CS598 Data Mining Capstone Project Final Report

Ashish Kumar

Summary of Project Activities

The data mining capstone project provided us opportunity to use the knowledge and skills of data mining to solve a real world problem. We were also required to participate in online seminar to present a text analytics/mining based research paper to our fellow students and teaching staff.

Project Work

The capstone project was to analyze and mine a large restaurant review dataset offered by Yelp. The project was divided into six distinct tasks and we were required to submit and peer review report for each task. At a high level, the deliverables were as follows:

1. Cuisine map construction
2. Discovery of popular dishes for a cuisine
3. Opinion visualization
4. Discover popular dishes for a given cuisine type
5. Recommendation of restaurants based on dish type instead of cuisine type
6. Hygiene prediction based on the restaurant reviews

Algorithms, Libraries and Results

In this section, I have outlined the libraries and algorithms used for various tasks. The results and the corresponding usefulness are discussed in the Highlights section below.

Task 1

I implemented topic modelling to discover main topics and top 10 to 15 words in each topic. I used Gensim's LDA library and Sklearn's Gensim library for this exercise. The output was set of visual diagram for topics discovered by the Sklearn and Gensim libraries. We compared the the performance of two models and results were different.

Task 2

I used leverage TF/IDF for vectorization and added sublinear option to further improve the quality of cuisine maps. The output of task 2 was set of visuals of cuisine maps. The chart was created for different algorithms such as TF without IDF, TF and IDF and LDA topic modeling. The visuals showed better clustering of cuisine types for the LDA topic modeling algorithm. As expected, the clustering was relatively poor when I used TF without IDF. I also ran clustering algorithm on two similarity matrix. I was able to discover pattern of cuisine similarities even when number of cluster varied.

Task 3

I used unsupervised topic mining methods: TopMine and NMF library to improve the dish names for a cuisine type. I ran TopMine and NMF libraries to discover topics from cuisine data. I was able to use the output of these algorithms to identify positive and negative samples in the provided label file and improve the accuracy by increasing the iteration count. I was also successful in identifying and adding new dishes names for a given cuisine type.

Task 4 and 5

I created TF/IDF vector for the review corpus. The vectors became the backbone for the follow up calculations such as average review. I was able to identify popular dishes by calculating frequency of dish reviews and joining it with average of star ratings of the corresponding reviews. I also created a restaurant recommendation algorithm for a given dish instead of generic cuisine type. The input for algorithm were: ids and names of restaurants serving the dish, average review ratings for a given dish. I extrapolated the sentiment of the review through rating provided by the reviewer. The output was a bar chart showcasing top 40 restaurants for a given dish name.

Task 6

I used various libraries to train the model to predict the outcome of hygiene test. I used NaiveBayes, Decision Tree and MLPClassifier. I used unigram and bigram methods for the CountVectorizer and TfIdfVectorizer libraries. The accuracy varied from 43-61% for various algorithms when just review text was used as input data. After adding features such as zipcode and average review, the accuracy improved and varied from 68 - 84%, a significant improvement.

Online seminar

I did the presentation for research paper titled **Demographics, Weather and online reviews: A study of restaurant recommendations** by Saeideh Bakhshi, Partha Kanuparthy and Eric Gilbert.

Presentation is available at https://youtu.be/xdqYKvJMJ1U

The online seminar was a great opportunity to learn about how to read, digest and dissect the research outlined in a research paper. It took me few iterations but in the end, I was able to appreciate the importance of research done but more importantly, I was able to critically analyze the findings and offer my own critic.

Highlights

All the tasks outlined in the Project work section enabled me to put my skills to practical use. I was able to experiment with different libraries and tools to solve a given problem. I gained much deeper understanding about the strengths and weaknesses of the tools and libraries used for a given task.

Usefulness of Results

Topic discovery in task 1 was useful because it can help an analyst identify important keywords and extrapolate them to identify patterns. A topic with related words such as fries, burger, bacon, pizza etc will strongly is a strong indicator of presence of american food restaurants and significant consumer base eating at these restaurants.

