```
from sklearn.feature_extraction import DictVectorizer
from sklearn.model_selection import train_test_split
from sklearn.svm import SVC
from google.colab import drive
import json
from sklearn.metrics import accuracy_score
```

```
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
actual_tags = []
with open ('/content/drive/MyDrive/Cleaned_Sentences_Task/tags_original.txt', 'r') as file:
    actual_tags = json.load(file)
```

```
removed = []
with open ('/content/drive/MyDrive/Cleaned_Sentences_Task/sentences_without_tags.txt', 'r') as file:
    removed = json.load(file)
```

```
print(len(removed))
```

```
331364
```

```
def extract_features(sentence, index):
  return {
      'word':sentence[index],
      'is_first':index==0,
      'is_last':index ==len(sentence)-1,
      'prefix-1':sentence[index][0],
      'prefix-2':sentence[index][:2],
      'prefix-3':sentence[index][:3],
      'prefix-3':sentence[index][:4],
      'suffix-1':sentence[index][-1],
      'suffix-2':sentence[index][-2:],
      'suffix-3':sentence[index][-3:],
      'suffix-3':sentence[index][-4:],
```

```
    'prev_word':'' if index == 0 else sentence[index-1],
    'next_word':'' if index == 1 else sentence[index+1],
    'has_hyphen': '-' in sentence[index],
    'is_numeric': sentence[index].isdigit()
  }
```

```
def transform_to_dataset(sentences):
  X, y = [], []
  for sents in sentences:
    for index in range(len(sents)):
      X.append(extract_features(sents, index))
  return X, actual_tags[0: 22000]
```

```
X_, y_ = transform_to_dataset(removed[0: 11000])
for i in range(0, 100):
  print(X_[i], " ----------> ", y_[i])
```

```
{'word': 'आग', 'is_first': True, 'is_last': False, 'prefix-1': 'आ', 'prefix-2': 'आग', 'prefix-3': 'आग', 'suffix-1': 'ग',
{'word': 'की', 'is_first': False, 'is_last': True, 'prefix-1': 'क', 'prefix-2': 'की', 'prefix-3': 'की', 'suffix-1': 'ी', 's
{'word': 'लिए', 'is_first': True, 'is_last': False, 'prefix-1': 'ल', 'prefix-2': 'लि', 'prefix-3': 'लिए', 'suffix-1': 'ए', '
{'word': 'सांडिआ', 'is_first': False, 'is_last': True, 'prefix-1': 'स', 'prefix-2': 'सा', 'prefix-3': 'सांड', 'suffix-1': 'आ
{'word': 'की', 'is_first': True, 'is_last': False, 'prefix-1': 'क', 'prefix-2': 'की', 'prefix-3': 'की', 'suffix-1': 'ी', 's
{'word': 'किरण', 'is_first': False, 'is_last': True, 'prefix-1': 'क', 'prefix-2': 'कि', 'prefix-3': 'किरण', 'suffix-1': 'ण
{'word': 'जानी', 'is_first': True, 'is_last': False, 'prefix-1': 'ज', 'prefix-2': 'जा', 'prefix-3': 'जानी', 'suffix-1': 'ी',
{'word': 'चाहिए', 'is_first': False, 'is_last': True, 'prefix-1': 'च', 'prefix-2': 'चा', 'prefix-3': 'चाहि', 'suffix-1': 'ए',
