amazon-review

May 26, 2022

[ ]:
```
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.pipeline import make_pipeline
```

[ ]:
```
model = make_pipeline(TfidfVectorizer(), MultinomialNB())

train_data_all = pd.read_json("dataset_ja_train.json", orient="records",␣
 ↪lines=True)
train_data = train_data_all["review_body"]
train_target = train_data_all["stars"]

len(train_data_all)
```

[ ]: 200000

[ ]:
```
model.fit(train_data, train_target)
```

[ ]:
```
Pipeline(steps=[('tfidfvectorizer', TfidfVectorizer()),
                ('multinomialnb', MultinomialNB())])
```

[ ]:
```
test_data_all = pd.read_json("dataset_ja_test.json", orient="records",␣
 ↪lines=True)
test_data = test_data_all["review_body"]
test_target = test_data_all["stars"]

len(test_data_all)
```

[ ]: 5000

[ ]:
```
print('Train accuracy = %.3f' % model.score(train_data, train_target))
print(' Test accuracy = %.3f' % model.score(test_data, test_target))
```

```
Train accuracy = 0.992
 Test accuracy = 0.318
```