Final Report
CSE -5334 Project

Help new business to develop faster by analyzing trends, discovering good and bad in already well-established similar ventures using yelp data.

Team members: Jvalin Dave, Mudit Pradhan

Objective and motivations:

Owners of the new ventures such as hotels, restaurants or home services need a comprehensive study on latest trends in the location. for e.g. if we take restaurant business, restaurant types such as Asian, Lebanese or Mexican can be trending in area based on density of people from one origin in that location. It is important to know the latest trends for which people really care about in a particular business, and this has to be overall and also location specific. This would help him establish business with the appropriate product and services. Apart from this, the new owners would also want to know good and bad qualities/ products and food items in already established similar ventures in the same location, this would help them develop faster and also lessen the risk of failure. for e.g. if a new branch of Mc Donald's needs to be opened some another country like U.K., they need to find out the what the local established burger restaurants are doing and also what people looking for in terms of services and product in the reviews.

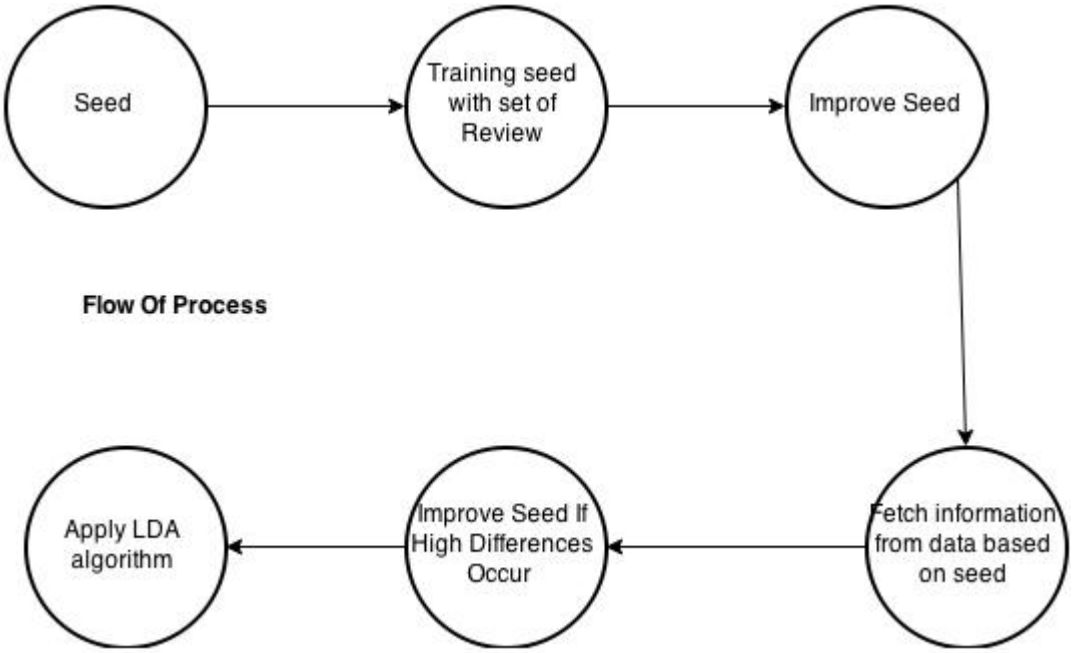
Data Mining tasks/ Analysis tasks:

One of the major data mining tasks that we perform in this project is information extraction. The data set available from yelp has a number of reviews, which contain a lot of information about a particular location. Applying data mining tasks on the reviews from on one location could certainly result into discovery of the new features and information. In this task we planned to fetch the local flavors, food items and products in general. It could be also a comprehensive list of such items where each word describes a different item. Fetching Information structures [1] could help in this case, for this we read several reviews and try to find out some common structures (if existing) with the food items/flavors whenever they are mentioned in any context. It was quite challenging to frame some initial sentence structures as the review data is plain text with each review coming from a different user, moreover determining rules or structures over a large context over large amount of items seemed to be more challenging and we were not sure if we could have good results from that. The next approach was to make a self-learning algorithm which would initially start from a food, items, flavors supervised from the programmer. The system then learns more words matching with the initial seed [1]. This involves training from an initial dataset and information extraction from the entire data reviews using the trained seed on the reviews for restaurants in a particular city.

