10/12/2019

KIT606 Assignment

Tweet

Lianxue Zhang 516785

Jiaxin Ye 513389

Cong Liu 518681

Table of Contents

[Problem Statement and Background 2](#_Toc26894968)

[The Structure and Characteristics of Tweet Dataset 2](#_Toc26894969)

[Descriptive Problem 2](#_Toc26894970)

[Predictive Problem 2](#_Toc26894971)

[Methods 4](#_Toc26894972)

[Data transformation 4](#_Toc26894973)

[Generate attributes with existing attributes 4](#_Toc26894974)

[Text mining 4](#_Toc26894975)

[Data cleaning 7](#_Toc26894976)

[Imputation for missing data 7](#_Toc26894977)

[Detect and remove outlier 10](#_Toc26894978)

[Inconsistent data 11](#_Toc26894979)

[Feature selection and design 11](#_Toc26894980)

[Descriptive feature selection 11](#_Toc26894981)

[Predictive feature selection 11](#_Toc26894982)

[Exploratory Data analysis 12](#_Toc26894983)

[Descriptive Data Mining 15](#_Toc26894984)

[DBSCAN Clustering 15](#_Toc26894985)

[K-Means Clustering 16](#_Toc26894986)

[Predictive Data Mining 17](#_Toc26894987)

[K-NN 18](#_Toc26894988)

[Decision Tree 18](#_Toc26894989)

[Naive Bayes 18](#_Toc26894990)

[Neural Network 18](#_Toc26894991)

[Evaluation 20](#_Toc26894992)

[Evaluation Methodology 20](#_Toc26894993)

[Parameter Selection 20](#_Toc26894994)

[Performance measure used 21](#_Toc26894995)

[Analysis of results 22](#_Toc26894996)

[Tools 23](#_Toc26894997)

[Lessons Learned 23](#_Toc26894998)

[Team Contribution 23](#_Toc26894999)

# Problem Statement and Background

## The Structure and Characteristics of Tweet Dataset

There are 8 attributes in the dataset below.

* Tweetid represents the primary key of Tweet.
* Coordinate type has only one value which is point.
* Geo.coordinate is represented by the format of latitude concating longitude.
* Place.country is represented by different language formats.
* Place.name is the column that being predicted.
* Userid is the column indicating who posted the Tweet.
* Tweet Language represents the language of Tweet, which can be a mix of languages.
* Tweet is content of Tweet.

The data types of all the attributes are polynominal except Tweetid which is real.

## Descriptive Problem

The descriptive problem of this assignment is to accurately identify location clusters by analysing longitude and latitude. Both DBSCAN and K-Means are used to solve this problem. Parameters, such as epsilon, min points, and k values, are manually selected to test clustering performance. The method with the best performance will be finally used to perform clustering.

## Predictive Problem

The predictive problem of this assignment is how to accurately predict the *place.name* where tweets are posted. The purpose of predicting place is to help twitter developer to understand the geographical distribution of tweets post frequency. This can eventually help it to select the most optimised server locations. For example, if the majority of tweets are posted in the same place, they can set up new servers in that place, thus resulting in faster information transformation and lower transformation costs. Also, by knowing the frequency of tweets posting in each place, it can help the company determine where to focus the advertising.

In order to solve this problem, it is necessary to explore the relations between *place.name* and other attributes (including *place.country, useid, language* and *geo coordinates*). Tweetid, coordinate type, place.country and Tweet columns are removed from the analysis since they are irrelevant and therefore do not help in the prediction of place.name. However, such columns can be used for data transformation.

Theoretically, although the coordinates can represent the place that tweets posted to some extent, there are many mistakes that could happen in real practice, which would result in bias in prediction if only using coordinates to predict the place. For example, the GPS of the phone may not update on time when user post the tweets due to the signal issues or broken hardware, which lead to wrong prediction by only using coordinates to build a predictive model. Therefore, it is necessary to train a model with applying other attributes besides coordinates like user id and language to obtain a better prediction accuracy.

# Methods

## Data transformation

### Generate attributes with existing attributes

For coordinates, it contains two parts - latitude and longitude. To get more accurate results, it is necessary to separate these two parts into two different attributes by applying functions when generating new attributes. After this process, we obtain two more attributes named “latitude” and “longitude” with setting them both as numerical data type.

For attribute “Country”, there are some countries expressed in different languages. To resolve this issue, we unified all different languages into English by using functions.

Following is the screenshot of the functions that are used for data transformation:

A screenshot of a cell phone

Description automatically generated

### Text mining

Tweet text in the data set contains text information but text mining is not suitable to combine with other attributes analysis. For example, text processing always produces different wordlists for different dataset, it is not possible to combine the wordlist while applying the model which was trained by other irrelevant attributes. However, text mining could be used for exploring the hot topics of a particular place, therefore it still contributes to the prediction of the place and is necessary to transform the complex text into word vectors with numerical values by using text mining methods.

A screenshot of a cell phone

Description automatically generated

To do text mining, only the text attribute should be analyzed, therefore, we selected only tweet and tweet id (as each tweet’s unique identifier) to feed the following text mining process.

Considering that all text mining functions only work on attribute with text data type and the original data type for tweet is polynominal, we transform the attributes with nominal data type into text.

Next step is to process the document into data with numerical values

A picture containing screenshot

Description automatically generated

Inside this subprocess, there are few separate methods to process the text: tokenize text into several words, transform all cases into lower case, filter stop words out, filter tokens by specified length, find the stem for each token.

1. Tokenize and transform cases
2. Filter stop word

Because the tweet text contains four different languages which are English, French, Portugunes, and spanish, it is necessary to filter stopwords with a dictionary that contains all four languages' stop words. Therefore, we combined four language stopword dictionaries into one document and write this document into filter stopword operator.

1. Filter tokens by length

Only the tokens with lengths between 4 chars and 15 chars would be filtered out to get rid of too short or too long tokens.

A screenshot of a cell phone

Description automatically generated

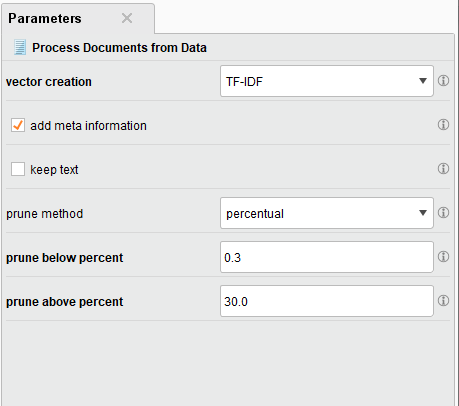
1. Find the stem word

Similarly, we sequentially stem the words using Snowball algorithm since it support all four different languages.After text processing, we obtained many word vectors as new attributes.

