Text mining using LDA and LSA

My D. Coyne

November 13, 2015

Introduction

It is very exciting that Yelp makes available the **Yelp Challenge Dataset** to the Data Science Specialization Capstone project for data analysis purpose. The Yelp Challenge Dataset comprises of 5 files – business, review, user, check-in, and tip, where there are 1.5 million reviews of 61K business; the reviews from 366K users with 495K tips.

Using the *business and review datasets*, I aim to answer the following questions for my text analysis. It seems to be natural that a reviewer gives a high score then it is an indication of a good service rendered at a business; similarly, a low review score accompanied with uncomplimentary service received by the reviewers. **Would it is possible to qualitatively determine the review score using text analytics? In other words are there particular terms indicative of the score? The application of the analysis may be used to help Yelpers to retrieve better results from searching reviews. The analysis focuses on business that provides medical services or health centers.**

The paper is organized as follows: Methods and Data section discusses the process of obtaining, cleaning, and preparing data analysis. The Results section discusses findings. Finally, the Dicussion section will interpret the reuslts and answer the above questions.

The source code and this report are also on Github, https://github.com/mdcRed/Capstone. Details of the data elements are found at http://www.yelp.com/dataset_challenge.

Methhods and Data

Preparing Data

The Yelp data is in JSON format, use the function *getJsonData.R* to read a JSON file and "flatten" it into data frames. There are 5 data frames, but the paper only focuses on the business (*businessDf*) and review (*reviewDf*). In this project, since the "raw" and process data are not only large, but also takes long computing time to produce; the intermidiate data R objects are persisted into file system and loaded into R global environment as needed. Database tables are also utilized for peristence and queries as well.

The following table shows top three states that have the largest number of medical doctor, healthcare center, dentists, chiropractor business registered in Yelp. For the scope of this paper, review text for Medical related businesses in the **state Arizona (AZ)** will be analyzed.

```
##   state number of health related business
## 1    AZ                              1614
## 2    NV                               987
## 3    NC                               128
```

AZ_df is the dataframes that contain the merged reviews and business data items for the healthcare related business. R code that generates these the dataframes is found in *combineBusinessReview.R. For state of Arizona, there are 1,614 medical and healthcare providers, with 10,827 review text to form the corpus.*

Preparing data for text mining

In order to prepare text for mining purpose, all review text is used to generate a *corpus*. After which, text is converted to lower case and removal of punctuations, numbers and English stopwords. Stopwords are words like 'a', 'the', 'ive','we', etc. Then, the words in the corpus are stemed. Stemming is a proces to reduce all words in the corpus to its stem. For example, the text may read "The boy's cars are different color"; after stemming process, the text may be converted to "the boy car be differ color" Refer to function *clean.R* for details of this process. R *tm* package provides all functions to prepare text for the mining process that follows.

After the corpus is 'cleaned', the using *tm* package to generate document term matrix (DTM) or term document matrix (TDM). DTM is a matrix of (m-terms by n-documents); whereas the TDM is a transpose matrix of DTM, or (n-documents by m-terms). Refer to *prepCorpus_dtm()* and *prepCorpus_tdm()* for the R code.

Sparsity Both dtm's and tdm's for this text corpus is 100% sparse. When a matrix is sparse, it contains mostly zeroes in its cells. After number of trials in removing the sparsity, the dtm's (hence tdm's) sparsity is reduced to 78-79 percent. At this level of sparsity, the vocabulary contains from 47-53 terms.