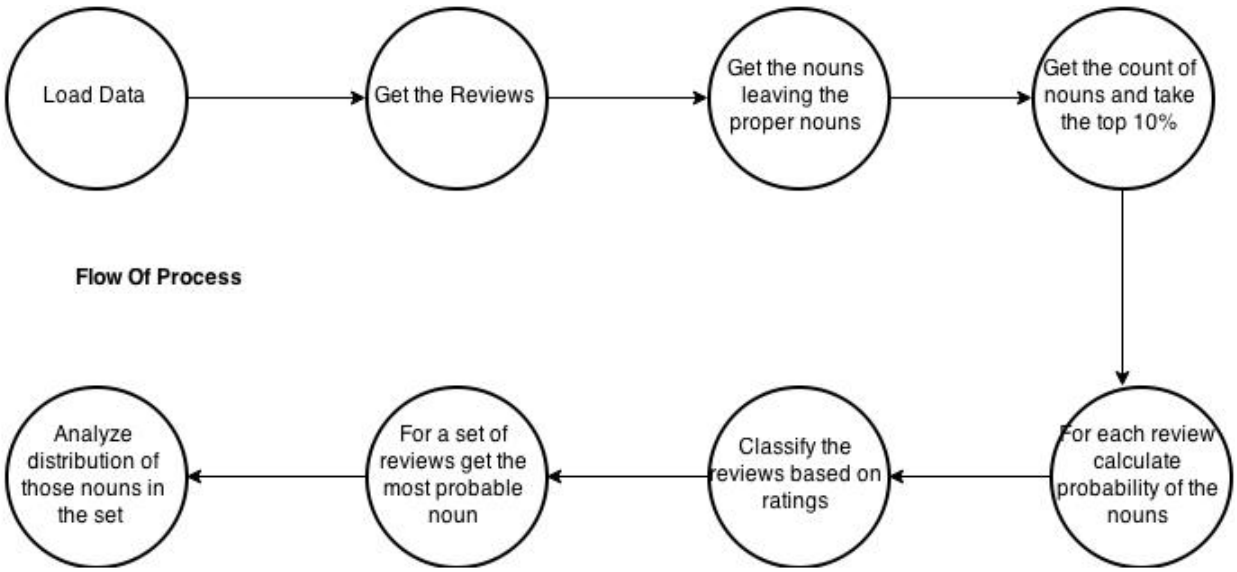
Another data-mining task we have tackled is generating trends from the user reviews and knowing the most trending restaurant type in that particular city. For this we have used the approach of finding the cosine similarity on a review and the trained sees [1]. Again the trained seed contains good amount of knowledge about a particular food type for restaurants and can be used to find out with the cosine similarity measures applied to get the number of reviewers who have actively mentioned about a restaurant type in their reviews. The restaurant types could be Mexican, Asian or American based on the type of food they serve.

Final Report
CSE -5334 Project

Design for Trend analysis and information extraction:



Design for Term Distribution:



Final Report
CSE -5334 Project

Design of Methods for trend analysis and information extraction:
We have used LDA algorithm to generate the subtopics being mentioned in a given set of reviews.
Gensim LDA package has been used to get the set of subtopics.
Gensim LDA algorithm is applied over only those reviews which are potentially higher in content in terms of food items, products and flavors.
The design of the algorithm flows in the following manner:
1) fetch the reviews for restaurants in a particular City.
2) Choose a seed [1], which has high information about the food items which are supervised by the the programmer.
3) Train the Seed [1] using the set of reviews from the same city to fetch more information about the local items.
4) Apply the seed [1] which is now improved and having more information about the local flavors and products to fetch the filtered reviews which potentially contain high source of knowledge about the local items.
5) Train the seed [1] again in real time if a highly knowledgeable source is encountered and use the trained seed from the next time.
6) Apply subtopic modelling to fetch the most important subtopics from the filtered data set, which means applying topic modelling algorithm only on the desired dataset to fetch more information from that dataset.
7) Again filter the outcome to keep only the desired outcome and discard the subtopics generated out of context.

For Trend analysis following approach has been used:
Analyzing trends is also based on the similarity between the trained seed and the reviews. It matches the reviews for a particular restaurant type such as Mexican, Asian and Indian. When a review talks about particular food, it could be inference that the reviewer cares about the food type and it is a topic of interest for him. When a group of user talk about a food type, it can be inference that the more the user, the more population goes is interested in that food type in a location. Finding the similarity measures of review with a specific food type would help to find population amount which has mentioned about the food type in review.