A screenshot of a cell phone

Description automatically generated

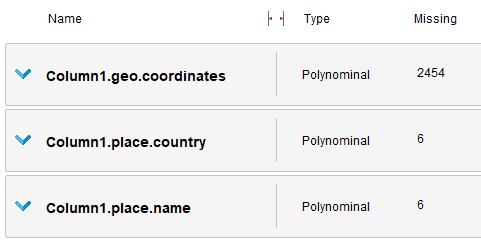
However, due to the reason that the data set is very large, there would be a huge amount of word vectors to be generated as attributes, which could potentially increase the cost of running the process, so we use prune function by setting parameters as follows.



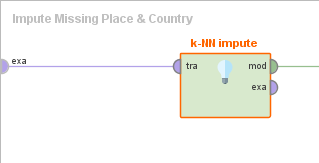
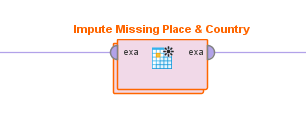
## Data cleaning

### Imputation for missing data

According to the structure of “Tweet” that shows in RapidMiner, there are many missing values in attributes “country”, “place”, and “coordinates” (which will be replaced by “latitude” and “longitude”).



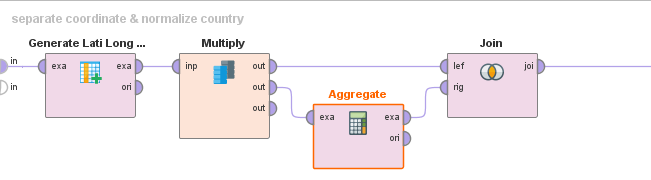
Considering that the data type of “country” and “place” are polynominal, k-NN method is selected to impute the missing values of these two attributes to find the nearest neighbours of the examples that contains missing values and impute the missing tuples with corresponding values of “neighbours”.

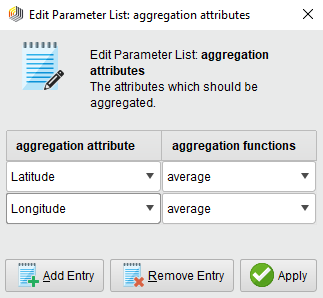
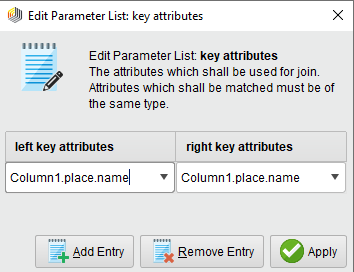


Meanwhile, it is simple to implement and use k-NN, thus saving the cost of the process. In addition, it is also used to explain prediction.

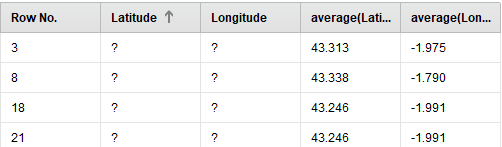
To impute the value of “latitude” and “longitude”, the close relation between “place” and these two attributes are noticed – one place indicates a certain range of “latitude” and “longitude”. Based on this inference, our team decided to retrieve the average value of “latitude” and “longitude” group by “place” and replace the missing value with the average value in that place.

To achieve this goal, the first step is to aggregate the average value of “latitude” and “longitude” group by “place”, then left join the new table with average values with the original data set by the key “place”.



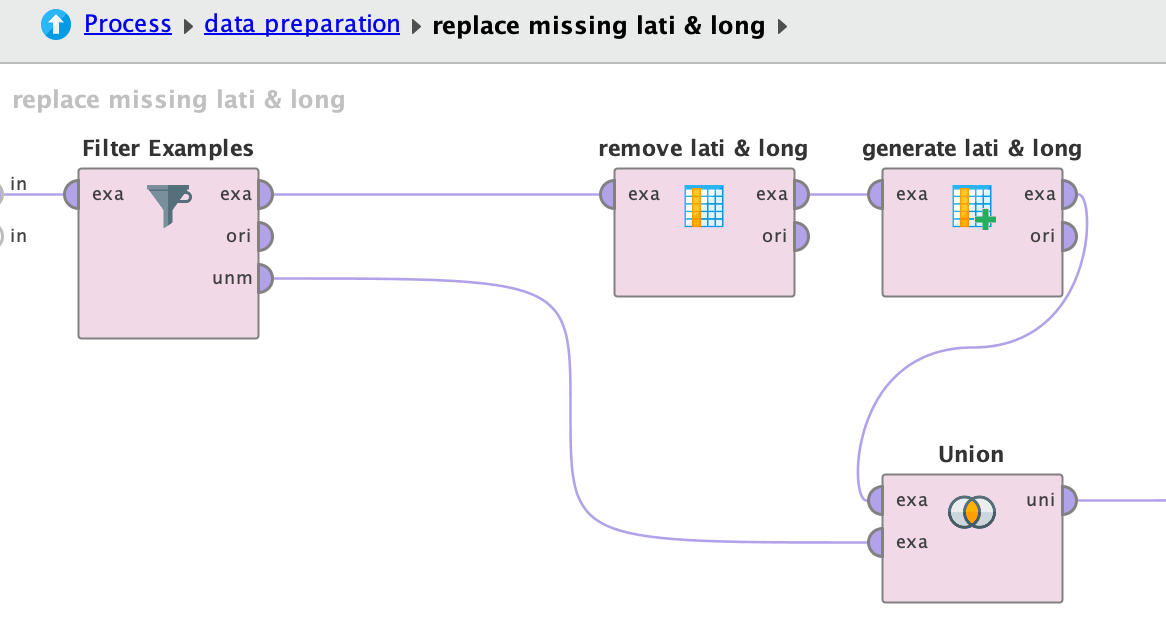
 

The first few results are as follows:



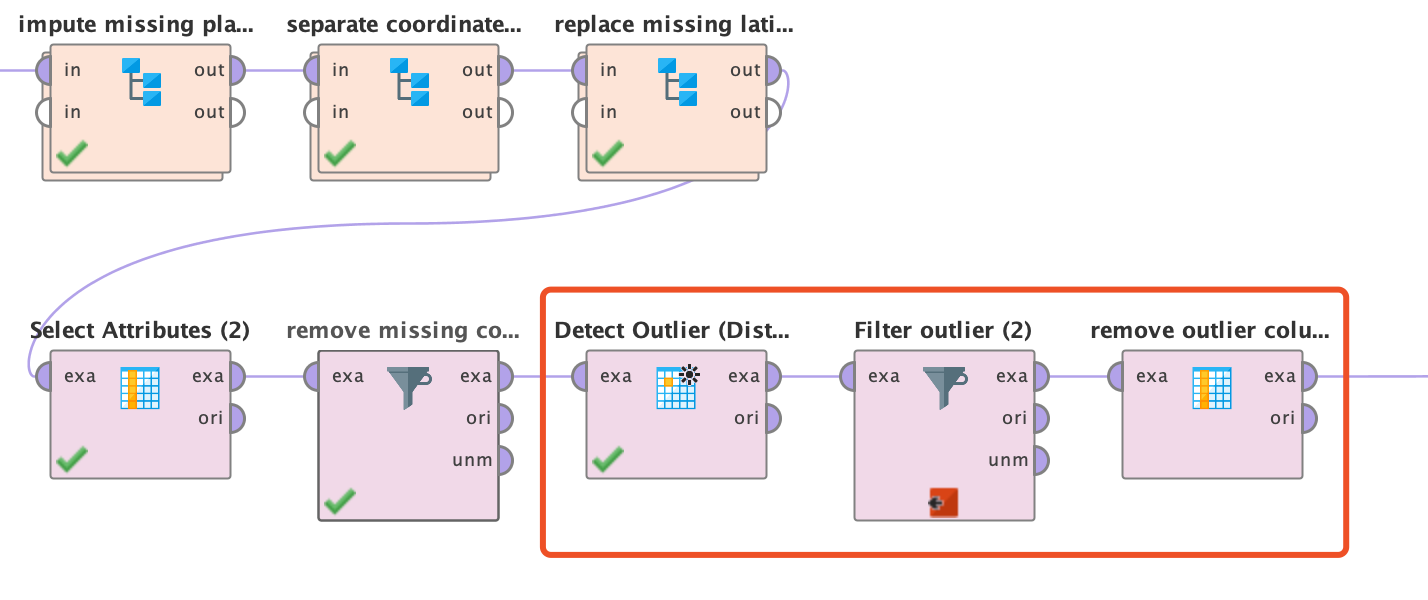
Now the average latitude and longitude are obtained for the missing values as two new attributes.

