Exploring the Yelp Dataset

Looking for trends in reviews and stuff

Carter Reid†  
 Applied Computer Science University of Colorado Boulder  
 Boulder, CO U.S.A.  
 reidcc@colorado.edu

Maxwell Reynolds  
 Applied Computer Science  
 University of Colorado Boulder  
Boulder, CO U.S.A  
mare3521@colorado.edu

Surya Jatavallabhula  
 Applied Computer Science  
 University of Colorado Boulder  
Boulder, CO U.S.A  
 suja3865@ colorado.edu

PROBLEM STATEMENT

Yelp is a business directory service and crowd-sourced review forum. That is, the business revolves around the connections made between the consumers who read and write reviews and the local businesses that they describe. (source 1) Since the company’s founding in 2004, it has grown to include 4.6 million active claimed business locations and 192 million cumulative reviews for those 4.6 million business locations. (source 1: <https://www.yelp-ir.com/overview/default.aspx>)

Despite the simplicity of the service offered, there are many attributes tracked and related to each other in the dataset allowing a vast opportunity for data mining.

Yelp.com has an extensive dataset gathered from their online review services, and Yelp has made this dataset available to students in the form of a contest aimed at encouraging students to explore their data and discover novel trends and relations among their reviewers and businesses. This contest has a cash incentive and is on its thirteenth iteration, ending in December 2019.

Some potential questions we would like to answer are:

* Can external factors such as local weather or global economy have an effect on review ratings or sentiment?
* Can characterize potential pitfalls/areas of improvement of a restaurant based on its reviews.
* Can we characterize the most important aspect of a restaurant to a given regional population (i.e. what do reviewers in Austin, TX seem to value the most?).
* Are “funny” reviews considered more or less helpful? What are some characteristics of “funny” reviewers?

**LITERATURE SURVEY**

Previous work using this dataset has included a positivity estimator based on review text and key words, an automatic review generator that generates a review from an initial small text such as ‘They have the best.’ A json to csv converter specific to the dataset. A public repository containing these previous contest submissions can be found here: <https://github.com/Yelp/dataset-examples.A> full list of google scolar citations that used the Yelp dataset is found here: <https://scholar.google.com/scholar?q=citation%3A+Yelp+Dataset&btnG=&hl=en&as_sdt=0%2C5>

Some of these projects include a system for identifying local “experts” (Source: <https://www.ideals.illinois.edu/handle/2142/78499>), or predicting whether a restaurant would close down or stay open (Source: <https://dl.acm.org/citation.cfm?id=3229287>).

**PROPOSED WORK**

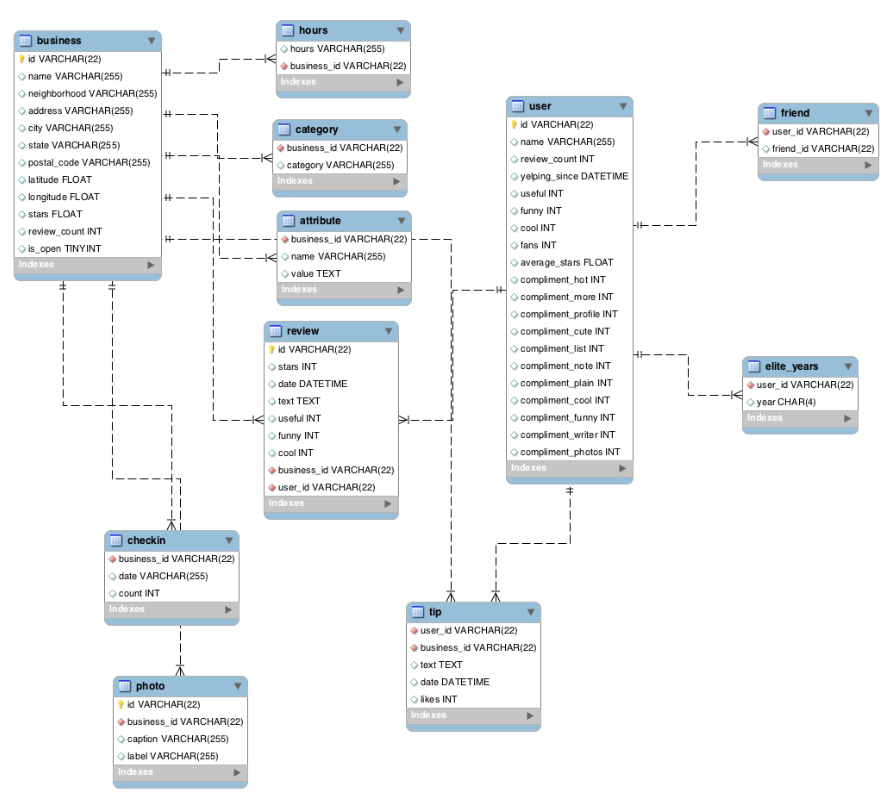
One of our initial tasks will be to design an SQL database to store the data and insert all of our data. We may need to clean aspects of the data such as outliers and normalize the data using Standard Deviation method and resolve if there are any formatting discrepancies. Once the data is relatively cleaned, we should perform an exploratory data analysis to identify trends and potentially form hypotheses regarding what we would like to investigate. After this step, we will be able to start mining the data, building models, and testing our hypotheses.

