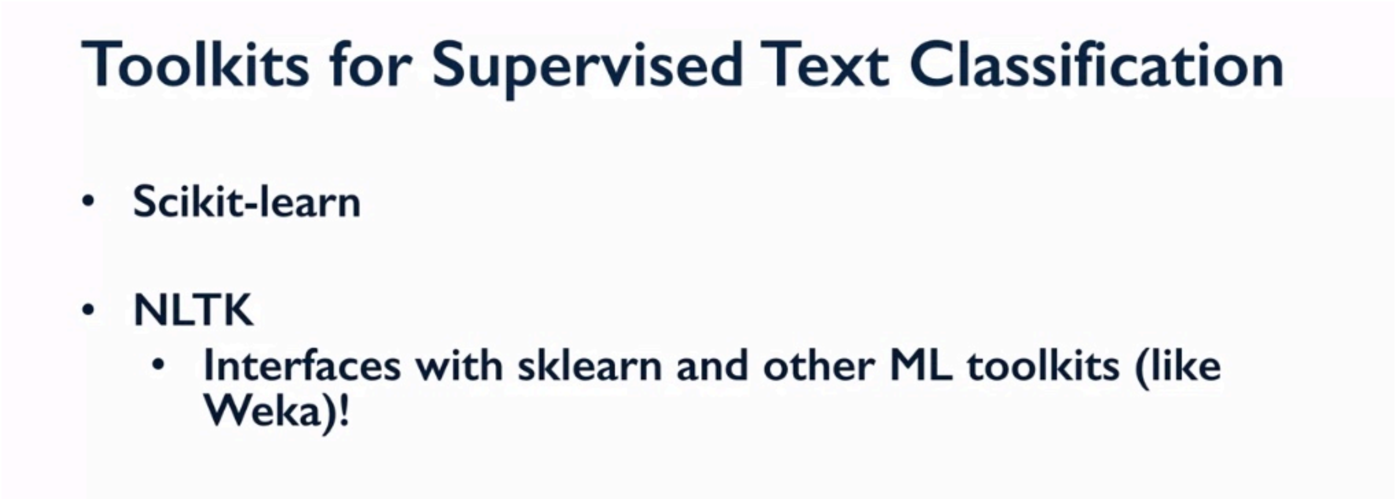
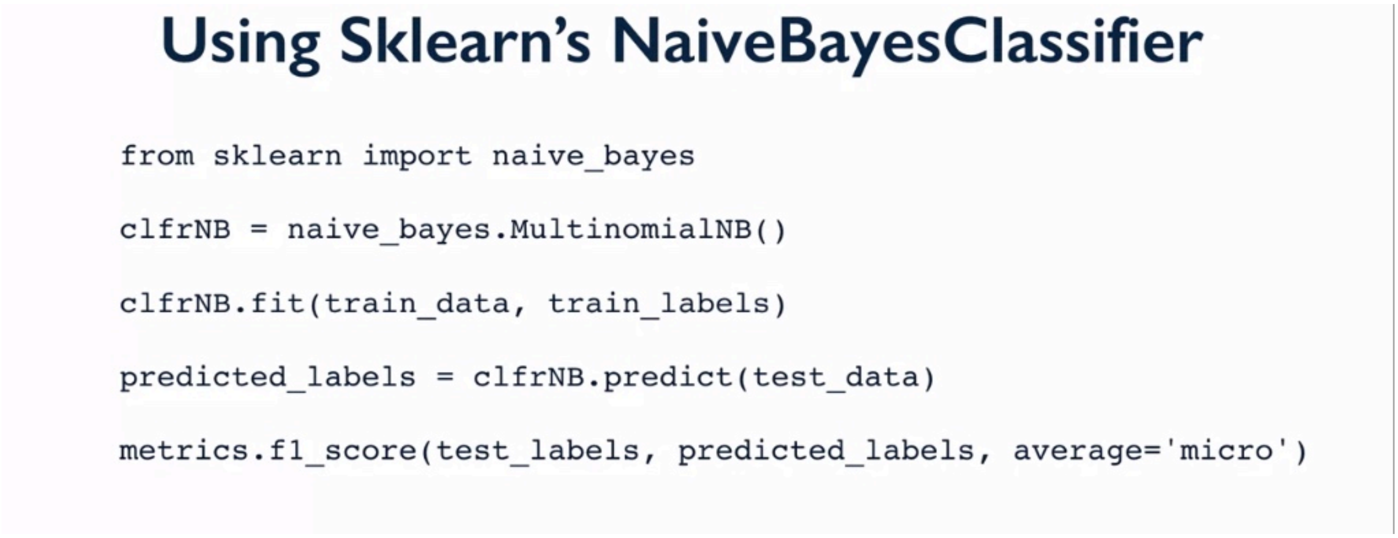
Learning Text Classifiers in Python saved