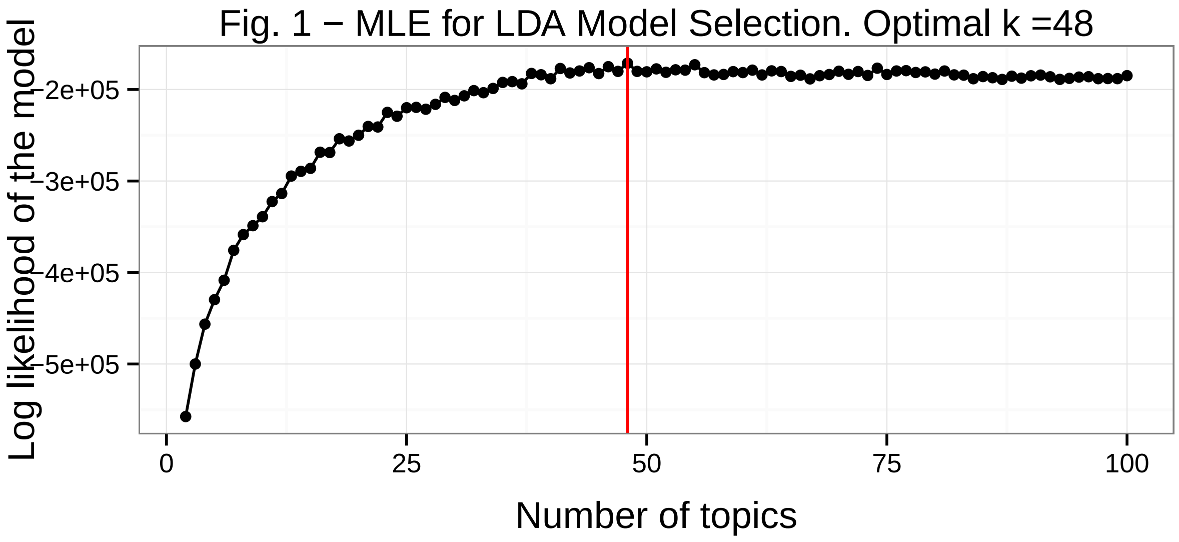
Methods

Latent Dirichlet Allocation (LDA) Topic Modeling

In this analysis, *LDA is used to generate the topics for all review text of all medical and healthcare business of AZ state as found in the Yelp Challenge Dataset.*. Topic comprises of terms. After all reviews were labeled by a topic (set of terms), by cross reference the reviews to the number of stars associated with the review and the business, one may find the association of terms to the review numbers. This is a qualitative way to explain the review number.

Topic Models are "[probabilistic] latent variable models of documents that exploit the correlation among the words and laten semantic themes" (1). Topic ("latent") is hidden; it is to be estimated. Topics link words in a vocabulary and their occurrence in a document. A document is seen to be a mixture of *topics*. LDA relies on the *bag-of-words* assumption–words in a document are exchangeable and their order, therefore, is not imporant.

For LDA model selection, the maxinum log likelihood estimate (MLE) occurs at the **optimum number of topics of 48**, as shown in the following figure.



Latent Semantic Analysis (LSA)

The Latent Semantic Analysis (LSA) model is a theory for how meaning representation might be learned from encountering a large samples of language without explicit directions as to how it is constructed. LSA assumes that the meanin of sentences is assumed to be the sum of the meaning of all the words occuring. Hence the meaning of multi-word phrases is more greatly determined by which words occur in the phrase, rather than how the words are ordered. A second assumption is that the semantic asssociatons between words is *latent* in a large sample of language and *eventually* the meaning is learned. LSA is used in this project to find the **similarity** of review text.

LSA is used in this analysis to find the similarities among the review text.

Results

LDA finds Latent Topics for Review Text

All 10,827 review text of 1,614 medical and healthcare business establishments in Arizona state are grouped into **38** topics. Following table shows the topic number and the five highest score terms that make up each topic; due to space constraint, only selected topics that are used in the discussion are shown.

```
##   lda_topic                          lda_terms
## 1         1   best,staff,friend,recommend,ive
## 2         8            call,back,one,well,feel
## 3         9 friend,staff,recommend,good,great
## 4        17    great,recommend,best,feel,help
## 5        26 staff,friend,great,recommend,feel
## 6        31 staff,friend,recommend,experi,get
## 7        34 experi,great,recommend,friend,now
```

Comparing reviews for business

Contrasting 5-star reviews and 1-star reviews for 5 and 1 star business establisments with the two graphs shown below, the topics that are high in 5-star review, but low in 1-star review are: #1, #9, #17, #31. *Topic 1 with terms "best, staff, friend, recommend"*, *Topic 9 with terms "friend, staff, recommend, good, great"*, *Topic 17 with terms "great, recommend, best, feel, help"*, and *Topic #31 with terms "staff, friend, recommend, friend, now"* are strong indication of good reviews.