The dataset itself contains other information as well which could reveal other interesting relationships and correlations. One aspect of the dataset which seems less explored than relating review scores with text, is comparing reviews with external data sources, such as economic data, or specific calendar dates like religious holidays. The yelp dataset provides dates which we plan to normalize and relate to public and religious holidays, economic data, and weather data. We also would like to explore relationships between user attributes and the attributes associated with the reviews they write (useful, funny, cool) to predict their review score of an establishment. We believe that predicting a review score based on other attributes could help direct establishments optimize their hours of operation and focus their marketing on specific demographics to increase average reviews and maximize marketing effectiveness. We might also explore what aspects of restaurants that reviewers care the most about across different regions.

**Data Set**

The available dataset is large. It is 8.69 gigabytes of business, user, and review data with another 7.67 gigabytes of business and customer photos. It is available form yelp directly with a valid school email address. Link here: <https://www.yelp.com/dataset/challenge>

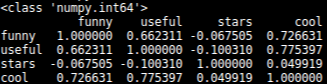
The data is packaged in the form of six json files, with multiple relations existing between the tables.



**Figure 1: A relational database model of the yelp dataset constructed in MySQL workbench. Source:** [**https://github.com/2gotgrossman/yelp-dataset-challenge**](https://github.com/2gotgrossman/yelp-dataset-challenge)

**Evaluation Methods**

Statistical analysis. Looking at correlation coefficients and other statistical metrics for the dataset in the beginning will help guide our progress though the dataset. Certain calculations will take time and rely on statistical constants such as mean, median, standard deviation, variance, correlation coefficients etc. Generating those constants now may save time and computation in the future.



**Figure 2: A correlation matrix of review attributes**

Bayesian analysis of certain attributes associated with user and reviews will be our method of predicting review score based on user attributes and review attributes such as user history and review likes. A Bayesian prediction algorithm seems appropriate for a large complex dataset allowing for unanticipated patterns to emerge as well as specific target patterns.

Additional data sources will be related to the yelp dataset through SQL, using a join on economic dataset date and yelp review date. Date format may need to be normalized between the two datasets in order to create a join, and weekends may need to be treated specially as the stock market is not open over the weekend and after hours trading has a different volatility than standard trading.

**Tools**

The current proposed tools for this exploration are python, github, SQL (sqlite3), and Tableau. Python will be used initially to load portions of the database into pandas dataframes due to ease of use. The dataset itself is too large to store the entire database and therefore a more scalable solution will need to be implemented, as solutions to the size constraint exist, they still leave a lack of optimization and are slow.

An SQL database is slated to be the scalable solution to the dataframe size and runtime constraint. The service for our relational database has yet to be determined but google cloud services appears to be a reasonable choice. This will also allow other data sources to be incorporated into the analysis using a join on date. This does require the date attribute to be normalized to the format used in the yelp dataset (YYYY-MM-DD).

Tableau will potentially used for visualization of the data and trends discovered.

**Milestones**

By November 1st we would like to have all of our data organized into a relational database so that it can be queried. By November 5th we would like to have our data cleaned and pre-processed. By November 15th we would like to have clear hypotheses for questions we will explore and/or models that we will create.

**TODO**

In this sample-structured document, neither the cross-linking of float elements and bibliography nor metadata/copyright information is available. The sample document is provided in “Draft” mode and to view it in the final layout format, applying the required template is essential with some standard steps.

These steps, which should require generation of the final output from the styled paper, are mentioned here in this paragraph. First, users have to run “Reference Numbering” from the “Reference Elements” menu; this is the first step to start the bibliography marking (it should be clicked while keeping the cursor at the beginning of the reference list). After the marking is complete, the reference element runs all the options under the “Cross Linking” menu.

For accuracy check of the structured paper, user can run the option **Manuscript Validation**. It informs the user of the wrong or missing values in the paper. The user must correct the paper as per validation messages and rerun **Manuscript Validation**.

Now, to generate the required layout of the paper, the user should select one of the template styles under the Define Template Style option (choose the required layout design, i.e. choose between Journals and Proceedings).