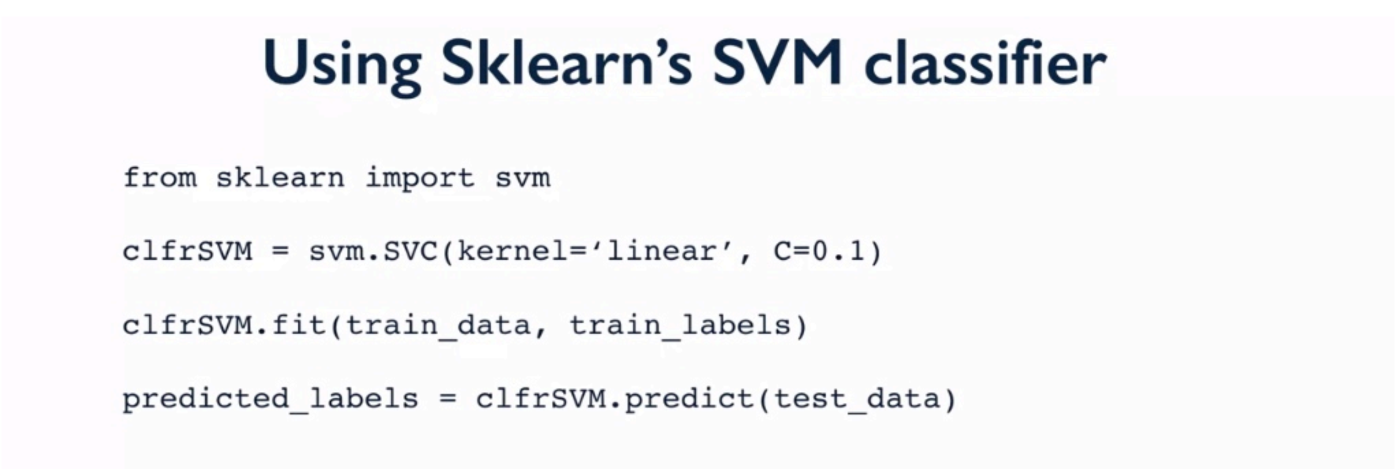
Toolkits for Supervised Text Classification

- Scikit-learn
- NLTK
  - Interfaces with sklearn and other ML toolkits (like Weka)!



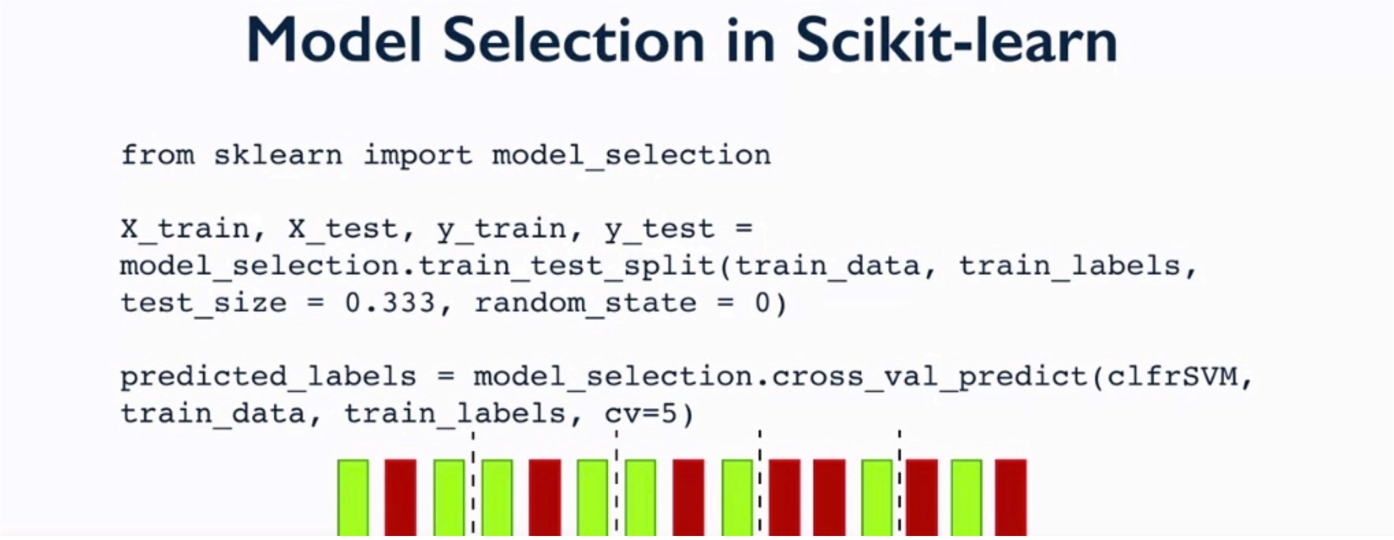
Using Sklearn’s NaiveBayesClassifier

```
from sklearn import naive_bayes

clfrNB = naive_bayes.MultinomialNB()

clfrNB.fit(train_data, train_labels)

predicted_labels = clfrNB.predict(test_data)

metrics.f1_score(test_labels, predicted_labels, average='micro')
```