Using kmean and spectral clustering produced very helpful visual charts depicting similarity of the cuisines. This data can be extremely useful for demographic analysis and it can be valuable input to predict success of a potentially new restaurant.

Topic mining models helped discover new dishes in the review corpus. This data can be used to identify potential popular dishes and restaurants can use it to differentiate themselves from the competition.

An ability to recommend a restaurant for a specific dish type can be an incredibly powerful tool for personalized recommendation. A highly personalized and accurate recommender tool can be an incredibly powerful utility for consumers. At the same time, it can offer unique insights to restaurant owners and marketing teams in the restaurant industry.

An ability to predict hygiene test results based on other features of restaurants will be a highly valued by anyone associated with the restaurant industry. This prediction model can also offer additional insights about the restaurants such as demographic pattern and customer base.

To summarize, each task produced a result relevant for either the restaurant owners, restaurant industry affiliates and most importantly the customers eating out at these restaurants.

Novelty of Exploration

Task 1

In the second part of the task, I focused on restaurant businesses with 3 and 4 star rating only. This modified strategy improved the quality of the results. Using D3 library to visualize the topics was the highlight of the first week’s task. The radial dial visualization was a powerful way to represent data and help user grasp the core finding with just a glance.

Task 2

The visual representation of the clustering of the similar cuisine map created during week 2 was a unique idea. The visual similarity matrix is a great tool for any use to recognize similar cuisines and critically interpret the results. The visualization also showcased the strength of clustering and helped picked the best clustering algorithm.

Task 3

I would have never guessed that one could use topic mining algorithms to discover new dishes from the review text corpus. For instance, this approach can be used to identify and co-relate multiple diseases to a specific clinical research, even if the summary cities one of two diseases.

Task 4 and 5

The strategy of creating a restaurant recommendation system based on a dish type by combining restaurant id, restaurant name, average review rating and sentiments was very effective. The algorithm for the dish type based restaurant recommendation is unique and it showcased intuitive results in the first attempt.

Task 6

The use of recurrent neural networks (RNN) to predict hygiene outcome for a restaurant is a novel idea. Again, it is against normal intuition but it produced reasonably good results.

Contribution of New Knowledge

Robust Data pipelines A robust data pipeline is extremely important for the large size text/data analytics projects.

Dish type Restaurant recommendation This project required us to implement an algorithm to create restaurant recommender system based on a dish type, instead of a cuisine type. Creating and implementing the algorithm was a significant contribution to new knowledge.

Feature Engineering We were required to improve hygiene prediction results by expanding features in the datasets. Identifying relevant features is an art as much as an science. I was able to dramatically improve the accuracy by picking right features.

Potential Enhancements

If I had more time, I would make following enhancements in order to build a production grade system capable helping people make dining decisions.

- Add a web based User interface and host it in cloud where anyone is able to access this application
- Create batch jobs to mine the data on regular basis and continuously update the mined results.
- Create an API to search the mined results in real time in response to user query.
- Create a web application that leverages API to get results in response to a user query and show the results to user in a real time and application must have millisecond response time.
- Filter, sort and store data for different geographic regions in different databases. If a user is looking for recommendation in San Francisco, it makes no sense to search database for the entire USA.
- Create a mechanism to continuously collect review data and sunset data older than a certain threshold. I would also like to create an algorithm that gives more weight to recent reviews and reduces influence of older reviews. A recent negative review should matter more than 10 year old positive review for a restaurant or a cuisine type.

Summary

This report outlines the tasks of the capstone project including the high level implementation details, usefulness of results, novelty of exploration and potential enhancements. The capstone project afforded me a great opportunity to use my text mining and analytics knowledge to solve a real world problem. I do wish that duration was 16 weeks and it would have allowed me to add more features and create a robust solution for general use.