin 2014 [7] in this paper they have developed text similarity computing algorithm based on hidden topics models and word occurrences was introduced. On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations2010 [8] paper shows https://www.yelp.com/dataset/challenge [9]. Four Experiments on the Perception of Bar Charts - Tableau Research 2014 [10] paper deals with bar plots extensively. Experience of data analytics in EDA and Test- principles, promises, challenges 2016 [11] paper reviews data mining methods and machine learning in electronic design automation

and test. The application of exploratory data analysis in auditing 2014 [12] paper, shows how EDA is been applied in auditing.

Visualization: - Word cloud explorer: text analytics based on word cloud (2014) [13], writes about visualization of text through word cloud and their tool called word cloud explorer. Word cloud visualization for multiple text documents 2015 [14] this paper deals with multiple text analysis by word cloud. (Kushal Dave, Steve Lawrence and David M.P ennock 2003) [15], had developed a method that automatically distinguishing between positive and negative reviews, where techniques like SVM with n-grams was used and performance was measured by metrics (Precision and recall). (Bo Pang, Lee, Sriva kumar 2002) [16], classified not by topic by the overall sentiment.

DATA

As this dataset is a schema type dataset, original dataset is in Json format but was converted to CSV format by the YELP team.

Files that we will be working on are as below: -

yelp_business.csv

yelp_business_attributes.csv

yelp_reviews.csv

check_in.csv

PROBLEM DEFINITION

Generally people don't like at looking a column of numbers in a dataset and determine statistics, looking at numbers can be difficult and time consuming more and not so user friendly, yelp dataset has many tables, So EDA has been used to solve this problem and Business entity wants to know what are reviewers talking about so word cloud was applied, more so most business look trending topics for this reason Topic modeling was used on single entity after Exploratory data analysis.

Exploratory data analysis: Bar plots provide numerical information in a very specific structure

This has several advantages:

- plots can have very high information density, sometimes with no loss of data. By contrast, stating only the mean and standard deviation provides a summary that loses information about, for example, the number and position of outliers.
- plots allow rapid assimilation of the overall result. However, graphs also have some disadvantages, especially if done badly:
- plots take up a lot of space if showing only a few data points. Hence, they are best not used if there are only a few numbers to present.

1. Rating distribution: -

   We wanted to see the ratings Distributions of the reviews.

   Bar graph has been used on yelp business information which has ratings column

   Figure 2 shows the ratings distribution by taking stars ratings and number of businesses

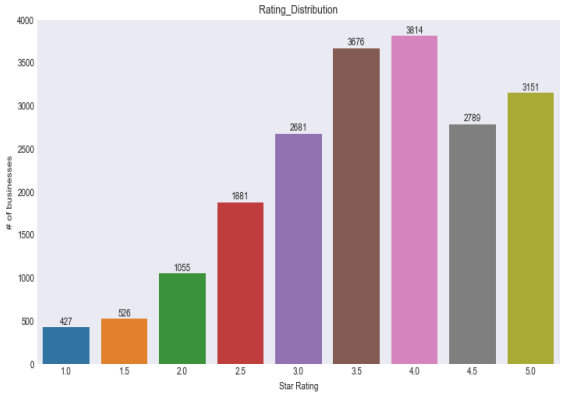

Figure 2. Star ratings VS Number of businesses

2. Top categories:-

   We wanted to know what are the top categories in the dataset. Bar plot was used for this purpose.

   Figure 3 shows the top categories in the yelp business dataset.