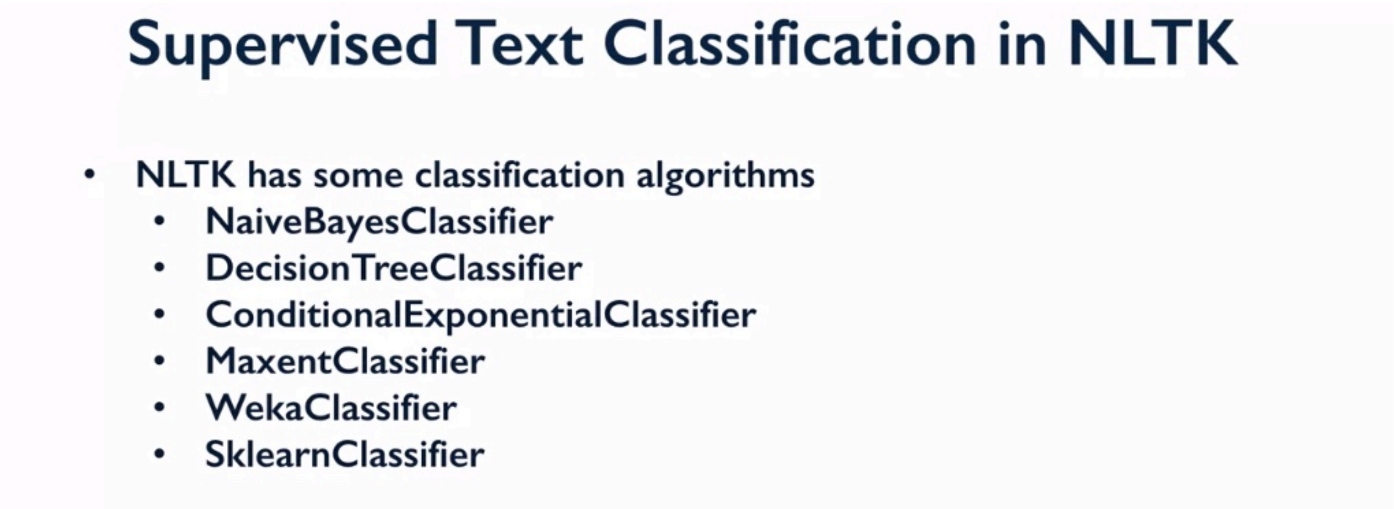
Using Sklearn’s SVM classifier

```
from sklearn import svm

clfrSVM = svm.SVC(kernel=‘linear’, C=0.1)

clfrSVM.fit(train_data, train_labels)

predicted_labels = clfrSVM.predict(test_data)
```



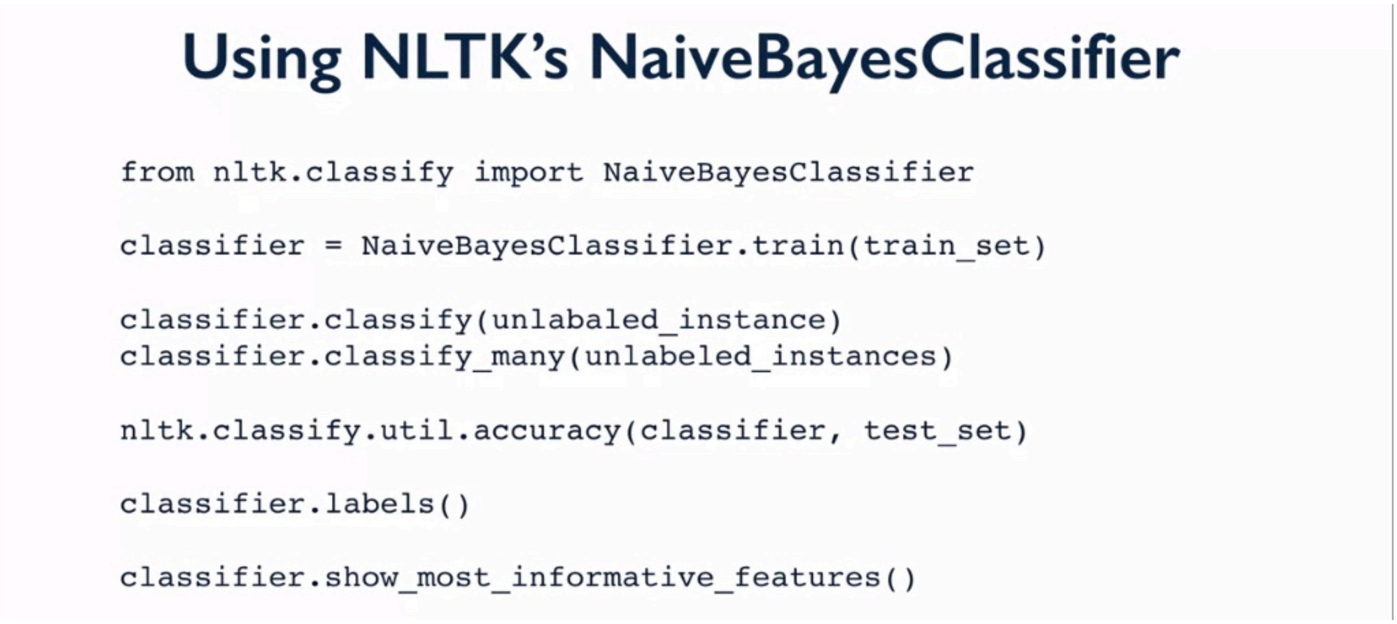
Model Selection in Scikit-learn

```
from sklearn import model_selection

X_train, X_test, y_train, y_test =
model_selection.train_test_split(train_data, train_labels,
test_size = 0.333, random_state = 0)

predicted_labels = model_selection.cross_val_predict(clfrSVM,
train_data, train_labels, cv=5)
```



