Restaurant Review Rating Prediction

Multiclass Classification Using Spark MLlib

Boris Korotkov

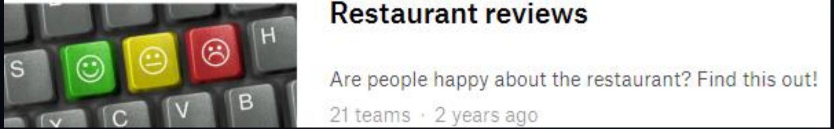
SCS-3252-008

Objective

- Predict Review Rating

Data

kaggle

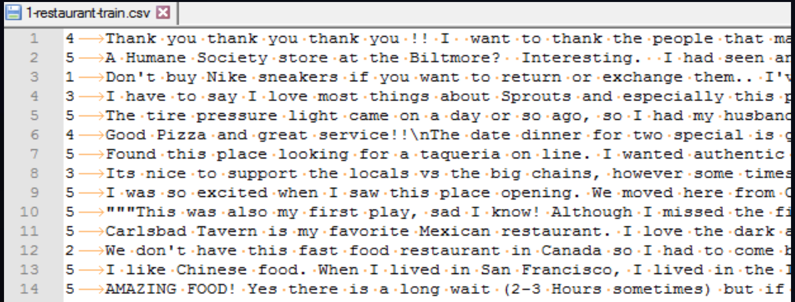


Restaurant reviews

Are people happy about the restaurant? Find this out!

21 teams · 2 years ago

82, 065 records



1-restaurant-train.csv

```
4	Thank you thank you thank you !! I  want to thank the people that ma
5	A Humane Society store at the Biltmore?  Interesting.  I had seen ar
1	Don't buy Nike sneakers if you want to return or exchange them.. I'v
3	I have to say I love most things about Sprouts and especially this p
5	The tire pressure light came on a day or so ago, so I had my husband
4	Good Pizza and great service!!\nThe date dinner for two special is g
5	Found this place looking for a taqueria on line. I wanted authentic 
3	Its nice to support the locals vs the big chains, however some times
5	I was so excited when I saw this place opening. We moved here from C
5	"""This was also my first play, sad I know! Although I missed the fi
5	Carlsbad Tavern is my favorite Mexican restaurant. I love the dark a
2	We don't have this fast food restaurant in Canada so I had to come b
5	I like Chinese food. When I lived in San Francisco, I lived in the I
5	AMAZING FOOD! Yes there is a long wait (2-3 Hours sometimes) but if 
```

Feature Engineering

- Tokenizer
- StopWordsRemover
- HashingTF
- IDF (Inverse document frequency)

Model selection and training

- 80% Training / 20% Test

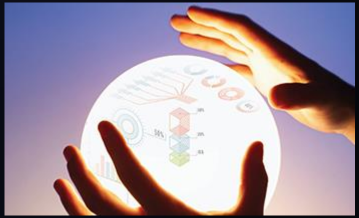
- Random Forest Classifier

- Cross Validation training
  - 5 folds
  - 2 parameters: Max Depth & NumTrees

Model evaluation

- MulticlassClassificationEvaluator
- The accuracy on Test set is 35.08%

Improvement opportunities



- Additional feature engineering
- Different models
- Advanced hyperparameter tuning