Next, the missing values of latitude and longitude are replaced by average values by following way: firstly filter rows with missing latitude and longitude (because we only want to replace the missing values), secondly remove latitude and longitude columns, then generate two columns with the same name and finally union the new table with previously unmatched table.

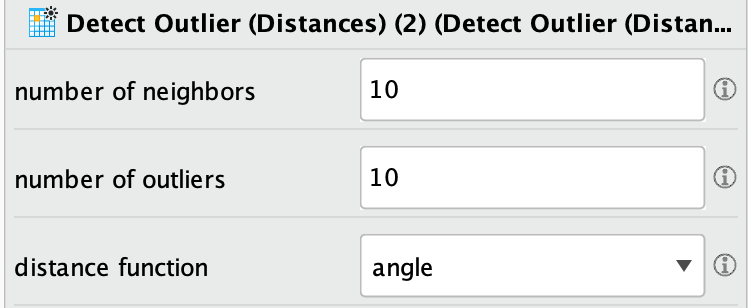


However, it is impossible to find the average value of “latitude” and “longitude” in one place when all the values of “latitude” and “longitude” in that place are missing. *Therefore, we remove those values that are unable to be imputed by this way because they only take a very small proportion in the whole data set hence do not influence the result.*

### Detect and remove outlier



The graphic above indicates the process of detecting and removing outliers.



Our team chose to detect outliers by the distances between data points rather than by density manner, because the density distribution is irregular. As for parameters, we set the number of the outlier as 10. Other distance functions also have been tried but angle can achieve best performance that can be evidenced in visualization. Following is the scatter plot in 3D after detecting outlier with using angle distance function:

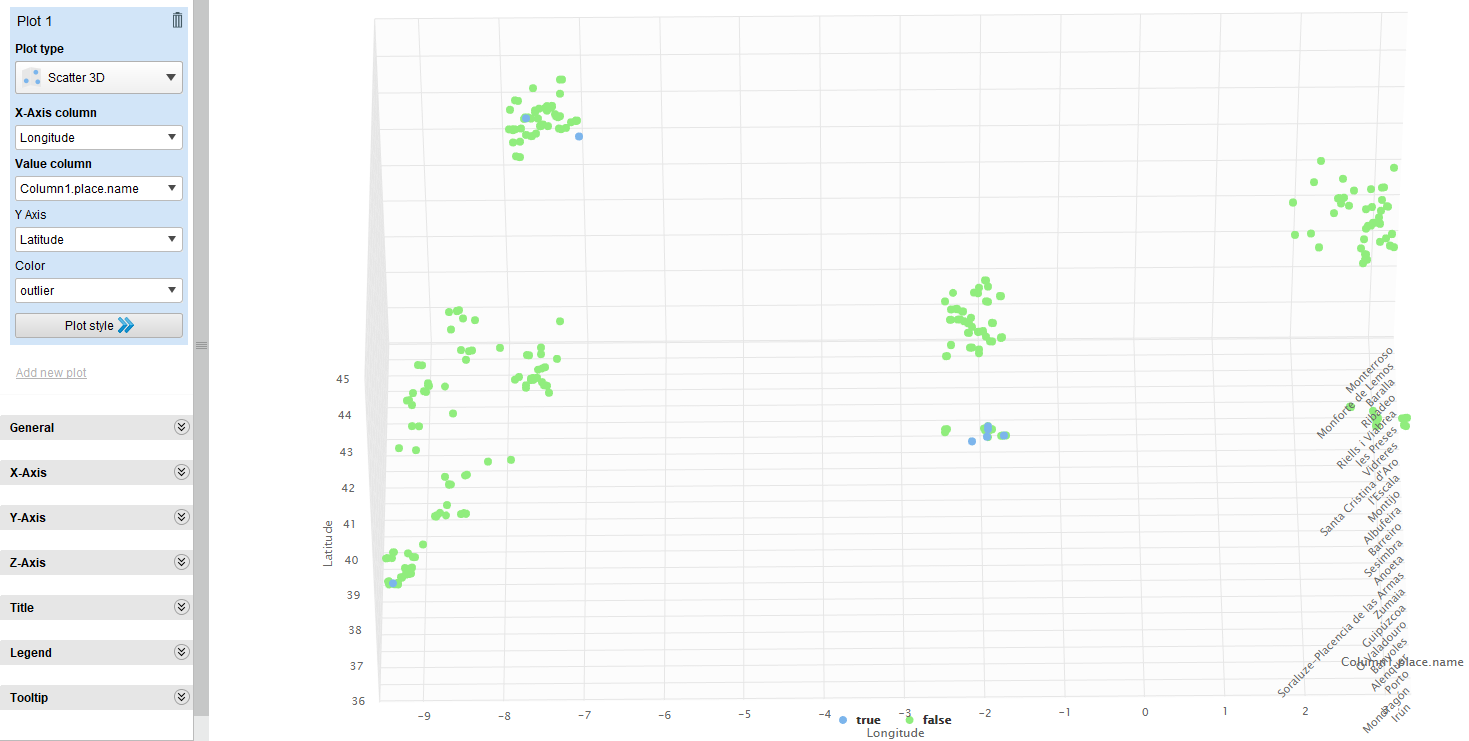


Figure 1. outlier distribution

*All the blue points which stay far away from other points are detected as outliers.*

Next, the outliers will be removed by filtering examples.

### Inconsistent data

Values of *place.country* are considered as inconsistent data because the same country is represented by different language forms. For example, Spain is represented by Spain, Espainia, España and other languages.

To deal with the inconsistent data, we generated a new attribute with values in English corresponding to the original values using the if else condition.