∗Article Title Footnote needs to be captured as Title Note

†Author Footnote to be captured as Author Note

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

*WOODSTOCK’18, June, 2018, El Paso, Texas USA*

© 2018 Copyright held by the owner/author(s). 978-1-4503-0000-0/18/06...$15.00

https://doi.org/10.1145/1234567890

Some specific values are required to create a standard layout by choosing a template for the journals or proceedings. So once the user chooses one of the template layout styles, the respective Journal or Conference details dialog box (i.e. **journal/conference acronym, DOI, ISBN, copyright, year, etc.**) will appear as a prompt during the Define Template Style functionality. The user should fill these values, after which the template creates the desired layout of the paper. The user can now create a PDF of his/her manuscript using the “*Save as PDF*” option.

If the user is adding any new data, they should make sure to style it as per the instructions provided in previous sections. Carry out the steps for Cross-linking, Fundref data, adding Document History (specific to journal submission), and finally, Manuscript validation and placing the respective metadata (Bibstrip/copyright text)[[1]](#footnote-2) while applying the required template.

CCS CONCEPTS

• Relational databases • Bayesian theorem   • Croud Sourcing

KEYWORDS

Croud-Sourced review forum text, Insert keyword text, Insert keyword text, Insert keyword text

ACM Reference format:

FirstName Surname, FirstName Surname and FirstName Surname. 2018. Insert Your Title Here: Insert Subtitle Here. In *Proceedings of ACM Woodstock conference (WOODSTOCK’18). ACM, New York, NY, USA, 2 pages.* https://doi.org/10.1145/1234567890

1 Insert Heading Level 1

The updated template, user manuals, samples, and required fonts, all are available at the URL <https://www.acm.org/publications/proceedings-template>. It contains said information for all three versions of MS Word (Windows and 2 versions of Mac). There are also separate links to the user guide, which can be referred to by the user. This URL also contains some useful video links, which describe how to add the template, structure the paper, and generate the layout, in different clips. **Display Formula with Number**

 (1)

**Continuation part of Paragraph Text** The user must style this paragraph in **ParaContinue** style, which follows immediately after the **DisplayFormula** (numbered equation). The **DisplayFormula** style is applied only in case of a numbered equation. A numbered equation always has a number to its right. Insert paragraph text here. **Display Formula without Number**



The **DisplayFormulaUnnum** style is applied only in case of an unnumbered equation. An unnumbered display equation never contains an equation number to its right, and this unique property distinguishes it from a numbered equation.



Figure 1: Figure Caption and Image above the caption [In draft mode, Image will not appear on the screen]

**Theorem/Proof/Lemma.** Insert text here for the enunciation or Math statement. Insert text here for the enunciation or Math statement. Insert text here for the enunciation or Math statement. Insert text here for the enunciation or Math statement. Insert text here for the enunciation or Math statement.

....Insert text here for the Quotation or Extract, Insert text here for the Quotation or Extract, Insert text here for the Quotation or Extract, Insert text here for the Quotation or Extract, Insert text here for the Quotation or Extract, Insert text here for the Quotation or Extract.

1.1 Heading Level 2

In the below paragraph, it is explained how alt-txt value is placed in **MS Word 2010**. To add alternative text to a picture in Word 2010, follow these steps:

1. In a Word 2010 document, insert a picture.
2. Right click on the inserted picture and select the **Format Picture** option.
3. Select the **Alt Txt** option from the left-side panel options.
4. In the "Title:" and "Description:" text boxes, type the text you want to represent the picture, and then click "Close".

Below are steps to place alt-txt value in **MS Word 2013/2016**. To add alternative text to a picture in Word 2013/2016, follow these steps:

1. In a Word 2013/2016 document, insert a picture.
2. Right click on the inserted picture and select the **Format Picture** option.
3. In the settings at the right side of the window, click on the "Layout & Properties" icon (3rd option).
4. Expand **Alt Txt** option.
5. In the "Title:" and "Description:" text boxes, type the text you want to represent the picture, and then click "Close".

*1.1.1 Heading Level 3.* Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here.

*1.1.1.1 Heading Level 4.*Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here.

ACKNOWLEDGMENTS

Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here.

REFERENCES

[1] FNM Surname (2018). Article Title. Journal Title, 10(3), 1–10.

[2] F.N.M Surname, Article Title, https://www.acm.org/publications/proceedings-template.

[3] F.N.M Surname and F.N.M Surname, 2018 Article Title, The title of book two (2nd. ed.). Publisher Name, City, State, Country.

[4] Ian Editor (Ed.). 2018. *The title of book two* (2nd. ed.). University of XXX Press, City, Chapter 100. DOI: <http://dx.doi.org/10.1000/0-000-00000-0>.

Conference Name:ACM Woodstock conference

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00

1. The existing Bibstrip data, copyright text and permission block in the sample file are dummy values, so the user needs to provide the correct values required for the submission in the metadata dialog box. [↑](#footnote-ref-2)