**Project 5 – Advanced Data Mining Applications**

**CS548 / BCB503 Knowledge Discovery and Data Mining - Fall 2017**

**Prof. Carolina Ruiz**

**Student: Tianyu Wu**

|  |  |
| --- | --- |
| 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.  *All data mining techniques you use in this project for pre-processing, mining and evaluation must have been covered in class during this semester.* | /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:**

Datasets sometimes contains large amount of texts information that hard to KDD, even though human can read and understand them. Also, some datasets contains texts that would take long time for human reading and translating which is not efficient for human resources. So text mining technics are nesessary for this senario to make conclusions or find patterns.

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

In this dataset, the description are human sommeliers’s reports which decribe wines taste, look, feel, etc, and the machine cannot understand. I used vetorizer technic with tf-idf term weighting in this project, to find best description of varieties of wines that machine can understand and analyse, also to find some other interesting patterns in this dataset.

1. **Dataset Name: Wine Reviews**
2. **Where found: Kaggle**
3. **Dataset Description:**

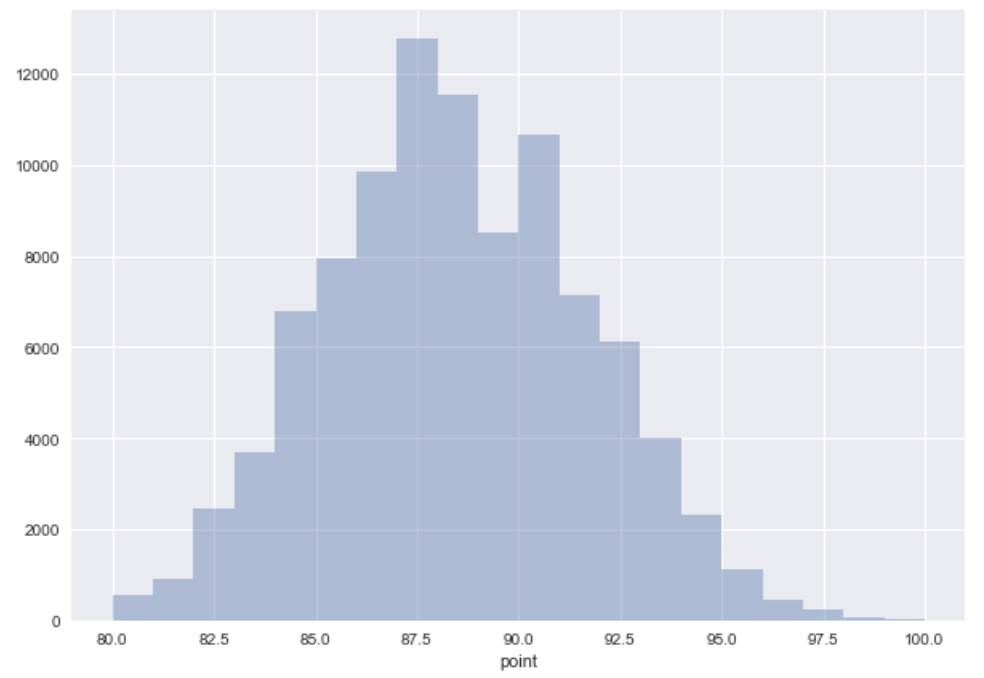
This wine reviews dataset was scraped from WineEnthusiast during the week of June 15th 2017, it contains 150930 instances along with 10 attributes. The data consists of 10 fields: ***Country***(the wine is made from), ***Description***(from sommeliers describing taste, smell, look, feel, etc), ***Designation***(where the grapes are from), ***Points***(rated on a scale of 1-100), ***Price***(cost of one bottle), ***Province***(province or state that made the wine), ***Region\_1***(wine growing area), ***Region\_2***(more specific region), ***Variety***(the type of grapes), and ***Winery***(winey made the wine).