{'word': 'बिहार', 'is_first': True, 'is_last': False, 'prefix-1': 'ब', 'prefix-2': 'बि', 'prefix-3': 'बिहा', 'suffix-1': 'र',
{'word': 'की', 'is_first': False, 'is_last': True, 'prefix-1': 'क', 'prefix-2': 'की', 'prefix-3': 'की', 'suffix-1': 'ी', 's
{'word': 'मोहम्मद', 'is_first': True, 'is_last': False, 'prefix-1': 'म', 'prefix-2': 'मो', 'prefix-3': 'मोहम', 'suffix-1': 'द
{'word': 'हकीम', 'is_first': False, 'is_last': True, 'prefix-1': 'ह', 'prefix-2': 'हक', 'prefix-3': 'हकीम', 'suffix-1': 'म
{'word': 'ने', 'is_first': True, 'is_last': False, 'prefix-1': 'न', 'prefix-2': 'ने', 'prefix-3': 'ने', 'suffix-1': 'े', 'suffi
{'word': 'बताया', 'is_first': False, 'is_last': True, 'prefix-1': 'ब', 'prefix-2': 'बत', 'prefix-3': 'बताय', 'suffix-1': 'ा',
{'word': 'को', 'is_first': True, 'is_last': False, 'prefix-1': 'क', 'prefix-2': 'को', 'prefix-3': 'को', 'suffix-1': 'ो', 's
{'word': 'खत्म', 'is_first': False, 'is_last': True, 'prefix-1': 'ख', 'prefix-2': 'खत', 'prefix-3': 'खत्म', 'suffix-1': 'म',
{'word': 'के', 'is_first': True, 'is_last': False, 'prefix-1': 'क', 'prefix-2': 'के', 'prefix-3': 'के', 'suffix-1': 'े', 'suf
{'word': 'मरने', 'is_first': False, 'is_last': True, 'prefix-1': 'म', 'prefix-2': 'मर', 'prefix-3': 'मरने', 'suffix-1': 'े', '
{'word': 'कभी', 'is_first': True, 'is_last': False, 'prefix-1': 'क', 'prefix-2': 'कभ', 'prefix-3': 'कभी', 'suffix-1': 'ी',
{'word': 'कुछ', 'is_first': False, 'is_last': True, 'prefix-1': 'क', 'prefix-2': 'कु', 'prefix-3': 'कुछ', 'suffix-1': 'छ',
{'word': 'लादेन', 'is_first': True, 'is_last': False, 'prefix-1': 'ल', 'prefix-2': 'ला', 'prefix-3': 'लादे', 'suffix-1': 'न',
{'word': 'की', 'is_first': False, 'is_last': True, 'prefix-1': 'क', 'prefix-2': 'की', 'prefix-3': 'की', 'suffix-1': 'ी', 's
{'word': 'रहे', 'is_first': True, 'is_last': False, 'prefix-1': 'र', 'prefix-2': 'रह', 'prefix-3': 'रहे', 'suffix-1': 'े', 'su
```

```
{'word': 'हैं', 'is_first': False, 'is_last': True, 'prefix-1': 'ह', 'prefix-2': 'है', 'prefix-3': 'हैं', 'suffix-1': 'ं', 'suffi
{'word': 'लिए', 'is_first': True, 'is_last': False, 'prefix-1': 'ल', 'prefix-2': 'लि', 'prefix-3': 'लिए', 'suffix-1': 'ए', '
{'word': 'बचाने', 'is_first': False, 'is_last': True, 'prefix-1': 'ब', 'prefix-2': 'बच', 'prefix-3': 'बचान', 'suffix-1': 'े',
{'word': 'कांग्रेस', 'is_first': True, 'is_last': False, 'prefix-1': 'क', 'prefix-2': 'का', 'prefix-3': 'कांग', 'suffix-1': 'स',
{'word': 'में', 'is_first': False, 'is_last': True, 'prefix-1': 'म', 'prefix-2': 'मे', 'prefix-3': 'में', 'suffix-1': 'ं', 'suffi
{'word': 'युवाओं', 'is_first': True, 'is_last': False, 'prefix-1': 'य', 'prefix-2': 'यु', 'prefix-3': 'युवा', 'suffix-1': 'ं', '