## Feature selection and design

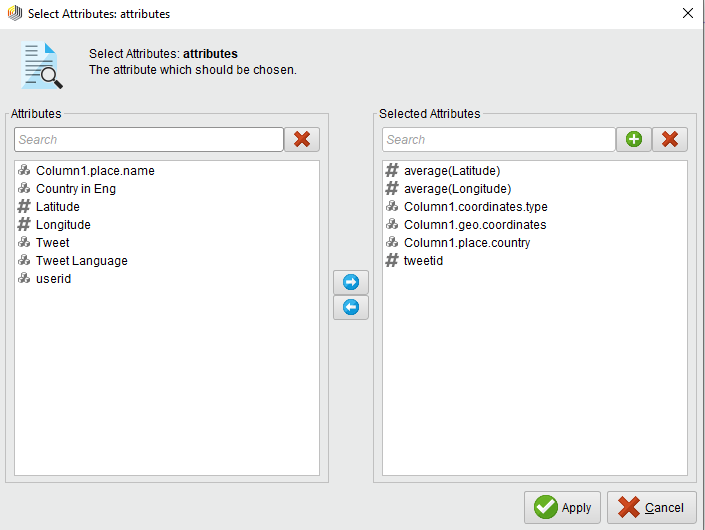
### Descriptive feature selection

During the data preparation, our team adopt text mining methods to generate new attributes (word vectors) in order to analyse the key word frequency in different places, which help to explore the data set better.

Our team also generates a new attribute named “Country in Eng” to unify the language of Country and make it clearer for exploratory data analysis.

### Predictive feature selection

Original data set contains some useless attributes to make predictions. In addition, many new attributes, such as average latitude & longitude, are generated during the process of data transformation and imputation, which makes some original attributes become unnecessary. Therefore, dimensional selection(reduction) is a crucial step which must be done before modelling.





As we can see from the above screenshot, average latitude and longitude become unnecessary after using them to replace the missing values of latitude and longitude. As for geo.coordinates, because “latitude” and “longitude” already covers all information of this attribute, it became useless and should be removed before modelling. Similarly, country also contains repetitive information of “Country in Eng”, which also should be removed.

The reason to remove tweetid and coordinates.type is that they are irrelevant to predict “place” because there is no evidence showing that these two attributes have any relation with “place”.

## Exploratory Data analysis

A screenshot of a cell phone

Description automatically generated

Figure 2. Frequency of Tweets in different Places

This above histogram uses *tweetid*, *place.name* and *Tweet Language* features which is preliminarily created without data mining. The Y-axis stands for the number of Tweets posted in a specific place, while the X-axis represents place names. The color stands for Tweet Language. It can be clearly seen that there are three main languages used to post Tweets, which are es, ca and pt respectively. The majority of tweets for each place are posted in the same language. For example, over 95% tweets in San Sebstian are posted in Spainish. In addition, this graphic also shows the range of places where using the same language.

As for the number of tweets posted, it can be found that the highest number is in Lugo, while the second highest number is in San Sebstian. This indicates that these two places have large number of active users and high post frequency.

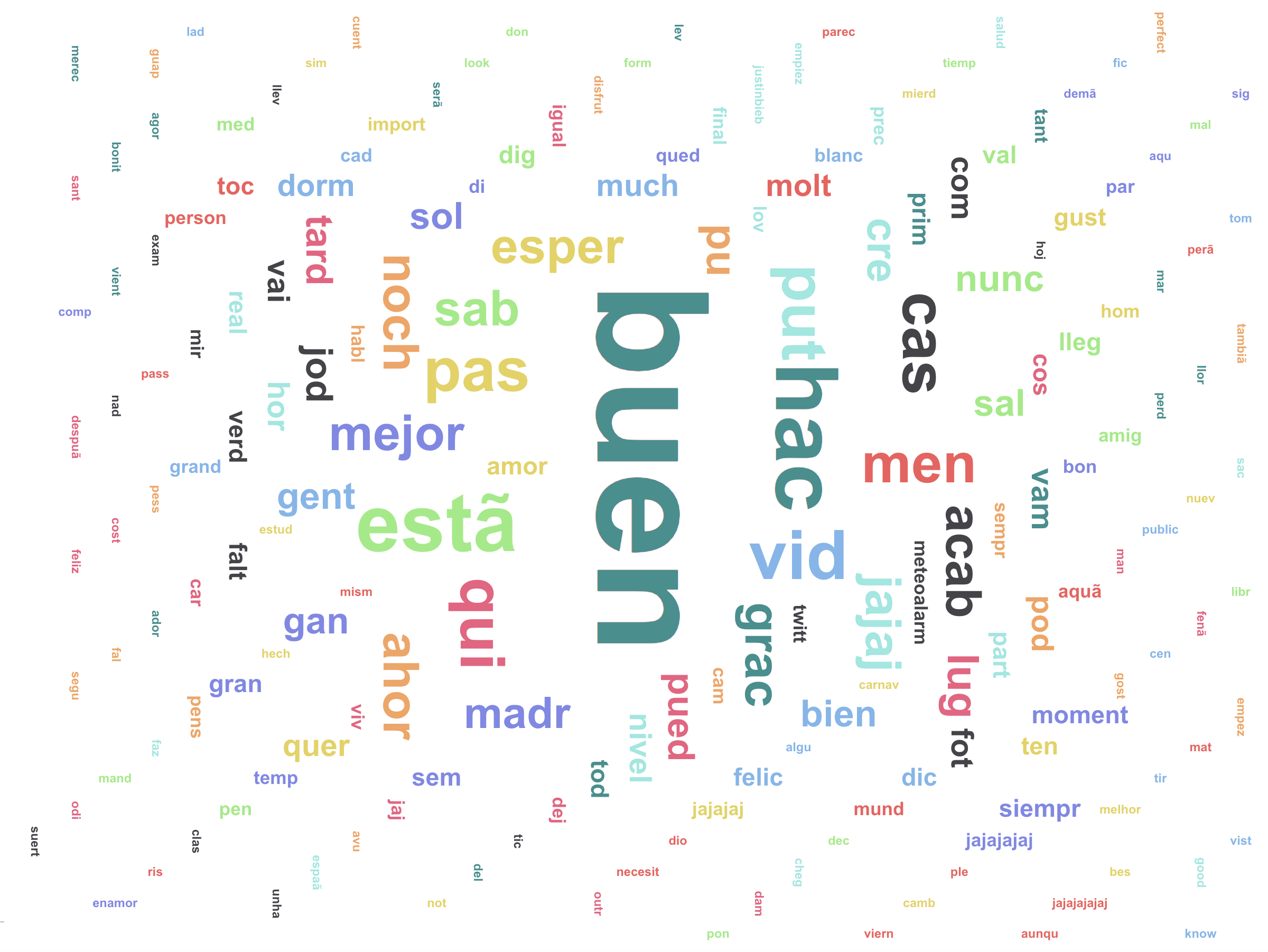


Figure 3. Word frequency in Tweet content

All the words showing in this wordcloud picture are the results after processing including tokenization, filtering stopwords, and extracting stem. Only the words with high frequency in tweet content are displayed above. Bigger the word is, higher frequency it represents.

