**Data source:**

Tweets

**Context**

Dataset contains the information of tweets. For each tweet, there are tweet id as unique identifier, coordinate types, geo coordinate details, country, place, user id, tweet language, and tweet text. The data types for all attributes are polynominal except tweet id which is real.

**Problem**

the main problem we plan to address is to predict the value of “place” by exploring the relationships between place and other attributes.

To achieve this goal, the distribution of the “tweets” data and “language” data in different places also needs to be addressed. For tweets text, key words that present most frequently in each place would also contribute to the final model.

In the process, we will use descriptive method to train the model by clustering while using predictive method to predict the value of “place”.

**Analysis tool**

RapidMiner is used to analyze each attribute and to generate the features to cluster datasets in order to train the model to predict the target attribute.

**Data product**

The deliverable will be a well-trained model which can be used to predict value of “place” in high accuracy.

**Preparation**

In the preparation phase, the attributes which have no contributions to model training will be removed, e.g. tweet id, coordinate type. K-NN method will be used to impute missing values. Later on, outliers will be either removed or smoothed.

A screenshot of a cell phone

Description automatically generated

**Preliminary statistics/exploration**

The following bar plot illustrates the relationship between the frequency of Tweets and place. Colored by languages.

A screenshot of a cell phone

Description automatically generated

From the plot, it can be concluded that San Sebastian and Lugo have the highest frequency of tweet posting. It is also obvious that places are using the same language in majority.