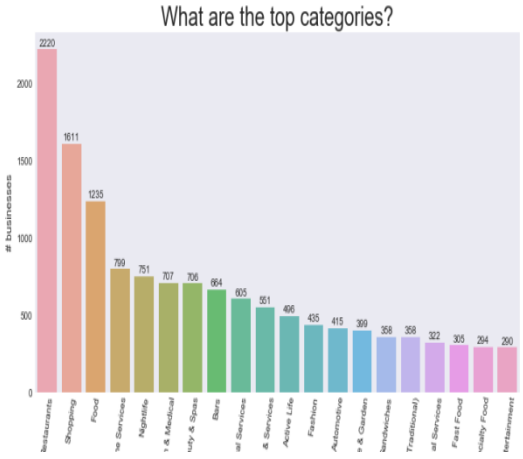

Figure 3. Categories VS Number of businesses

3. Cities with most reviews:-We wanted to know from where the reviews are coming from.

   Figure 4 shows the cities with most reviews in the dataset.

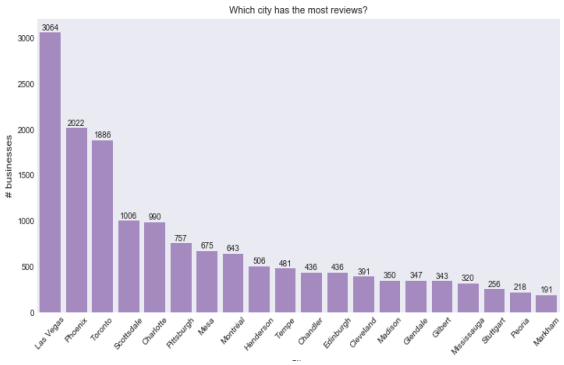

Figure 4. cities VS number of businesses

4. User statistics on review:- It's important to understand the on an average how many reviews are been given by the users and by using this information future analysis is done. Figure 5 shows the cumulative average reviews by users

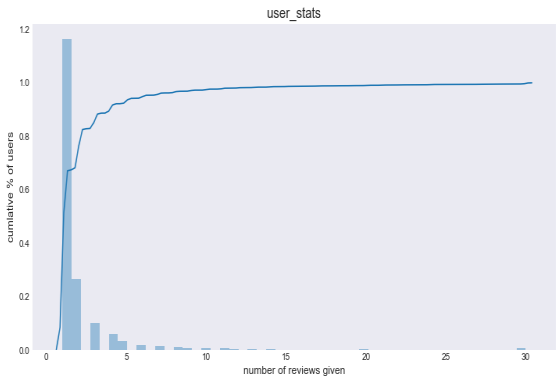

Figure 5. number of reviews given by the users VS cumulative %

5. Check-ins:- Check-ins is an very important parameter for a lot of business and they can use this information to come up with marketing strategies and attract the customers.

hour	Mon	Tue	Wed	Thu	Fri	Sat	Sun
0	1527	633	779	972	682	1086	1559
1	766	661	839	861	809	1162	1310
10	41	24	40	18	34	50	53
11	48	72	69	44	68	127	42
12	142	113	149	164	71	85	67
13	180	230	221	197	328	235	214
14	198	158	213	180	518	254	213
15	223	268	228	447	243	520	503
16	326	431	406	482	463	746	414
17	521	534	425	451	574	1407	1117
18	620	913	600	748	1072	1401	744
19	610	937	736	699	767	1462	1088
2	801	645	938	803	1057	1277	1128
20	754	689	646	548	763	920	1154

Figure 6. table format of check-ins

Descriptive statistics:- Descriptive statistics help in summarizing give the dataset, Descriptive statistics are divided into measures of central tendency and measures of variability. We applied descriptive statistics on restaurants and review count: By taking the review stars and group by business id further sorting the values we got the following results:-

1. From the restaurants and review count, most reviewed are "Gordon Ramsay BurGR" followed by "The Cosmopolitan of Las Vegas"
2. Even for the five star rating also, we find that "Gordon Ramsay BurGR" tops.

Deep dive analysis on Gordon Ramsay BurGR

A 'word cloud' is a visual representation of word frequency. The more commonly the term appears within the text being analyzed, the larger the word appears in the image generated. Word clouds are increasingly being employed as a simple tool to identify the focus of written material. They have been used in politics, business and education, for example, to visualize the content of political speeches. In the Health Board that I support, word clouds have been applied to analyze the content of Board committee papers to see whether sufficient attention is being given to the core business of the organization.