According to this picture, people can build a clear overview about what are the topics that twitter users care most by simply comparing the size of the word in picture. Additionally, this picture also implies the relations between hot topics and language. For example, the word “tard” means “late” in French. Therefore, we could infer that French speakers like discussing about time delay issues. Furthermore, the sentiment context can also be discovered according to some words. For example, the most frequent word “buen” means “good” in Spanish. Based on this information, the inference that most twitter users prefer to post events on positive side could be made.

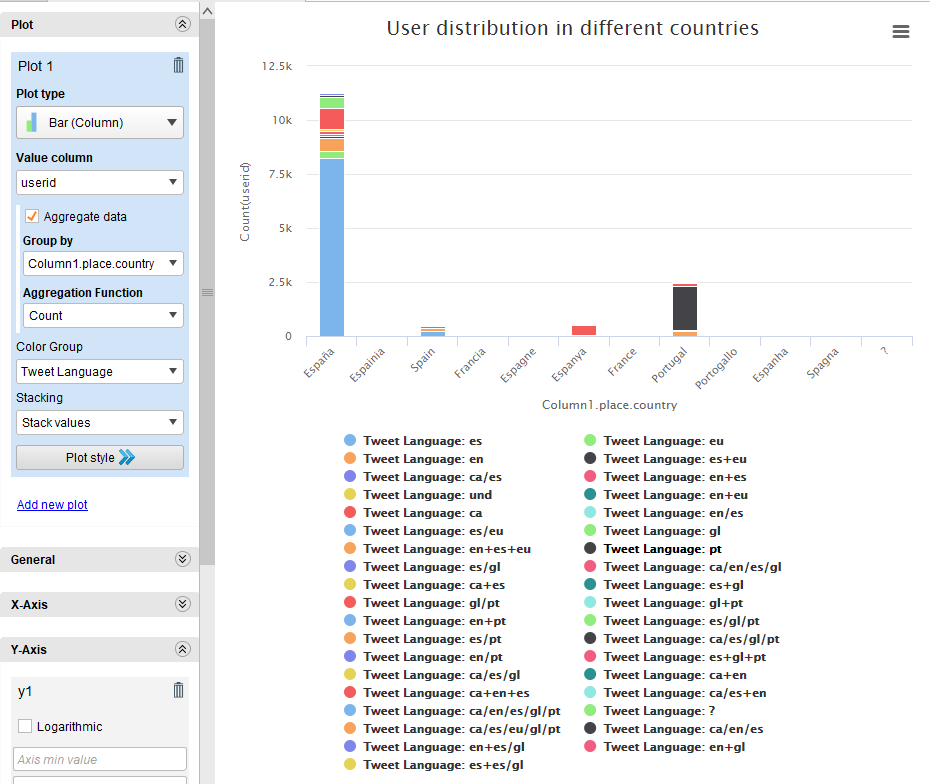


Figure 4. User distribution in different countries

The above bar chart uses *userid*, *place.country* and *Tweet Language* features, and the relation between the number of users and country is revealed. It can be clearly found that Espana has the highest number of users while Portugal has the second highest number of users. Colours represent different languages in the plot. Espana has a wide variety of languages while the language for Portugal is relatively single.

## Descriptive Data Mining

### DBSCAN Clustering

As we can see from the map below, if esplion=0.5 and minPts=10, DBSCAN generates 6 clusters based on other attributes. However, cluster\_0 (black) shows a low data density. This indicates that there are lots of outliers in cluster\_0 (black). Similarly, this problem also incurs in cluster \_3 (earth-yellow). In reality, cluster\_0, cluster\_2 and cluster\_3 are from the same country. In this case, they are labelled as different clusters due to the drawbacks of DBSCAN clustering, e.g. density sensitive.

This is the best result that my team found by manually selecting epsilons and min points though there are still some errors in both cluster\_0 (black) and cluster \_3 (earth-yellow). We have tried esplion=1 and minPts=10, esplion=0.5 and minPts=10, esplion=0.3 and minPts=5 and so on, however, none of them produces a better clustering from the location distribution.