By removing the business ranking from the above two graphs, Fig.4 shows 5-star and 1-star reviews for *all* businesses vs topics. The graph **confirms that Topic 1, topic 9, and topic 17 are indicative of best ranking reviews.**

Fig. 2 − Review Topics for 5−star Healthcare business in Arizona

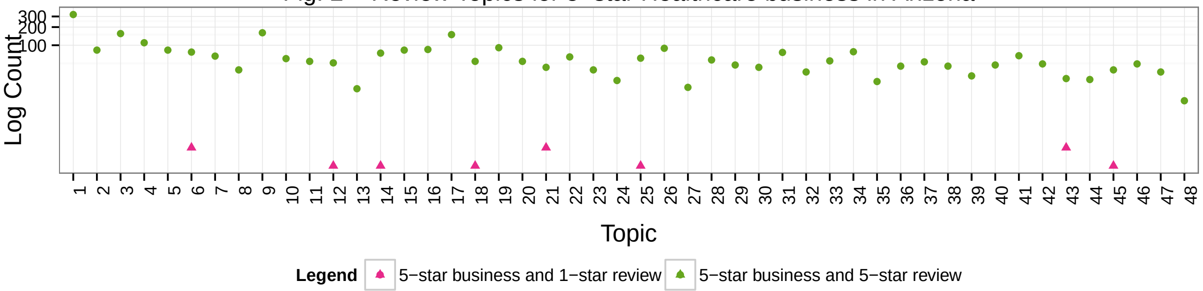


Fig. 3 − Review Topics for 1−star Healthcare business in Arizona

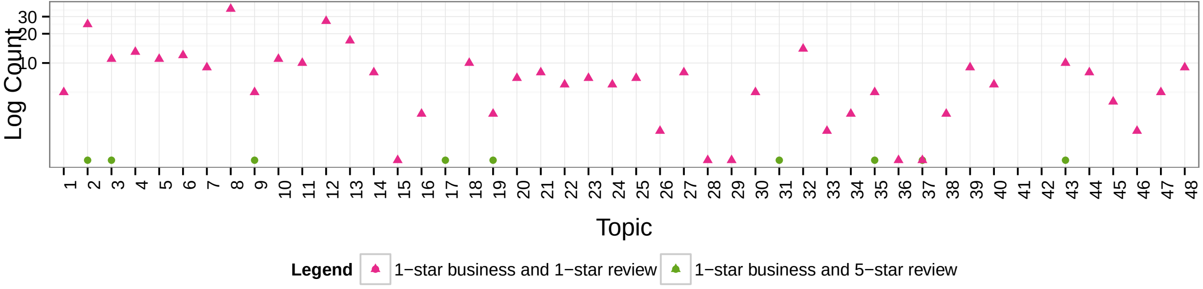
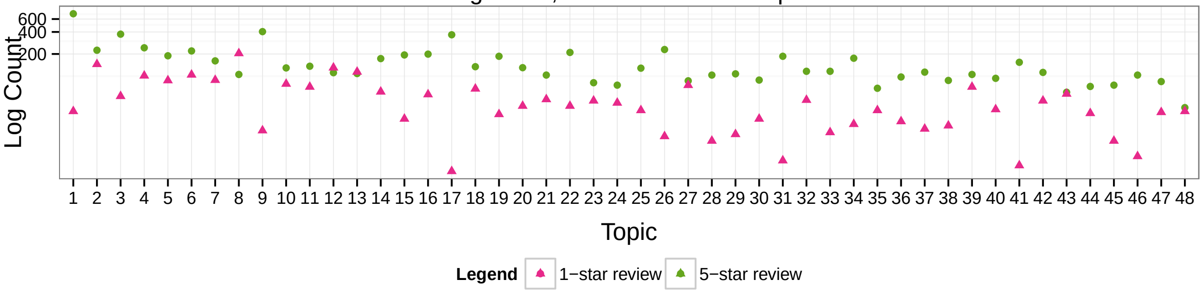