Negative words from Gordon Ramsay BurGR:-

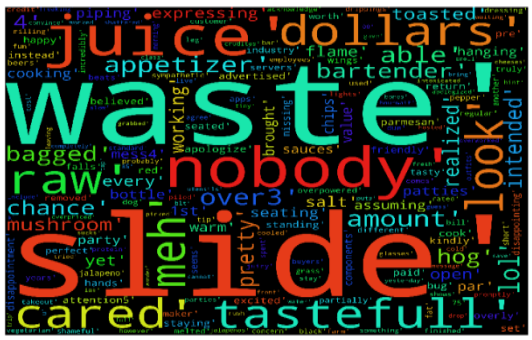

Figure 7. word cloud of negative words

Positive words from Gordon Ramsay BurGR:-

Figure 8. positive words

Latent Dirichlet Allocation (LDA)

First introduced by David Blei, Andrew Ng, and Michael Jordan in 2003, Latent Dirichlet Allocation (LDA) is a technique for topic discovery. LDA belongs to a generative probabilistic model family in which topics are represented as the probability that each of a given set of terms will occur. Documents can in turn be represented in terms of a mixture of these topics. A unique feature of LDA models is that topics are not by definition distinct and words may occur, with differing probabilities, across several topics, before apply LDA we applied text preprocessing and wrangling.

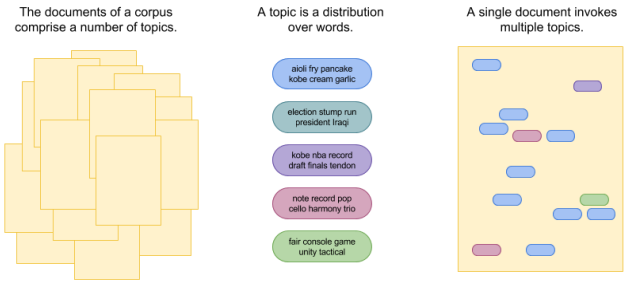

Figure 9. flow sheet of Latent Dirichlet Allocation

Text Preprocessing and Wrangling

Breaking Down the sentences:- We can think paragraphs as the unit of documents structure, it is useful to see sentences as units of discourse, paragraphs has one complete idea in the same way sentence has complete language structure and we can encode and analyze. We used sent_tokenizer from NLTK.

Tokenization and identifying tokens:- Tokens are syntactic units, tokens in the corpus will be sequences of characters that appears in one or more documents and are grouped together to encode some semantic information beyond characters.

Parts of speech tagging:-Parts of speech (e.g. verbs, nouns, prepositions, adjectives) indicate how a word is functioning within the context of a sentence. In English as in many other languages, a single word can function in multiple ways, and we would like to be able to distinguish those uses (for example "building" can be either a noun or a verb). Part-of-speech tagging entails labeling each token with the appropriate tag, which will encode information both about the word's definition and its use in context.

Stop words:- It is typical to exclude high-frequency words (e.g. function words: "a", "the", "in", "to"; pronouns: "I", "he", "she", "it").

Text Normalization: - For text, normalization is intended to reduce the number of features by eliminating or combining different tokens into a single class.

Figure 10 shows the output of Latent Dirichlet Allocation on Gordon Ramsay BurGR.

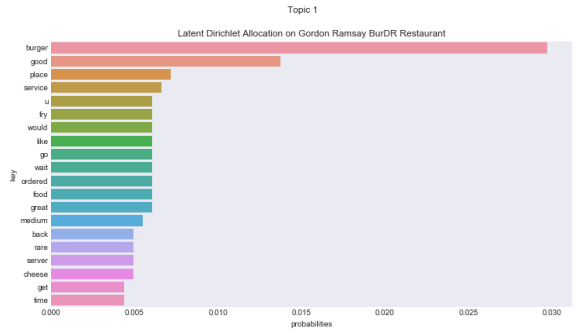

Figure 10. LDA Probabilities VS Words

So the top 4 topics from LDA are burger, good, place, service.