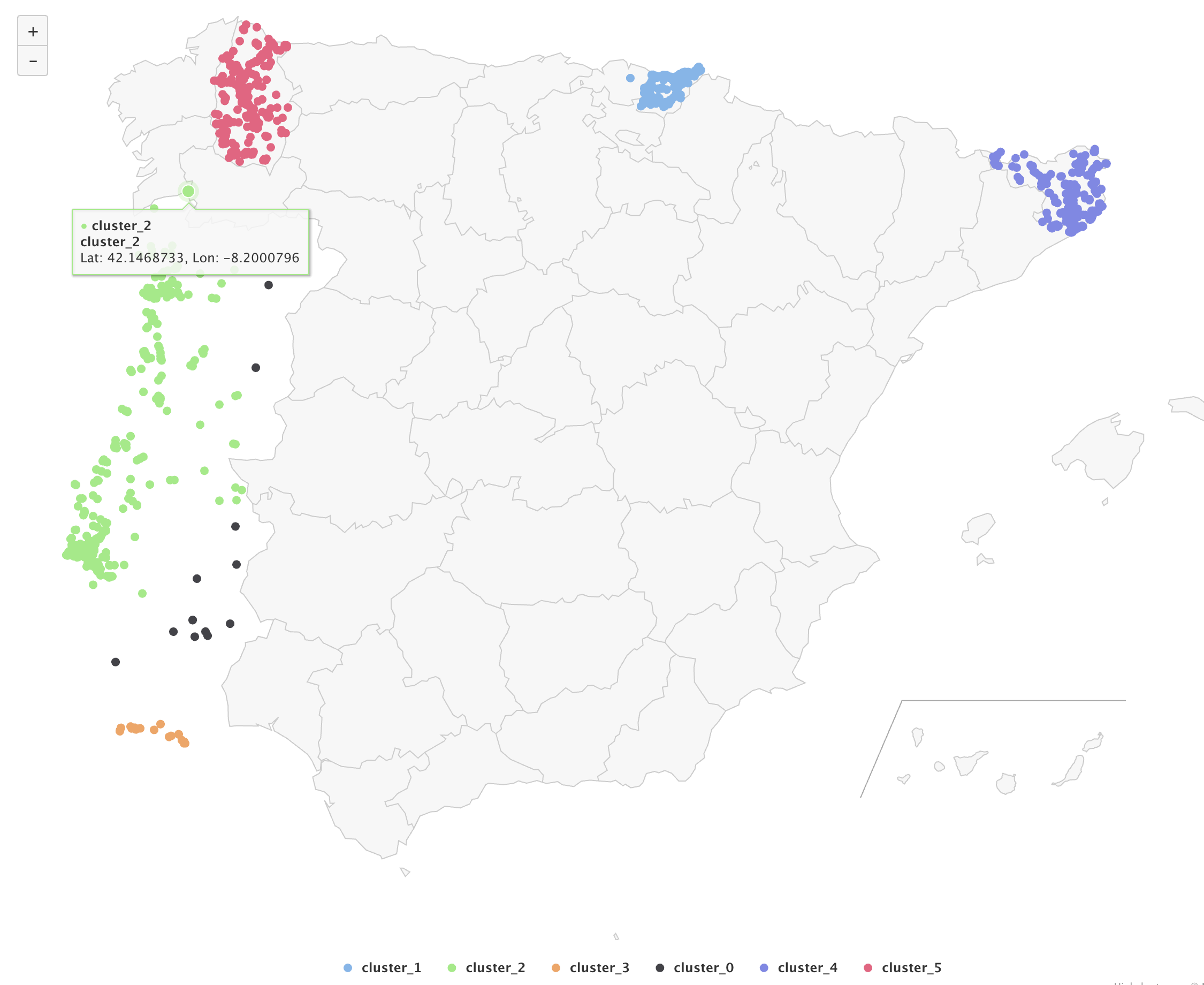


Figure 5. DBSCAN with epsilon=0.5 and minPts=10

### K-Means Clustering

However, if using K-Means, it can be seen that cluster\_0 (orange) and cluster\_3 (green) do not present reality accurately. Specifically, some data points which should belong to cluster\_3 (green) are clustered into cluster\_0 (orange). Additionally, the distances between data points in cluster\_3 (green) are very long, especially for those on the right-hand side. This may imply that cluster\_3 (green) contains many outliers which results in a wrong centroid. It may also indicate that k=4 is not appropriate in this case.

We also tried other k values like 3, 5, 6, 7, but none of them produces a better clustering compares with k=4.

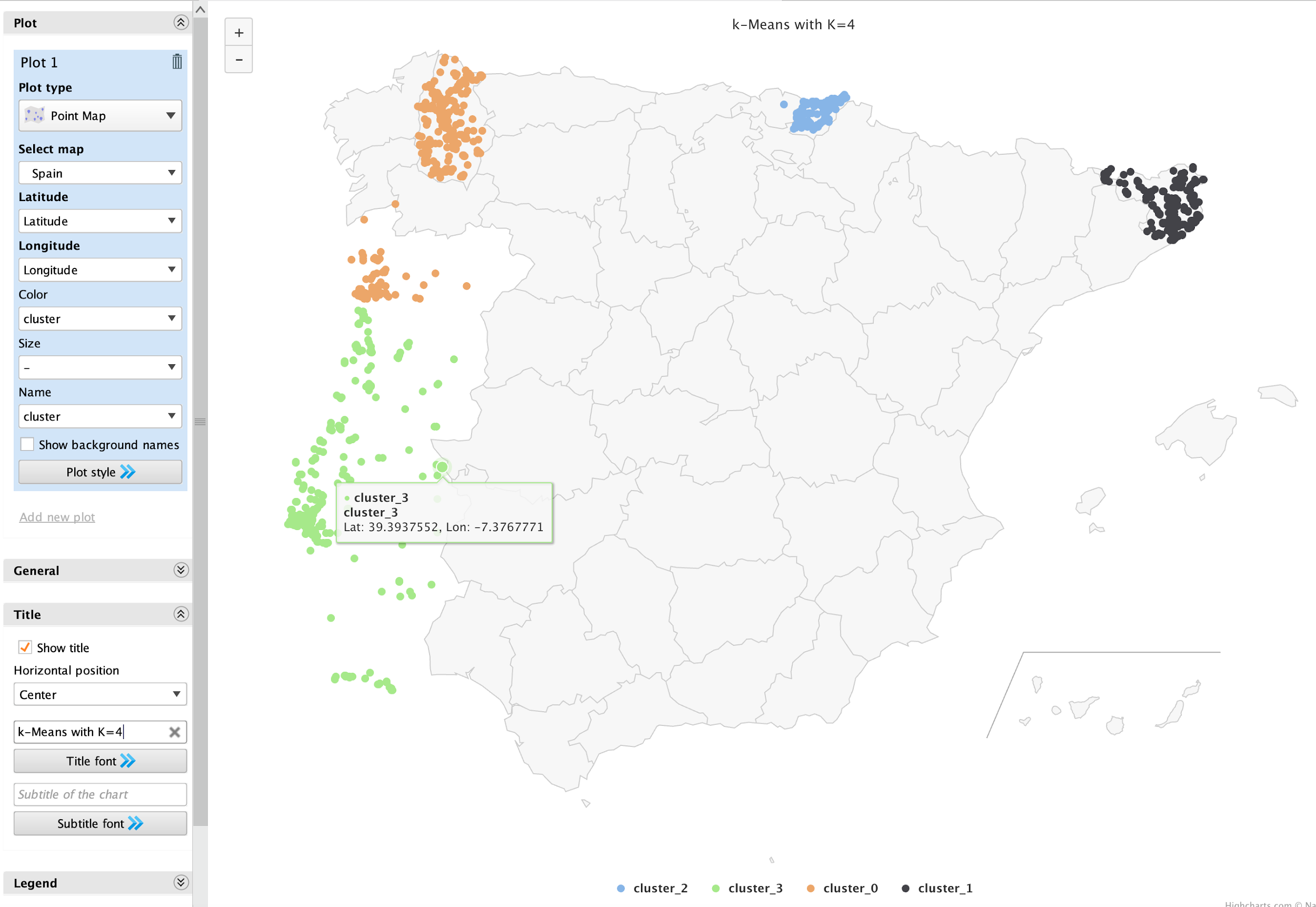


Figure 6. K-Means with k=4

By comparing the two maps above, it can be found that the clustering result of DBSCAN is closer to the ground truth than that of K-Means. In terms of functions, DBSCAN has four main advantages. Firstly, DBSCAN determines the number of clusters automatically. However, if using K-means, initial centroids or k value will be difficult to be determined. Secondly, DBSCAN can handle large dataset efficiently, while the complexity and time of data processing are much higher if K-Means is chosen, because the distances between centroids and all other data points need to be calculated by iteration. Thirdly, DBSCAN eliminates interference of noisy data, while K-Means is very sensitive to noisy data and outliers. More importantly, DBSCAN can handle clusters of different shapes and sizes, while K-Means does not work well for non-convex or differing size structures.

Based on the analysis above, it can be concluded that DBSCAN is the better manner to cluster geographical coordinates.

## 

## Predictive Data Mining

According to the advantages and disadvantages of each algorithm, k-nn, decision tree, naïve bayes and neural network are elected as training algorithms. Because SVM is only suitable for binary classification and linear regression is only used for numerical value while the data type for most columns is ploynominal, so they both are not suitable for training models with selected features.