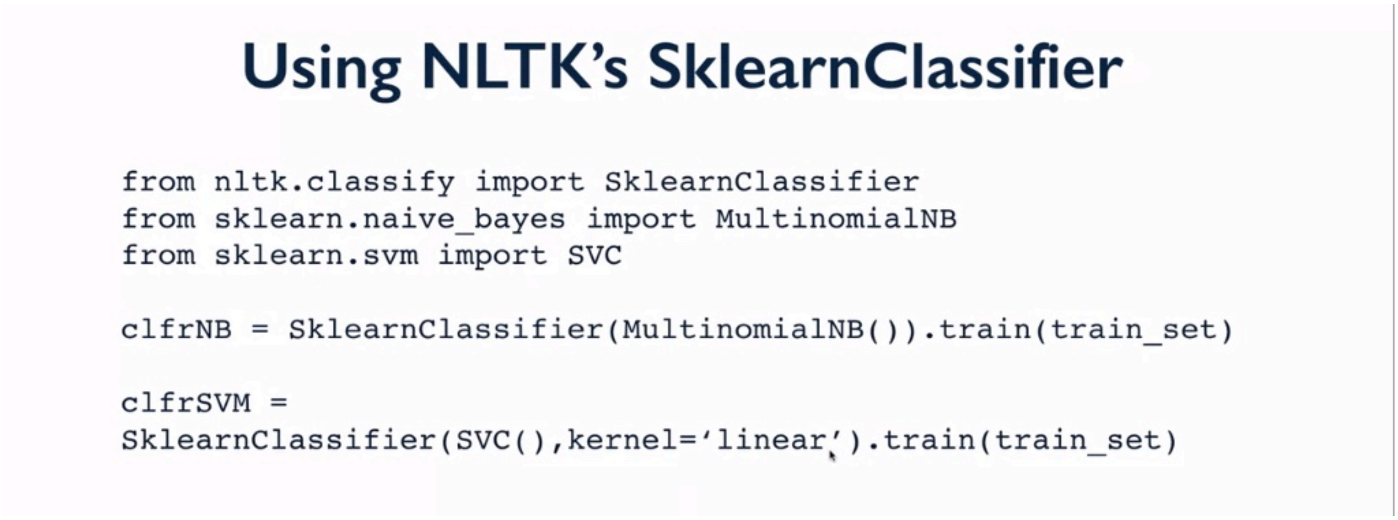
Supervised Text Classification in NLTK

- NLTK has some classification algorithms
  - NaiveBayesClassifier
  - DecisionTreeClassifier
  - ConditionalExponentialClassifier
  - MaxentClassifier
  - WekaClassifier
  - SklearnClassifier



Using NLTK's NaiveBayesClassifier

```
from nltk.classify import NaiveBayesClassifier

classifier = NaiveBayesClassifier.train(train_set)

classifier.classify(unlabaled_instance)
classifier.classify_many(unlabeled_instances)

nltk.classify.util.accuracy(classifier, test_set)

classifier.labels()

classifier.show_most_informative_features()
```



Using NLTK’s SklearnClassifier

```
from nltk.classify import SklearnClassifier
from sklearn.naive_bayes import MultinomialNB
from sklearn.svm import SVC

clfrNB = SklearnClassifier(MultinomialNB()).train(train_set)

clfrSVM =
SklearnClassifier(SVC(),kernel=‘linear’).train(train_set)
```