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

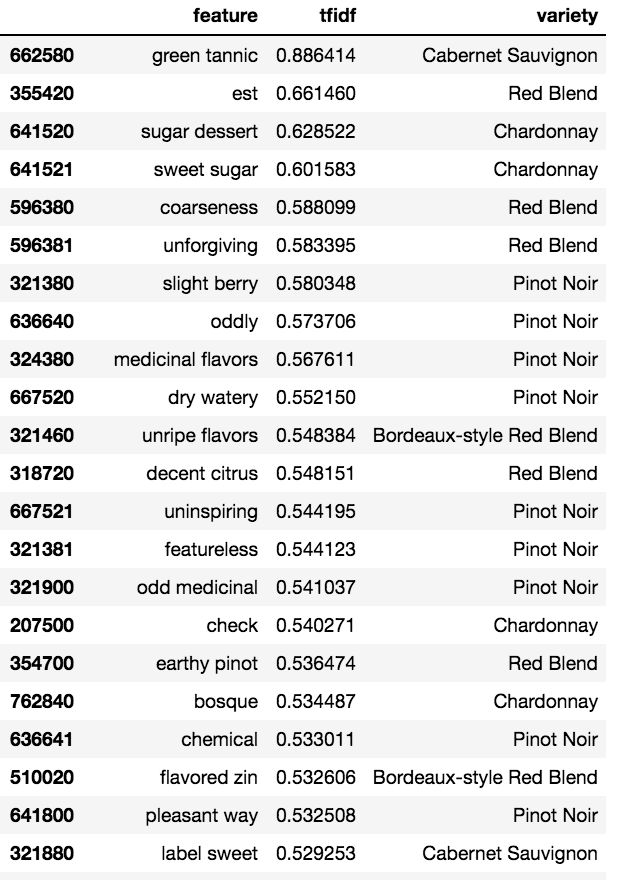
Missing value only appears in attribute ***Designation***(30.37%), ***Price***(9.07%), ***Provice***(0.78%), ***Region\_1***(16.60%), and ***Winery***(0.02%), after removing missing value and duplicate instances, I finally got 97156 instances along with 9 attributes, namely, ***Country, Description, Designation, Points, Price, Province, Region\_1, Variety, and Winery.*** I tried to use online translating APIs to translate French, Italian, and Spain, cause those language were not encoded as Ascii, and only convert them into UTF-8 would break the words. Unfortunately, popular translater APIs are not avaliable currently, this is the flaw of this project.

1. **Three Guiding Questions about the dataset domain:**
2. What are the best description of different varieties of wine?
3. What are the bext description of different country’s wine?
4. What are the

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Summary of Experiments.** *At most 2 page.* | | | | | | | |
| **Tool** | **Pre-process** | **Mining**  **Technique** | **Results**  **Variety: Description Features** | **Time**  **taken** | **Evaluation**  **TF-IDF score:** | **Observations about experiment**  **Observations about visualization**  **Interpretation results** | |
| Python | Stemmed and punctuation | Text mining | Chardonnay: green tannic  Cabernet Sauvignon: superoak, featureless, unforgiv, raspberryflavor  Red Blend: sugar dessert  Pinot Noir: medicine flavor | 115s | [0.886]  [0.785, 0.696, 0.660, 0.658]  [0.644]  [0.600] | Stemming and punctuation leaded to huge different results, that’s because forign landuage | |
| Python | No | Text mining | Cabernet Sauvignon: green tannic  Red Blend:est, coarseness, unforgiving  Chardonnay: sugar dessert,sweet sugar  Pinot Noir:slight berry,oddly,medicinal flavors | 127s | [0.886]  [0.661,0.588,0.583]  [0.629,0.602]  [0.580,0.574,0.568] |
| Python | Stemmed and punctuation | Text mining | Italy:green tannic,light rose,unforgiv  France:superoak,dilut tart  Spain:featureless,promis fruit,way toast  US:raspberryflavor,sugar dessert,cherryberri spice | 237s | [0.886,0.839,0.660]  [0.786,0.657]  [0.696,0.688,0.565]  [0.658,0.643,0.642] |  | |
| Python | No | Text mining | Italy:green tannic,light rose,promise fruit  US:wine company, est,sugar dessert,barrel  France:slight berry,medicinal flavor,viognier tropical  Spain:rosemount,syrupy cherry,juice tannins | 245s | [0.886,0.840,0.676]  [0.691,0.661,0.629,0.614]  [0.580,0.568,0.556 ]  [0.527,0.481,0.480] |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |







**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.

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