A screenshot of a social media post

Description automatically generated

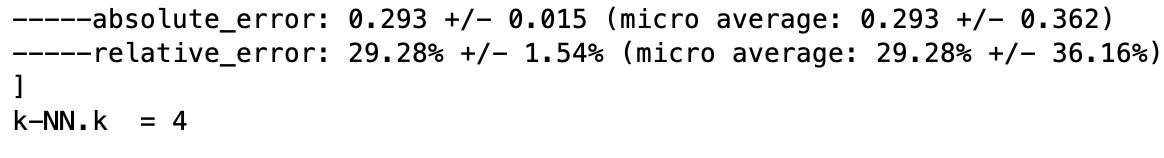
### K-NN

The k-NN algorithm is one of the most frequently used learning algorithms. It makes predictions by

comparing the distance from an unlearned example to existing examples and finds the ones that are the closest.

The goal is to predict *place.name* from varies attributes, k-nn is especially performing well in a multi-dimension distribution. Also, it is simple to implement and use while easy to explain the prediction.

In order to get a k value with best performance, it is embedded in the optimize parameters operator. The loop starts from 1 and ends with 10. It outputs a good result which is 78.38% +/- 2.08%.



### Decision Tree

Decision tree is selected because it can be seen graphically in the Graph View and is easy to explain to non-technique stakeholders. With a relatively higher depth, it will produce a good accuracy. Also, decision tree does not require normalization of data and the missing values does not affect the building of decision tree.

The main parameters for decision tree are maximan\_depth, criterion, confidence. Different parameters can cause very big difference in performance. Therefore, combined with optimizing parameter methods, it can give a good performance with optimized parameters.

### Naive Bayes

Compared with other algorithms, Naïve Bayes is less expensive and gives relatively good results. Besides, it is insensitive towards irrelative attributes and is very fast to train model, which is suitable for small workplace. Therefore, it is suitable for large data set processing thanks to its’ high efficiency in cost and speed.

### Neural Network

Neural network algorithm has a lot of advantages. It can used in many data science problems, e.g. clustering, classification. And it does not need to be re-programmed. Thus, it saves time if a model is already trained well.

However, neural network requires a rather powerful hardware. In other words, it is obviously resource consuming. During the training phase, it takes more than 7 hours to train a model. Though the parameter is adjusted to some extent, it still consumes a lot of time and CPU compare with other algorithms. Therefore, it is not suitable for quick implementation.

A screenshot of a cell phone

Description automatically generated

Based on above reasons, Neural Network is abandoned to train predictive model.

# Evaluation

## Evaluation Methodology

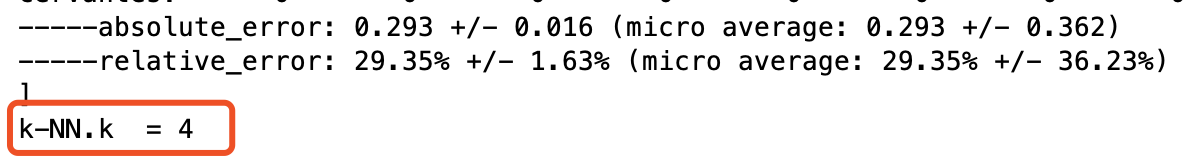
The k-fold cross validation is selected as the methodology with k=10. Compared with hold-out validation, this methodology has a relatively higher accuracy because it takes an average accuracy of k tests as the result while keeping a similar efficiency.

In order to achieve a more precise accuracy, the evaluation methodology is embedded into the Optimize Parameters operator. Four predictive algorithms are being tested which are k-nn, decision tree, naive bayes and neural network. It is found that neural network is the most costly and least accurate method, taking hours of training time but only achieves a 10% to 20% accuracy. After comparing the other three methods, we decided to take decision tree as the final training method. Performance for each of them are list as below (from one of the test results):

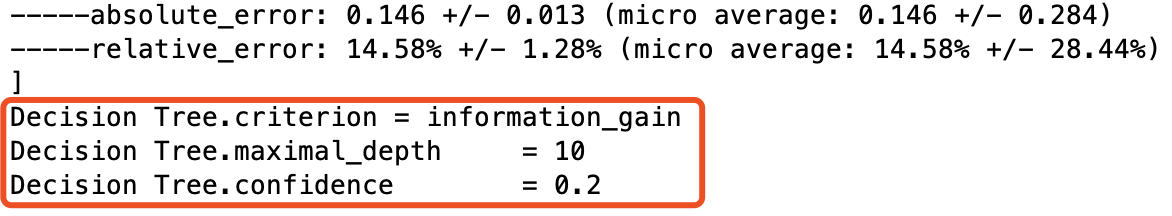
|  |  |  |  |
| --- | --- | --- | --- |
|  | K-NN | Decision Tree | Naive Bayes |
| Accuracy | 78.24% +/- 2.69% | 88.82% +/- 1.88% | 79.44% +/- 1.24% |
| Absolute Error | 0.293 +/- 0.016 | 0.146 +/- 0.013 | 0.200 +/- 0.011 |
| Relative Error | 29.35% +/- 1.63% | 14.58% +/- 1.28% | 20.03% +/- 1.07% |
| Weighted Mean Recall | 29.08% +/- 1.69% | 35.21% +/- 1.57% | 36.24% +/- 1.31% |
| Weighted Mean Precise | 28.54% +/- 1.81% | 33.74% +/- 1.69% | 35.07% +/- 1.23% |

