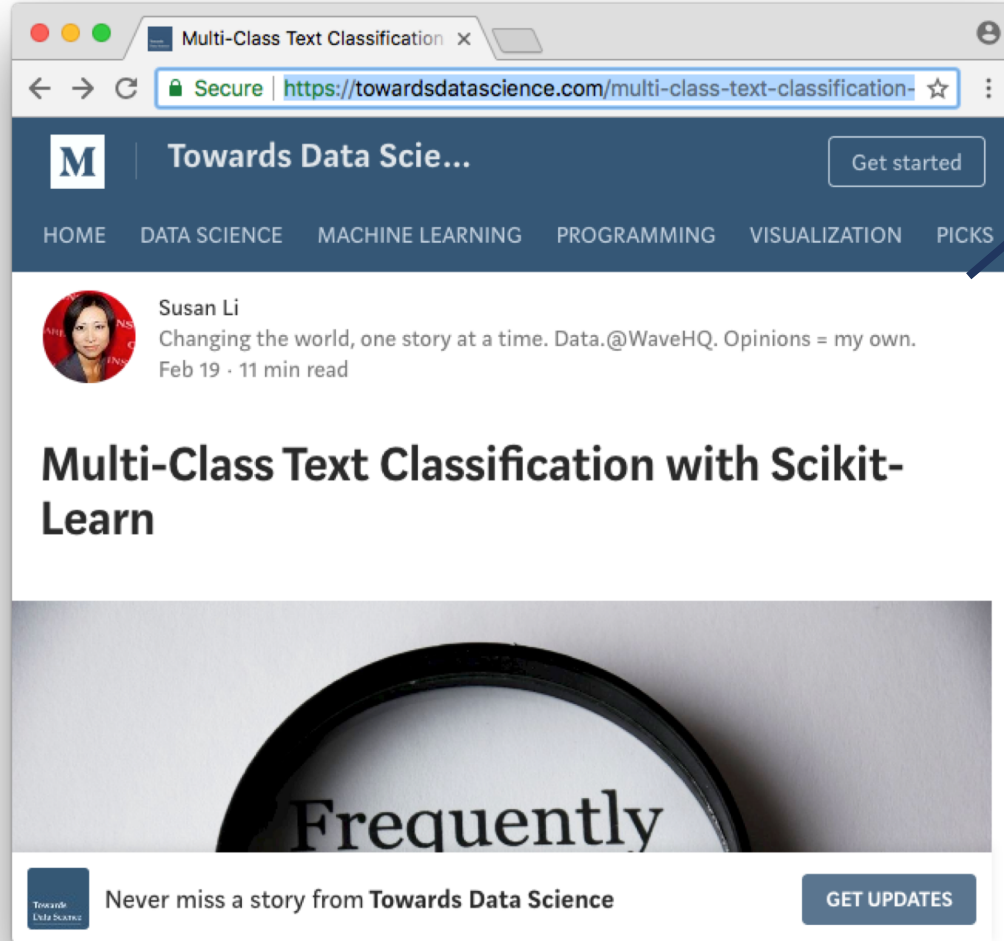


Scrapy Project

Multi-Class Text Classification with Scikit-Learn

Felipe da Silva Santos

Introduction

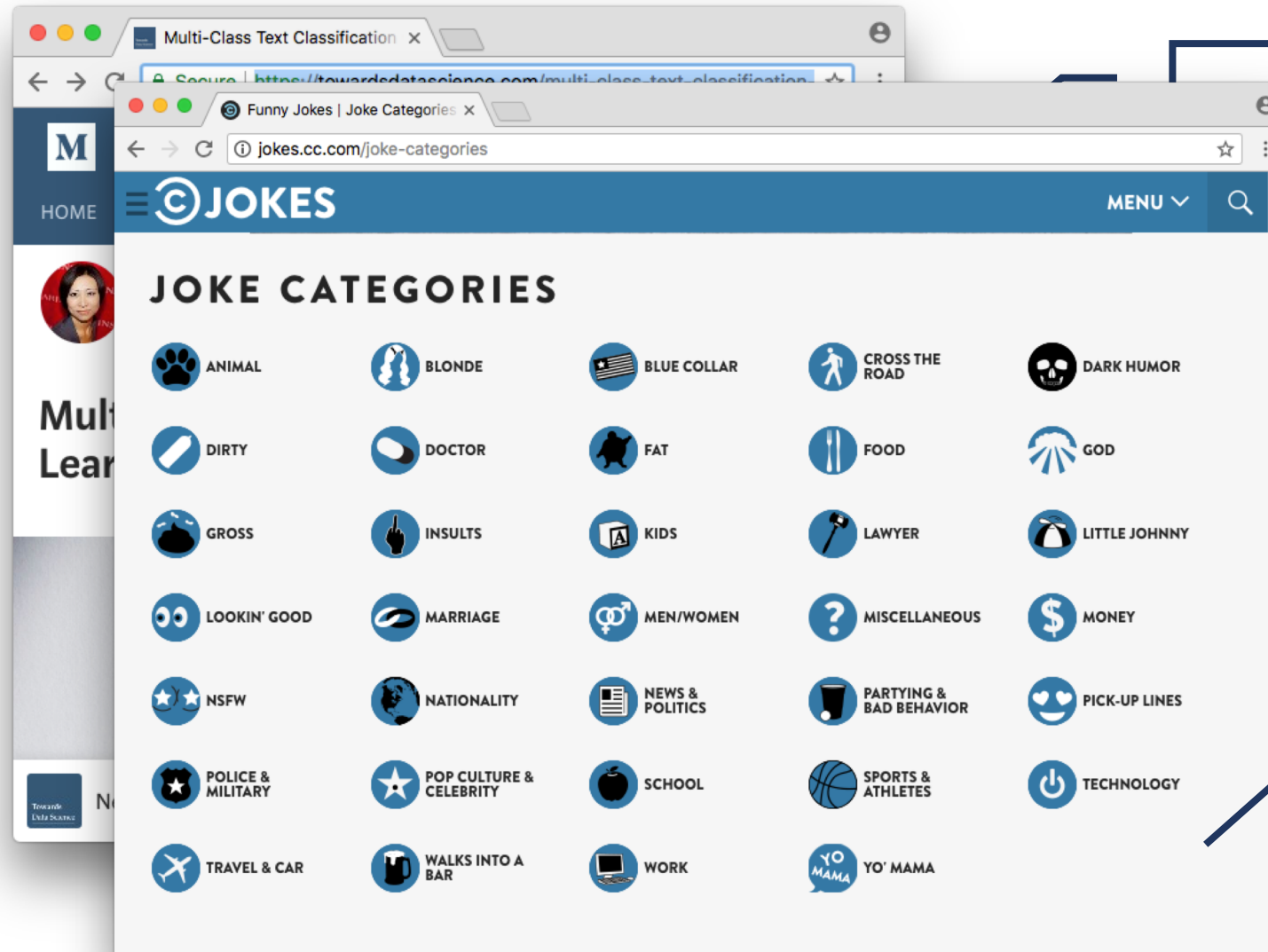


Susan's article

Classifying customer finance complaints into 12 pre-defined classes.

Does it work in different situations?

Introduction



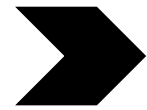
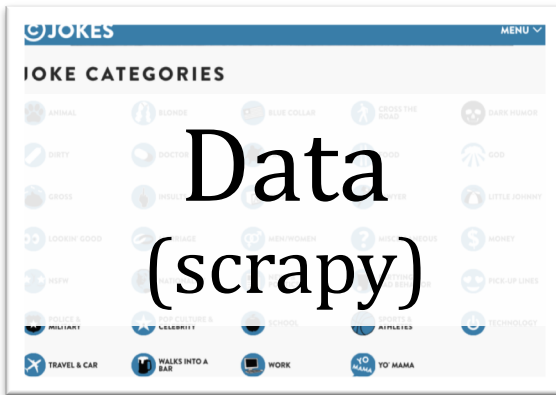
s article

ing customer finance complaints
pre-defined classes.

it work in different situations?

Scrapy
Source of data.

