jupyter JupyterNotebook Last Checkpoint: 12 seconds ago

File Edit View Run Kernel Settings Help

Markdown

0	wesley83 3g iphone 3 hrs tweeting riseaustin d...	Negative	Apple	[wesle
1	jessedee know fludapp awesome ipadiphone app y...	Positive	Apple	[jessedee
2	swonderlin wait ipad 2 also sale sxsw	Positive	Apple	[s
3	sxsw hope years festival isnt crashy years iph...	Negative	Apple	[sxs
4	sxtxstate great stuff fri sxsw marissa mayer g...	Positive	Google	[s›

[45]:
```
df['tokenized_text'].apply(type).value_counts()
```

[45]:
```
<class 'list'>    3123
Name: tokenized_text, dtype: int64
```

Modeling

For our modeling we iterated through a series of different models, lem
find out which combination gave the best results. We picked AUC ROC sc
best, rather than focus on any one label. Our business problem is not
reason we displayed the average AUC ROC scores for different models ac

[70]:
```
# Split the data into training and test sets
X_train, X_test, y_train, y_test = train_test_split(df['tokenized_text

# Define custom pre-processing functions for stemming and lemmatizatio
stemmer = PorterStemmer()
lemmatizer = WordNetLemmatizer()

class TextPreprocessor(FunctionTransformer):
    def __init__(self, method='lemmatize'):
        self.method = method
        super().__init__(validate=None)

    def transform(self, X):
        return X.apply(self._preprocess_text)
```