**Project 5 – Advanced Data Mining Applications**

**CS548 Knowledge Discovery and Data Mining - Fall 2016**

**Prof. Carolina Ruiz**

**Student: Mu Niu**

|  |  |
| --- | --- |
| Description of the particular problem within the selected data mining topic to be addressed in this project | /15 |
| Description of the approach used in this project to tackle the above problem | /25 |
| Description of the dataset selected | /15 |
| Appropriateness of the dataset selected with respect to this topic/problem | /10 |
| Guiding questions | /10 |
| Preprocessing | /10 |
| **Experiments:**   * Sufficient & coherent | /25 |
| * Objectives, Data, Additional Pre/Post-processing | /20 |
| * Presentation of results | /20 |
| * Analysis of results | /30 |
| Overall discussion, comparisons, and conclusions | /20 |
| TOTAL | /200 |

Total Written Report: \_\_\_\_\_\_\_\_\_\_\_\_\_\_/200 = \_\_\_\_\_\_\_\_\_\_\_/100

Class Presentation: \_\_\_\_\_\_\_\_\_\_\_/100

Class participation during project presentation: \_\_\_\_\_\_\_\_\_\_\_/100

*Do not exceed the given page limits for this written report*

**Topic: Text mining <at most 1 page>**

1. **Description of the particular problem within the selected data mining topic to be addressed in this project:**

Since the dataset is huge and contains just text, it would be difficult to come to any conclusion based on observation. So text mining technics are needed to make conclusion or prediction on both statistical way and data mining perspective. Transfer them into vectors can be a good way to evaluate and analyze dataset in a quantity way.

1. **Description of the approach used in this project to tackle the above problem:**

In order to find pattern and to predict, I used text to vector technic to do text vectorization, and then using PCA to reduce dimension into 2 before fitting K-means clustering.

I also used “bag of word” technic in this project. I built a dictionary of all material foods they can use from dataset, if the new recipe includes all material in one country’s dictionary, I can predict this recipe can cook a particular kind of food. I also use the same technic plus some statistical method to find out the most popular material to cook with.

In the third group of experiences are made for classification after vectored dataset.

1. **Dataset Name: What’s Cooking**
2. **Where found:** [**https://www.kaggle.com/c/whats-cooking**](https://www.kaggle.com/c/whats-cooking)
3. **Dataset Description:**

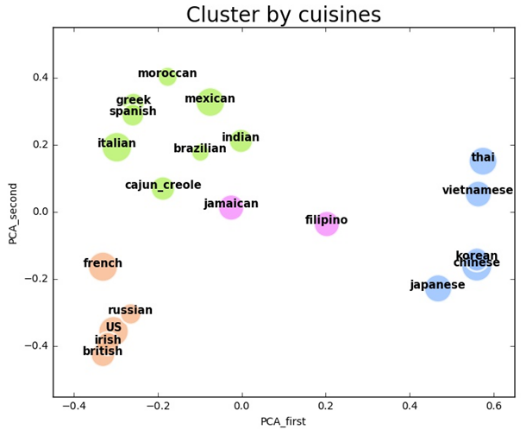
The dataset has train set and test set to be different recipe. Train set contains 39774 instances with recipe country, id and food material needed. The test set have in total 9944 instances with just id and food material needed. The goal of this kaggle project is to predict country where recipes come from. In my project, in order to evaluate performance of my experiments, I just used the train data. I separated the train dataset to have 90% (35796) instances to be train set and 10% (3978) instances to be test set.

1. **Initial data preprocessing, if any:**

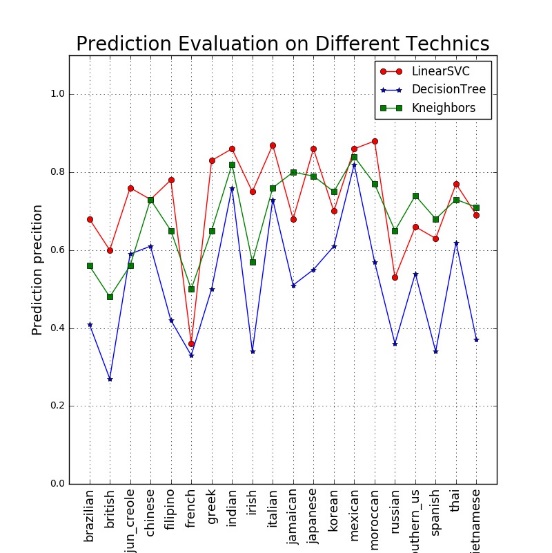
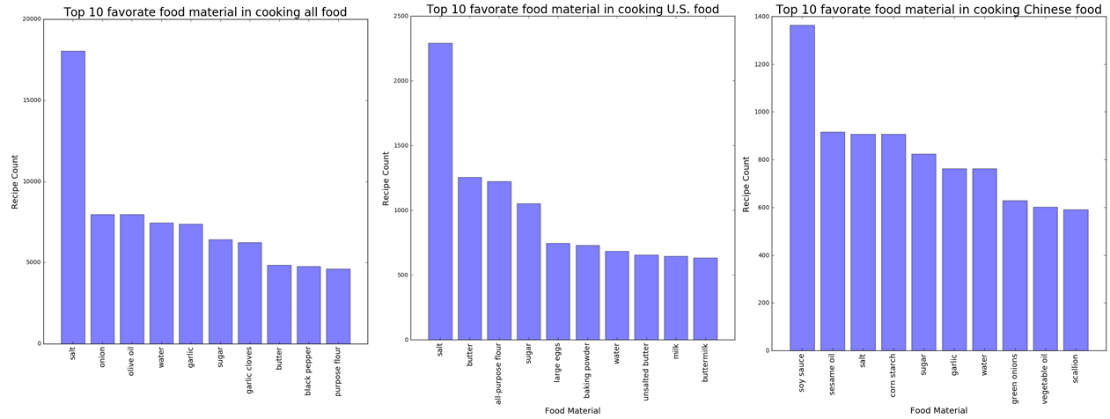
The dataset itself in clean, complete already without any missing values. So I don’t need to clean data. But JSON form didn’t fit well in all my experiments, so firstly I have to remove the brackets outside food material for doing further experiences with proper data structure. After that, I grouped data by country to have a dictionary for further explorations. For having other experiments, I also vectored data using TF-IDF approach.