Steps

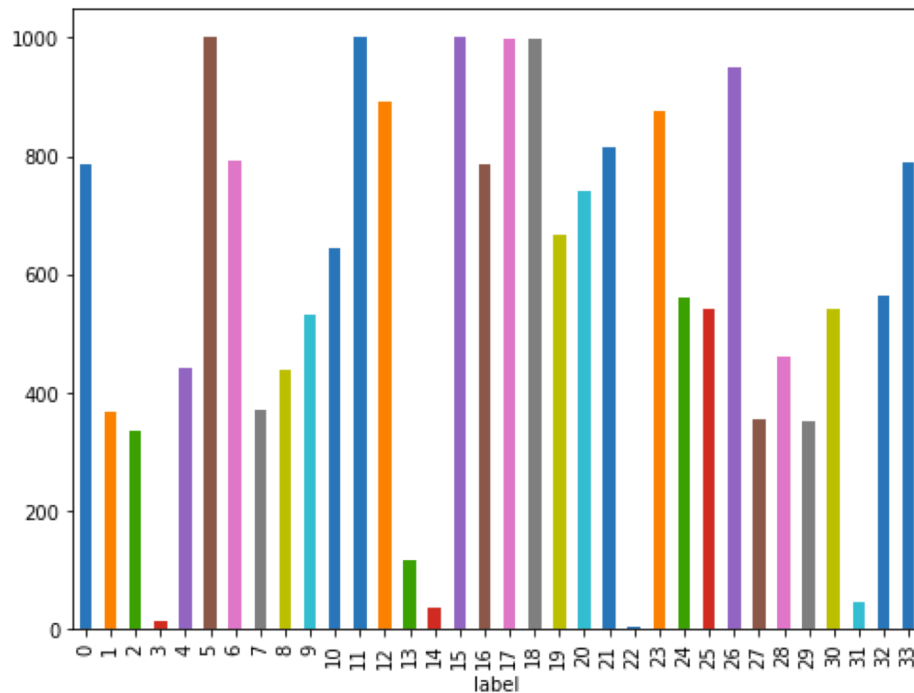


- Tf-idf term weighting
- Naive Bayes Classifier
- Random Forest
- Logistic Regression



Accuracy

34 Imbalanced Classes

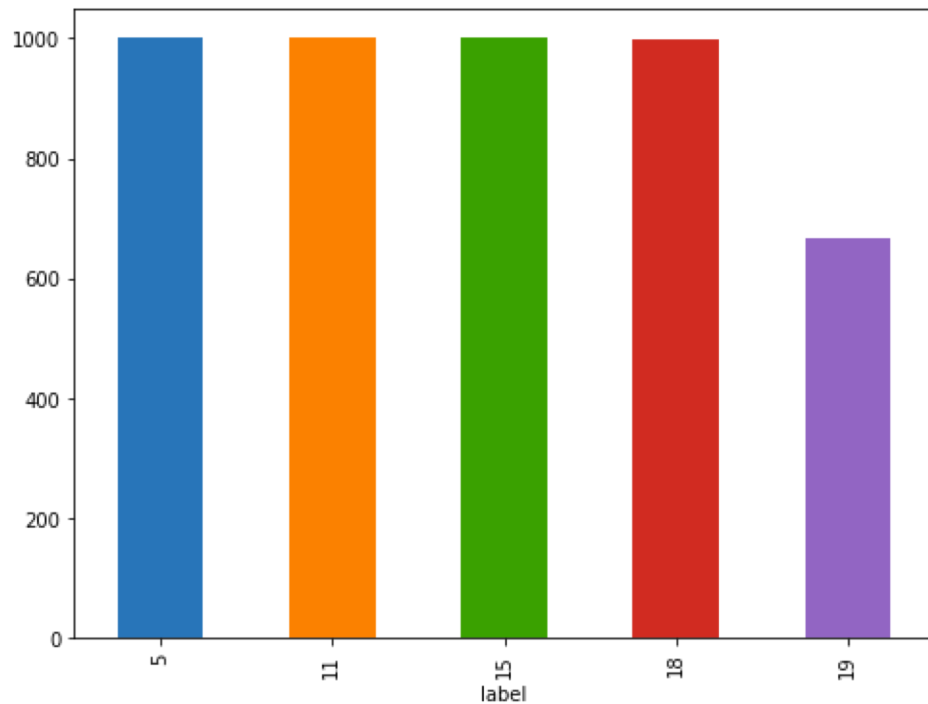


Accuracy

Model	Train	Test*
Naive Bayes Classifier	39%	17%
Random Forest	30%	15%
Logistic Regression	43%	18%

* Test size: 20%

5 Balanced Classes



Accuracy		
Model	Train	Test*
Naive Bayes Classifier	60%	38%
Random Forest	61%	35%
Logistic Regression	65%	39%

* Test size: 20%

Improvements

- Tuning the models parameters properly (GridSearchCV);
- Finding the appropriate metrics to compare the model's performance.