RESULTS

1. From the Rating distribution, we can see that 4 stars count is more and 1 star is the least.
2. Most reviews where from Los Vegas and followed by Phoenix.
3. Top categories are restaurant, followed by shopping.
4. From the cumulative average reviews of users we find that almost 85% of the reviews from users are below 5.
5. From the check-in's we find that max number of check-in's are on Sunday and followed by Monday then by Saturday.
6. From the restaurants and review count, most reviewed are "Gordon Ramsay BurGR" followed by "The Cosmopolitan of Las Vegas".
7. Even for the five star rating also, we find that "Gordon Ramsay BurGR" tops.
8. We were Able to visualize most negative and positive words.
9. From LDA we can see the most frequent topics in visual format.

Top 3 words in topic along with probabilities.

Words in Topic	Probabilities
Burger	**0.003**
Good	**0.015**
Place	**0.0055**

CONCLUSION AND FUTURE ENHANCEMENTS

We were able to successfully demonstrate Exploratory data analysis, identified positive and negative words and created a topic model for one single entity, our future work enhancements will be on topic modeling on all the reviews, creating sentiment analysis using traditional and deep learning methods and also build recommendation system.

REFERENCES

[1] Roger Peng Author, Exploratory data analysis(2012)

[2] Benjamin Bengfort author, Applied Text Analysis with Python: Enabling Language Aware Data Products with Machine Learning (2017)

[3] Allen B. Downey Author,Think Stats: Exploratory Data Analysis by (2014)

[4] Asghar, Nabiha. (2016). Yelp Dataset Challenge: Review Rating Prediction.

[5] I. J. Good, "The Philosophy of Exploratory Data Analysis," Philosophy of Science 50, no. 2 (Jun., 1983): 283-295.

[6] Rubayyi Alghamdi and Khalid Alfalqi, "A Survey of Topic Modeling in Text Mining" International Journal of Applications (IJACSA), 6(1), 2015. http://dx.doi.org/10.14569/IJACSA.2015.060121

[7] Shao, Minglai& Qin, Liangxi. (2014). Text Similarity Computing Based on LDA Topic Model and Word Co-occurrence. 10.2991/sekeie-14.2014.47.

[8] Arun, R., V. Suresh, C. E. VeniMadhavan and M. NarasimhaMurty. "On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations." ***PAKDD***(2010).

[9] https://www.yelp.com/dataset/challenge

[10] Talbot, Justin &Setlur, Vidya&Anand, Anushka. (2014). Four Experiments on the Perception of Bar Charts.

[11] Wang, Li-C. (2016). Experience of Data Analytics in EDA and Test - Principles, Promises, and Challenges. IEEE

[12] Heimerl, Florian, Steffen Lohmann, Simon Lange and Thomas Ertl. "Word Cloud Explorer: Text Analytics Based on Word Clouds." ***2014 47th Hawaii International Conference on System Sciences*** (2014): 1833-1842

[13] S. Lohmann, F. Heimerl, F. Bopp, M. Burch and T. Ertl, "Concentri Cloud: Word Cloud Visualization for Multiple Text Documents," ***2015 19th International Conference on Information Visualisation***, Barcelona, 2015, pp. 114-120. doi: 10.1109/iV.2015.30.

[14] Dave, K.; Lawrence, S. & Pennock, D. M. (2003), Mining the peanut gallery: opinion extraction and semantic classification of product reviews., in 'WWW' , pp. 519-528 .

[15] Pang, Bo & Lee, Lillian & Vaithyanathan, Shivakumar. (2002). Thumbs up? Sentiment Classification Using Machine Learning Techniques. EMNLP. 10.