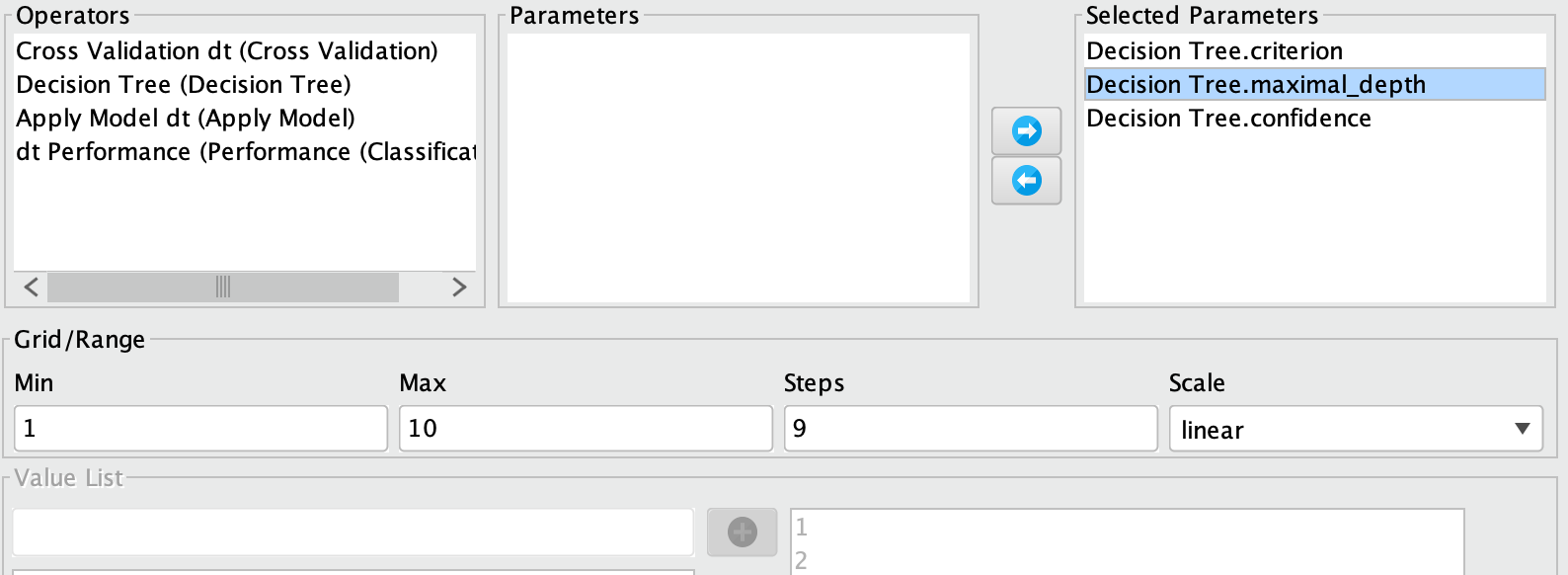
## Parameter Selection

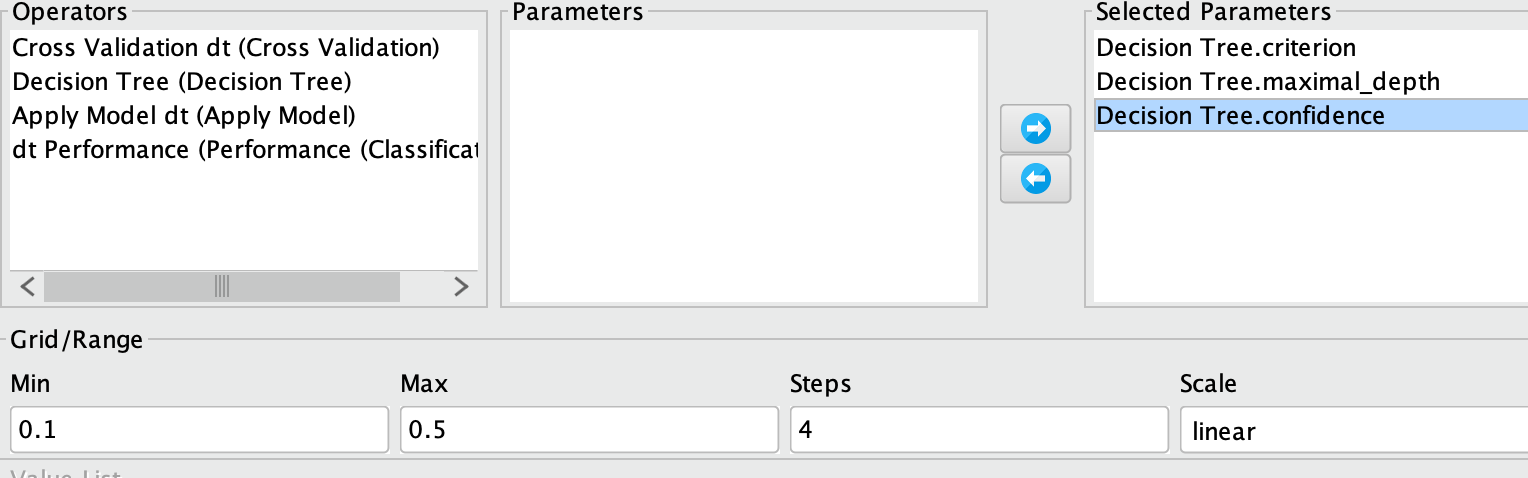
K-NN has 10 validation iterations with k value from 1 to 10. The above performance is achieved when K=4, which produces the highest accuracy among other values.



Decision Tree has 200 iterations. The default value of maximal\_depth is decreased from 100 to 10 because the higher depth produces a similar accuracy but consumes much more time and CPU resources to process the data. The default min of confidence is changed to 0.1 and the steps is set to 4. The above performance is achieved when maximal\_depth=10 and confidence=0.2.







Naive Bayes contains only one parameter which cannot be changed therefore it is left as default.

## Performance measure used

Accuracy is used as the measure to evaluate how good the model is. The reason is that classes of the predicted column in this case is moderately balanced when looking into the visualisation from the descriptive analysis, if cluster\_0 and cluster\_3 are treated as outliers.

A screenshot of a computer

Description automatically generated

Figure 7. moderately balanced group of *place.name*

## Analysis of results

The value of recall, precise and accuracy can be calculated from the confusion matrix with particular formulas. However, it is more accurate to take accuracy as the performance measure according to the balanced clustering of places.

Naive Bayes performs well under the assumption that all attributes are independent. In this case, there are several attributes and they dependent on each other, e.g. the column *Tweet Language* depends on *place.country* in reality, therefore it is not suitable though it has a high accuracy.

Since decision tree produces the highest accuracy while accuracy is considered as the performance measure, it is finally decided to use decision tree with parameter set maximal\_depth=10 and confidence=0.2 for predicting place.name of where tweet is posted.

# Tools

RapidMiner is used to process Tweets data. It offers a range of operators which can be used for data cleaning, data mining, text mining as well as data modelling. It is very easy to use and has a capability to visualize data by various plots. Moreover, it can automatically identify processing errors and gives proper solutions. The only drawback is that the processing efficiency may reduce significantly if there are large amounts of data or setting parameter values inappropriately.

# Lessons Learned

Through the whole process of training models and applying models, every team member has work together in each stage with taking responsibilities in different aspects. Therefore, we learned how to work in a team efficiently.

It is also important to take into careful account on predictive methods (e.g. K-NN, A-NN, K-Means or DBSCAN). Failure to do that may result in low data processing efficiency and accuracy. For example, we had initially used A-NN to build a mode, but this method took several hours to run the data due to fairly large amount of dataset.

Additionally, setting appropriate parameters is also challenging for us. We learn that we need to compare the performance of parameters before determining their values. By testing parameters, prediction accuracy improves significantly. For example, in DBSCAN clustering, we adjusted epsilons and min points many times in order to get more accurate clusters.

# Team Contribution

513389 Jiaxin Ye **33%**

At the developing stage, she took part in problem and process design. During the process of developing in RapidMiner, she works with other team members to work out each part and resolve the encountered problems. For both developing and report writing, she is responsible for data cleaning, data transformation, and feature selection including imputation for missing data, dealing with outliers, text mining, and dimensional creation and reduction.\

518681 Cong Liu **33%**

He worked together with teammates by using RapidMiner to perform data cleaning and data mining such as imputing missing values, removing outliers and setting parameters (e.g. k values, min points, or eslipons). He is responsible for descriptive statistical analysis by bar charts. He also did data clustering by using DBSCAN and K-Means. Additionally, he did preliminary data exploratory in EDA part.

516785 Lianxue Zhang **34%**

He is responsible to process data by RapidMiner especially optimizing the whole process performance. He also did evaluation part to validate the performance. As for report, he helps other team mates to find out the mistakes and modified them to improve the quality of the report. He is also responsible to ensure all operators functioning logically as expected.