Data Mining Specialization
Capstone Project - Final Report

Boudjema DEBBAH

This Data Mining Project allowed me to apply different knowledges learned during all those weeks and got into several courses such as Pattern Discovery, Clustering, Text Retrieval, Text Mining, Visualization, and finally complete those courses by this Data mining project to complete this Specialization. The final part of this Specialization, is very interesting to enable us to apply skills from the Data Mining Specialization to solve real-world data mining problems.

This Final Report is about the Project itself, and reflect on what I was able to accomplish and learn through working on all these tasks, to integrate all my work, and to highlight some of the most interesting/useful data mining results I have produced as well as any new knowledge I have discovered through experimentation.

1. Tasks & usefulness of the results.

From this Yelp reviews, goal was to understand what people care about when they go to a restaurant. This is very useful information, for many stake holders, and first for restaurants owners themselves. Depending on their business models, they would focus on how they can improve their businesses based on customers experiences. Mining thousands of data, can show how in particular they should focus on to avoid negative feedbacks (cooking technics, or long waiting times, etc…), or even how they can improve particular dishes they serve to customers.
Another interesting outcome for this information is how results can be useful for restaurant recommendations, based on a particular cuisine, or dish form this cuisine.

1. Exploring Yelp data set to understand the data, their characteristics, and how they look like, was a really good experience to start getting into how we can extract all the different topics from the reviews, and then **analyze them using topic models such as LDA & LSA** with different settings to get the best from the data. I applied those models to the text reviews to identify what people have talked about when they had their own restaurants experiences. Getting a particular visualization to show 'Highs' and 'Lows' rating over those restaurants, and finally provide insights to whoever may be interested in such information.
2. Mapping of cuisines is another technique to apply against this Yelp data to allow anybody to explore cuisines, either by experimenting/trying new cuisines experience, or within the well-known one if they want to keep up with their favorite one. Here we used variations of the **TF-IDF weighting scheme along with cosine distance** technique to represent similarity between cuisines. In the result visualization, we could observe several clusters that describe meaningful relations among cuisines (i.e : European cuisines – Mediterranean vs Italian).
3. Dish ranking and restaurant recommendations was an obvious area to explore using data mining techniques. Goal was to identify, from a particular cuisine, and know beforehand, types of dishes that are available for a cuisine, and then apply a particular ranking…But the main searched outcome here, was to answer a simple question – What is the best way of ranking dishes for a cuisine? … Here I used **pattern & phrase mining techniques** to obtain a result. **Using Segphrase & TopMine tools, combined, showed how to can get into mining quality phrases in massive text data**. From this mining, this led us to recommend good restaurants to those who would like to try one or more dishes in a cuisine. Here I have to find out **visualization strategies** to get this done using programming tool to extract the data, and then visualize in a such way that this can be easily reused with particular technology such as search engines.

4. Recommending a restaurant based on reviews analysis, and ranking is useful, but adding a particular detail to allow a specific identification such as hygiene inspection can be even better to add to recommendations (or dissuasion to eat there). However, such information is not always available and the only way to get into such information is to **make predictions based on a sample of data**. In that case, we used data mining to solve a real-world problem. Finally, to get into prediction techniques, I had to find out what would be the best classifier to perform on training/testing set of data. I learnt from this task to **choose the best way from text processing to learning algorithm experimentation.** I made a particular comparison between different learning algorithms such as Random Forest, SVM and Logistic regression, using different text representation of the initial dataset, separating the data using a training and a test set, and also adding additional features available such as ranking, or review number…, to enhance learning process. Such prediction could be finally used for different stake holders such as anybody who search such details to select a restaurant in a particular city, but also for any web sites who want to give additional services based on hygiene data, and finally for a given government who want to understand how restaurants deal with hygiene regulations.

2. Exploration methodology & new knowledge.

I have learnt to apply a set of tools for text mining along this project, and improve my usage of Python programming (which is one of the nice outcomes !!!). Added to this I tested different sets of parameters and data-algorithm combinations to seek for optimal results. But as for any experimentations, methodology is the key point to move on to a result (relevant or not).

First, manipulating data can be confusing if we don't have a clear understanding about what we want to achieve. Here are few approaches I have applied:

- Topic modeling: From a straightforward way (task1) to more complex way for dimensional reduction is LDA/LSA for example, or compute posterior probabilities.
- Representation of corpus using bag-of-words, and applying topic modeling to get a conceptual representation of the documents in a reduced space. And apply the best visualization to illustrate the results.
- Application of topic modeling in a such way that I got into finding out what would be the most probable meaning of a given term that related to the most probable topic of particular document.
- Remove edges among documents to be able to find similarities, and significant relations.
- Experiment differently the tasks descriptions to represent documents in a corpus, and then study the relationships among groups. And finally find the best way to visualize this finding.

