Yelp Dataset Challenge - Topic Mining
Executive Summary

-Venkat Gokul Reddy Palampally
-Ravi Teja Mulpuri

Introduction:

With the internet being accessible to everyone, a large amount of text data is being produced every second which can be leveraged to summarize large amount of text data, segment the data into different topics, identify important characteristics of each topic.

Probabilistic Latent Semantic Analysis(pLSA) is an unsupervised algorithm that helps us to mine inherent topics in documents. It uses a probabilistic framework to extract the underlying topic and word distribution.

To understand how this algorithm works in real world we have used the dataset that is made available by Yelp. Yelp shares datasets regularly on their website to encourage innovative uses of their dataset. For this project, we have focused only on restaurant reviews from the Yelp dataset to identify the topics using pLSA algorithm. Our main idea is to understand how the topics are distributed over the restaurant reviews and the word distribution in each topic. For example, if one of the topic identified is about the customer service, then what are characteristics that people focus/expect in good or bad service.

Cornell University research study shows that 27% of new restaurants fail within the first year and nearly 60% close by third year. Michael Luca's study on the effect of star ratings on the restaurants revenue, found that a one-star increase in Yelp rating leads to a 5-9 percent increase in revenue. Determining the topic and word distribution in restaurant reviews could serve as a proxy for the factors of restaurant success. In general, this information has several potential business benefits for restaurant owners by providing insights into the areas they need to focus on to succeed.
In this application of pLSA, our training dataset is a corpus of restaurant review documents from Yelp, which are represented in the form of a document-term matrix (this representation indicates the number of times each word occurs in each document) after data cleaning. pLSA algorithm uses this document-term matrix to extract "topics" that are associated with the documents.

Data Cleaning:

Following steps are implemented on every restaurant review in the dataset.
Tokenizing: Tokenization converts a restaurant review into several words using whitespace between words and punctuations.
Stopping: This step removes the stop words available in the tokenized list. Stop words appear frequently while writing a review and don't offer significant value.

Stemming: Reviews have different forms of same word. For example, words such as *recommends, recommendation, recommend* can all be grouped as *recommend* to reduce the related forms of a word.

pLSA Algorithm:

The following assumptions are considered about documents and words in each document.

- Each document is an unordered collection of words, also called bag-of-words.
- Each word in a document is independent of others (unigram representation of words).
- Each document is a mixture of underlying k topics

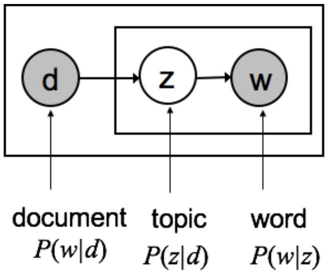

Figure 1 Topic and word distributions in a document

Each document, a restaurant review, can be considered as being a distribution of set of topics(k). Each topic has a probability of being associated with that document. Similarly, each topic has a word distribution with a probability of the word being linked to the topic.

$$p(w_i \mid d_j) = \sum_{k=1}^{K} p(z_k \mid d_j) p(w_i \mid z_k)$$

Probability of each word(i) belonging to a document(j) is calculated as the product of the probability of document(j) belonging to a topic(k) and the distribution of the word(i) in the topic(k) summed over all the topics.

Objective Function: The objective function calculates the probability of each word summed over all the words in a document and for all the documents present in the dataset. Objective of the pLSA algorithm is to maximize the log likelihood probability under a given set of probability constraints

$$\log p(C \mid \Lambda) = \sum_{d \in C} \sum_{w \in V} c(w,d) \log[\sum_{j=1}^{k} \Pi_{d,j} p(w \mid \theta_j)]$$

Constraints: For each topic and word distribution the sum of probabilities should be equal to 1.

- Sum of topic distribution probabilities for each document in the corpus

$$\sum_{j=1}^{k} \Pi_{d,j} = 1, \forall d \in C$$

- Sum of word distribution probabilities in a topic

$$\sum_{i=1}^{M} p(w_i \mid \theta_j) = 1, \forall j \in [1,k]$$

Iterative maximization: Expectation Maximization process is used at every iteration step to improve the objective value. Iteration process ends at a given number of steps or a specified tolerance between successive iterations

Expectation step calculates the probability that a word ‘w‘ occurring in a document ‘d‘ is explained by the topic ‘k‘ for a given parameter values

$$p(z_{d,w} = j) = \frac{\Pi_{d,j}^{(n)} p^{(n)}(w \mid \theta_j)}{\sum_{j'=1}^{k} \Pi_{d,j'}^{(n)} p^{(n)}(w \mid \theta_{j'})} \qquad z_{d,w} \in \{1,2,..., k\}$$

Maximization step calculates the parameter values based on the probability distribution of words from the Expectation step.

$$\Pi_{d,j}^{(n+1)} = \frac{\sum_{w \in V} c(w,d)\, p(z_{d,w} = j)}{\sum_{j'=1}^{k} \sum_{w \in V} c(w,d)\, p(z_{d,w} = j')} \qquad p^{(n+1)}(w \mid \theta_j) = \frac{\sum_{d \in C} c(w,d)\, p(z_{d,w} = j)}{\sum_{w' \in V} \sum_{d \in C} c(w',d)\, p(z_{d,w} = j')}$$

Conclusions:

pLSA algorithm was initialized to identify 'k' topics from restaurant reviews. The number of topics, in general, can be varied to identify the optimum value of the objective function. Due to the limitation in computing power and the volume of review of data, we have used two different number of topics and could identify the maximum objective value when the number of topics is set to 30 (k=30).

The results obtained from pLSA are visualized using an interactive html file to visualize the "topics" and respective word distributions.

The interactive plot depicts the Intertopic Distance Map and the Top-30 Most Salient Terms in each topic where the topics a represented as blue circles with numbers on them. The distance between two topics is a measure of the semantic relationship of the topics with each other.

There is also a flexibility in visualization plot to reduce the probability of words that have a very high document frequency to understand the topics in terms of unique words.

For instance, the topic 1 word distribution looks like "waitress, ask, took, care, server" which shows that the topic is about service in the restaurant. Similarly, topic 2 distribution is "call, counter, management, employee" which shows that the topic is about management and in the picture both topic 1 and topic 2 are both located close to each other which is intuitively true as both relate to each other., Now, let's take topic 23's word distribution, it looks like "breakfast, bagel, coffee, pancake" which says that the topic is about a breakfast place. It is located quite far from topics 1 and 2 which again is intuitively true.

The applications of the such topic mining are numerous. This algorithm can be applied to webpages to identify the relevant topics associated with the web page and then advertisements can be published on those web pages.

Another application is in the field of medicine where millions of medical reports can be scanned and tagged with topics so that doctors can retrieve to them efficiently in time of need.

Topic mining is an important area in the field of Natural Language Processing and algorithms like pLSA, LSA, LDA can be used to teach machines to process laborious tasks that have a lot of significance in our day to day lives.