1. **Three Guiding Questions about the dataset domain:**
2. **Is any countries share food materials in cooking?**
3. **What’s the most common cooking material?**
4. **Which country’ s food will be made based on materials customers have?**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Summary of Experiments.** *At most 2 page.* | | | | | | | |
| **Tool** | **# of**  **Question** | **Pre-process** | **Mining**  **Technique** | **Results** | **Time**  **taken** | **Evaluation** | **Observations about experiment**  **Observations about visualization**  **Interpretation results** |
| Python  Weka | 1 | Vectorization | PCA  K-means | * China and Korea, * US, Ireland and British, * Greece and Spain   have similar food structure. | 30s | SSE = 0.335 | After K-means clustering, I found these country have similar cooking material, which can lead to wrong predictions later. Jamaica and Philippines have quite unique materials to cook. |
| 106s | 65.3% accuracy |
| Python | 2 | Dictionary establish | Bag of words | For all countries, the top 10 popular foods are:  salt, onion, olive oil, water, garlic, sugar, garlic cloves, butter, ground black pepper, all-purpose flour, pepper | 263s |  | From the result of these experiment, I can figure out which kind of foods can be essential in a particular country. The guiding question is designed for this. For example, if I were a vegetable and flavoring seller, I would definitely stock based on both this results and the population structure of local area.  An interesting finding is that US have very similar popular food materials compared with the whole dataset, which can be an proof of cultural diversity. |
| Bag of words | For China, the top 10 popular foods are:  soy sauce, sesame oil, salt, corn starch, sugar, garlic, water, green onions, vegetable oil and scallion | 58s |  |
| Bag of words | For US, the top 10 popular foods are:  salt, butter, all-purpose flour, sugar, large eggs, baking powder, water, unsalted butter, milk, buttermilk | 54s |  |
| Python | 3 | Vectorization | Linear SVC | By using TF-IDF method, I transformed text information in to a “representative matrix”. The result would be a list of prediction on countries. Evaluation is made by comparing precision with real countries from test set. | 37s | Precision = 0.77 | This series of experiment explored which method would be more suitable for predicting target attribute. Based on precision score, the most suitable way would be linear support vector classification. K nearest neighbor classification could also be used but decision tree would be a bad idea. |
| Decision tree | 24s | Precision =  0.62 |
| K neighbors | 308s | Precision =  0.73 |
|  | Dictionary establish | Bag of word | If all food material fit in dictionary in a particular country, then I make the prediction. | 103s |  | The correct results lie in one of the choices, but there’s no further “narrow down” can be applied so that it would perform better if the recipe is long but |

**Analysis of Results: (at most 1 page)** 1. Analyze the effect of varying parameters/experimental settings on the results. 2. Analyze the results from the point of view of the Domain, and discuss the answers that the experiments provided to your guiding questions. 3. Include and explain (some of) the best / most interesting results you obtained in your experiments. 4. Include visualizations.

For guiding question 1, I calculated SSEs with different number of clusters, the “elbow” is at 4 clusters. So as we can see in plot, China and Korea, US, Ireland and British, Greece and Spain have similar food structure. Assuming data domain to be global food structure researcher, the results would be quite direct. If data domain to be vegetable suppliers, they can decided based on the result on which kind of customers can be potential customers with same kinds of stocks. I would have thought Russia have more different similar foods with US, Ireland and British because of the geographical distance. But they are unexpectedly similar.

For guiding question 2, I counted the most popular food materials among all dataset and by country. I just plot China and US as examples to make points. For data domain as food material suppliers, this results can show them what to stock for whom to buy. The most popular flavoring in the whole world is salt, as it is in US, but the most popular flavoring in China is soy sauce, which is really matching my cooking preference. Dairy products are widely used in global cooking instead of in China.

For guiding question 3, I used 3 classification technics, linear support vector K nearest neighbors and decision tree. As it shows in figure on the right, linear SVC model gives the best prediction. Decision tree gives the worst classification result among the three. I thought Jamaica and Philippines would have better predict accuracy since in cuisine clustering experience, they are close to no one and uniqueness should be a good thing in prediction, but the result showed otherwise. Since Chinese food and Korean food are too similar, wrong predictions are often made between these two.

**Summary of what you learned in this project:**

The most important thing I learn from this project is to combine text mining technics with other data mining technics, like classification, clustering and even regression, which is not very useful in this dataset but very import generally speaking. I found how important it is in being interested in data. And doing this project along make me realized how valuable others’ thoughts are in a team.