Then, mining data to discover relationships, or use collaborative filters to infer particular characteristics of the data. For example, in task 4 & 5, we used a user-based collaborative filter for a restaurant recommendation. But as I concluded in this task, we can extend our list or recommended restaurants using a user-based collaborative filter. And we know that this information is also available on Yelp for some restaurants. Therefore, in a real application we could further filter the ranking to keep only those restaurants that cook the dish, or add available restaurants menus from this Yelp data.

Finally, the prediction task over restaurant data for hygiene inspection was really interesting and new to me. But here the goal was to combine all the understanding of the previous results, along with skills & learning we got during the course to achieve this task, and ultimately be able to compute a classifier that predicts an event. Two important aspects of this were new to me, first one was about organizing documents & terms in a such way to be able to apply a supervised machine learning classification task, and second one was to be able to compute this efficiently, and using the best-in-class tools.

For each task, I have to keep in a mind that visualization is key to let anyone understand the outcomes of the assigned task, and also visualization is part of the data mining activity to let anybody understand the concept we wanted to share, coming from data terms considered as individual structures.

3. Novelty of Exploration

Along with tasks required for this overall project, we were given with some standard tools. For the most of this project, I have not used all of them, but in some case, looked for equivalent solutions within python environment (i.e.: Python package for classifier training). In addition to that, I have applied different approaches & methodologies from what was suggested in the different task's descriptions.

I will provide two different examples of how this project brought a slightly different way of using & exploring learning and skills.

1. Using Latent Dirichlet allocation (LDA) as a method for fitting a topic model is something quite normal and universal. Here I think I mainly used LDA tool for dimensionality reduction and feature creation, this not really a new idea/concept, but enough for me for such application of topic modeling. Studying relationships among groups was enabled, within LDA modeling representations, by the efficientness of such topic modeling of documents in a corpus.
2. During the search of similarity of cuisines, I had a different approach for visualization of the resulting similarity matrix. This was new to me in term of representing a sparse graph showing relationships from this similarity matrix not really correlated to the initial graph. What was new here and get me into a different exploration, is to find out how to represent nodes & edges weights representation of this matrix. Where nodes were documents (or reviews), and edges similarities between those documents, and to achieve this task, challenge was to finally keep only significant similarities after removing computed edges.

External contributions & Project references.

Literatures: (mentioned some of the references)

- Phan XH, Nguyen LM, Horiguchi S (2008). "Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-Scale Data Collections." In Proceedings of the 17th International World Wide Web Conference (WWW 2008), pp. 91–100. Beijing, China.
- Blei DM, Lafferty JD (2007). "A Correlated Topic Model of Science." The Annals of Applied Statistics, 1(1), 17–35.
- Lv, Y., Zhai, C., 2011. Lower-bounding term frequency normalization, in:. Presented at the Proceedings of the 20th ACM international conference on Information and knowledge management, ACM, pp. 7–16.
- Serrano, M.A., Boguna, M., Vespignani, A., 2009. Extracting the multiscale backbone of complex weighted networks. arXiv. doi:10.1073/pnas.0808904106
- Raghavan, U.N., Albert, R., Kumara, S., 2007. Near linear time algorithm to detect community structures in large-scale networks. Phys. Rev. E 76, 036106.
- Crain, S.P., Zhou, K., Yang, S.-H., Zha, H., 2012. Dimensionality Reduction and Topic Modeling: From Latent Semantic Indexing to Latent Dirichlet Allocation and Beyond, in: Aggarwal, C.C., Zhai, C. (Eds.), Mining Text Data. Springer US, Boston, MA, pp. 129–161. doi:10.1007/978-1-4614-3223-4_5
- Chang, J., Blei, D.M., 2009. Relational topic models for document networks, in:. Presented at the International Conference on Artificial Intelligence and Statistics, pp. 81–88.

Programming tools & knowledge sharing :

https://scikit-learn.org/ : Machine learning in Python

https://www.researchgate.net : recommenderlab: A Framework for Developing and Testing Recommendation Algorithms

https://towardsdatascience.com/ : Medium publication sharing concepts, ideas, and codes.