Implementation Details for term distribution:

Programming has been done in IPython Notebook using the libraries that I have used are nltk, json, matplot, and other modules of nltk. The process that I am following is better explained in the diagram.

Final Report
CSE -5334 Project

Loading the files: The files were loaded using json.load function. Then I loaded the review attribute values, followed by business attribute values. Then stored the values in different dictionaries for different attribute of a particular file. From dictionaries, I stored them in different lists in order to access them easily.

Joining the Files to get Relevant Data:

I extracted the reviews of restaurants in a particular city. The city that is in the code is 'Charlotte' which can be changed in the code and it will work. Then, I joined the two files of review and business, using business_id as primary key and got the reviews and star ratings for each of them.

Tokenization of Words in Reviews:

I tokenized the words I review using regular expression dictionary of python. After tokenizing the words I removed the stop words from the tokenized reviews.

Extraction of Nouns:

After tokenizing, I took universal nouns using nltk Treebank tagged words. I compared the words in those with that in reviews and the ones that matched, took them in a new list. Those were the nouns.

In nouns, all kinds of nouns come, so the next task was to remove at least proper nouns that is name of places and food items and other things. Names were not part of this design so removing them was important. I applied the concept of tagging using part of speech tagger in nltk and removed the words that were tagged as proper noun.

Counts of Nouns:

The next part was to get count of the nouns in all the reviews in order to get the more frequent nouns occurring in the reviews. It was also used to calculate prior probabilities of those nouns. The prior probabilities would be used to determine their importance after classification of reviews. The nouns having count of more than 4% count of that of total were extracted and used for further evaluation.

Classification Of reviews:

The reviews were classified on the basis of star rating i.e. in 5 sets with rating of 1, 2, 3, 4 and 5. Based on the frequencies, I chose the worst and best ones in order to get good results.

Determination of Distribution of Nouns:

Again we find the count of nouns in the classified reviews of rating 1 and rating 5. For a sequence of reviews, multiply it with the prior probability and get the review with maximum probability. The distribution in that particular review is taken by occurrence of terms in review.

Based on its results pie chart is plotted.

Final Report
CSE -5334 Project

Implementation of methods for trend analysis and information extraction:

1)training of the seed[1]:

Seed ==>trained with a set of reviews ==> improved seed ==> fetch information from the whole data set using seed ==> improve seed if high potential data encountered ==> apply LDA algorithm to generate subtopics from the filtered data

Cosine similarity has been is used to find the similarity between the text reviews and s)seed[1], nltk has been used to keep the seed clean of the unwanted words such as verbs and adjectives

2) filtering the reviews to get the specific reviews.

Use cosine similarity to find the similarity between the reviews and seed [1], then if cosine similarity is above a certain threshold, keep such reviews in filtered dataset.

Results and evaluations:

We found a list of local flavors and food items, which was much more enhanced from the list of initial seed that we used to generate this model. Our idea was to fetch more flavors and food items based on a set of initial items, which we could achieve with this model of supervised learning. The first program fetches food/flavors for the entire restaurant community and another program generates trends for local restaurants and also the food items/flavors being sold for each restaurant type respectively.

From trend analysis we noticed the difference in preferences of the local people of two different cities. Small cities have a huge difference in food preferences as compared to big cities.

We have also mined the 5 star reviews and 1 star reviews for finding out the difference in the terms, these terms carry importance. we could find that 1 star ratings differs from 5 star ratings in that, the terms such as "time" differs by 1%, which could be inferred as the "time “is an important key word in determine the service provided in a particular restaurant. Another word which holds importance is "place", which could be inferred to have problems with the place for ambience or people could like it for some other reason.

Presentation/Visualization of the Outcome:

1) the food items mined form the text reviews are presented as it is, like a list of important words. first list contains important local flavors covering all the restaurants types, the second list is a more comprehensive list specific to the restaurant type such as Mexican, American or Asian.

2) Trends are visualized on the time graph for different cities for the year 2014. Huge differences have been noticed between big cities and small cities.

3) Pie chart is used to visualize the term distribution.

Final Report
CSE -5334 Project

Instructions for running few programs:

1) dividejson.py produces the food/flavor list local to an area
2) trend_train.py produces trends w.r.t local restaurants such as mexican, asian, american.
3) Other text files have to be downloaded to the same folder as the programs before running the programs.

Project Web link: http://dtyelp.github.io/5334Project

References:

1. http://www.mathcs.emory.edu/~eugene/papers/dl00.pdf
2. Stack Overflow