Fig. 4 − 1, 5 Review Star vs Topic

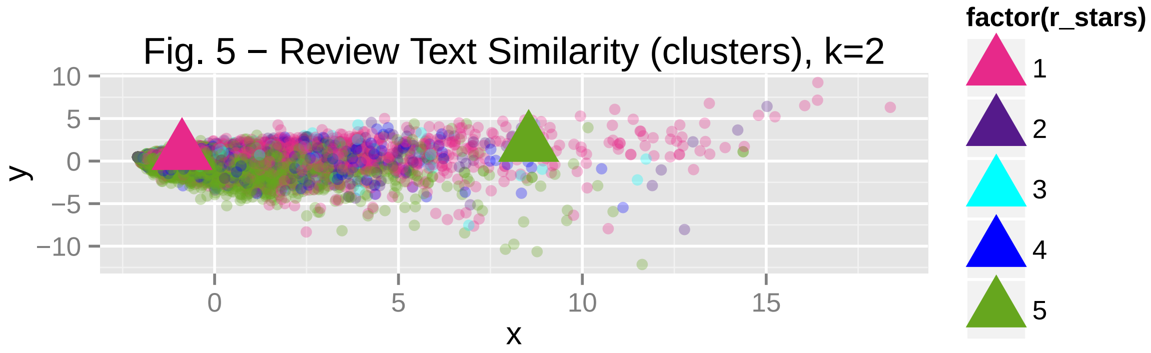


In addition, Fig. 4 also shows an interesting occurence: some 1-star reviews are labeled with *Topic 8 with terms "call, back, one, well, feel", as well as 5-star review.* This is to say that some terms are used indicating both good and undesirable services. By looking at the terms, it is not evident of the reason; one must look into the text. Below table show two specific cases where both reviews were placed in the same topic #8, but the reviews are completely different; yet the topic with terms 'call', 'back' fits well. Latent Semantic Analysis (LSA) is used to review the document similarities to perhaps complement to the LDA technique.

Business ID	Review ID	R star	Example text
1eCvpgvB4QA-0fSwb8-5Dw	NbcYFZRNBAlkzJHWtKgpZQ	5	Within 1 hour, he called me back
e4FM01_iF_2LLN_yiaYcHA	-d1Sl2KzWUIBsXOxH_0jdQ	1	heard nothing...called twice...nothing

LSA Document Similarity

Due to space limitation, only the result of LSA is shown below. All review text are used to form an LSA space, its distance matrix. LSA space dimension is reduced to 47 terms. Below figure shows an example of how the review text are 'clustered', or similarity. Note that **The above two text reviews, although placed in the same topic by LDA, are very far apart by LSA calculation.** The two review documents are marked by two emphasized triangles in the figure below. All code are found on github.



Discussion

1. It is qualitatively possible to sort out the review text with appropriate topic assigned. For the Yelp Challenge Dataset, the terms *“best,friend, recommend, good, great, help”* are strongly associated with great reviews. However LDA fails to detect negation, as seen in the cases when two different reviews are both labeled as topic 8 (“call, back, one, well, feel”). Although the labeling is correct, a reviewer gives 5-star because the doctor ’call back promptly as promised“; whereas the other case, when a user”did not get the call back, hear nothing" from the office, he gave a poor review of 1-star.
2. LSA helps in finding similarities or dissimilarities in document text. The review text is used to form a corpus, from which document term matrix (dtm) was built with 78% sparsity. At this sparsity, the vocabulary for the corpus conprises of 47 terms of the LSA space. Within the LSA space, it uncovers that the two text reviews are quite far apart in LSA space; they are not ‘close’ enough in distance to be in the same cluser. *LDA and LSA can be used together providing complementary analysis tool to each other.*
3. LSA can be used for reducing dimensionality and served as a tool to perform similarities between documents, hence clustering.
4. Although LDA and LSA can complement each other in text mining, it requires a considerable human effort in reviewing the results; they do not provide a turn-key solution.
5. This work can be further improved in applications such as:

   a. Intelligently Categorize business type, categories that are currently in Yelp using LDA
   b. Forming a search engine where users can use natural language and better information retrieval. In this applicaiton, review text are grouped into its topics; after carefully review of the topics and its grouping, the text can be served as examples to train LSA in forming indices for thes topics (or categories). Then Yelpers can query these indices using natural language (as opposed to keyword search) to find business needed. The query is compared to the trained examples, the most *similar* reviews in latent or hidden meaning will be retrieved and present to users.

References

(1) D. M. Blei and J. D. Lfferty. A correlated topic model of Science. Annals of Applied Statistics, 1(1):17-35, 2007

(2) R. Arun, V. Suresh, C.E. Veni Mdhavan, and M.Narashima Murty. *On Finding the Natural Number of Topics with Laten Dirichelet Allocation: Some Observations, PAKDD 2010, Part I, LNAI 6118, 391-402.