{'word': 'की', 'is_first': False, 'is_last': True, 'prefix-1': 'क', 'prefix-2': 'की', 'prefix-3': 'की', 'suffix-1': 'ी', 's
{'word': 'जारी', 'is_first': True, 'is_last': False, 'prefix-1': 'ज', 'prefix-2': 'जा', 'prefix-3': 'जारी', 'suffix-1': 'ी',
{'word': 'करने', 'is_first': False, 'is_last': True, 'prefix-1': 'क', 'prefix-2': 'कर', 'prefix-3': 'करने', 'suffix-1': 'े',
{'word': 'मसले', 'is_first': True, 'is_last': False, 'prefix-1': 'म', 'prefix-2': 'मस', 'prefix-3': 'मसले', 'suffix-1': 'े',
{'word': 'को', 'is_first': False, 'is_last': True, 'prefix-1': 'क', 'prefix-2': 'को', 'prefix-3': 'को', 'suffix-1': 'ो', 's
{'word': 'के', 'is_first': True, 'is_last': False, 'prefix-1': 'क', 'prefix-2': 'के', 'prefix-3': 'के', 'suffix-1': 'े', 'suf
{'word': 'लोगों', 'is_first': False, 'is_last': True, 'prefix-1': 'ल', 'prefix-2': 'लो', 'prefix-3': 'लोगो', 'suffix-1': 'ं', '
{'word': 'रही', 'is_first': True, 'is_last': False, 'prefix-1': 'र', 'prefix-2': 'रह', 'prefix-3': 'रही', 'suffix-1': 'ी', 's
{'word': 'है', 'is_first': False, 'is_last': True, 'prefix-1': 'ह', 'prefix-2': 'है', 'prefix-3': 'है', 'suffix-1': 'ै', 'suffi
{'word': 'को', 'is_first': True, 'is_last': False, 'prefix-1': 'क', 'prefix-2': 'को', 'prefix-3': 'को', 'suffix-1': 'ो', 's
{'word': 'गहरा', 'is_first': False, 'is_last': True, 'prefix-1': 'ग', 'prefix-2': 'गह', 'prefix-3': 'गहरा', 'suffix-1': 'ा',
{'word': 'अभी', 'is_first': True, 'is_last': False, 'prefix-1': 'अ', 'prefix-2': 'अभ', 'prefix-3': 'अभी', 'suffix-1': 'ी',
{'word': 'इसका', 'is_first': False, 'is_last': True, 'prefix-1': 'इ', 'prefix-2': 'इस', 'prefix-3': 'इसका', 'suffix-1': 'ा'
{'word': 'फांसी', 'is_first': True, 'is_last': False, 'prefix-1': 'फ', 'prefix-2': 'फा', 'prefix-3': 'फांस', 'suffix-1': 'ी',
{'word': 'लग', 'is_first': False, 'is_last': True, 'prefix-1': 'ल', 'prefix-2': 'लग', 'prefix-3': 'लग', 'suffix-1': 'ग', '
{'word': 'चढ़त', 'is_first': True, 'is_last': False, 'prefix-1': 'च', 'prefix-2': 'चढ़', 'prefix-3': 'चढ़त', 'suffix-1': 'त',
{'word': 'के', 'is_first': False, 'is_last': True, 'prefix-1': 'क', 'prefix-2': 'के', 'prefix-3': 'के', 'suffix-1': 'े', 'suf
{'word': 'साथ', 'is_first': True, 'is_last': False, 'prefix-1': 'स', 'prefix-2': 'सा', 'prefix-3': 'साथ', 'suffix-1': 'थ', '
{'word': 'समूचे', 'is_first': False, 'is_last': True, 'prefix-1': 'स', 'prefix-2': 'सम', 'prefix-3': 'समूच', 'suffix-1': 'े',
{'word': 'घुसा', 'is_first': True, 'is_last': False, 'prefix-1': 'घ', 'prefix-2': 'घु', 'prefix-3': 'घुसा', 'suffix-1': 'ा', 's
{'word': 'NULL', 'is_first': False, 'is_last': True, 'prefix-1': 'N', 'prefix-2': 'NU', 'prefix-3': 'NULL', 'suffix-1': 'L',
{'word': 'सफल', 'is_first': True, 'is_last': False, 'prefix-1': 'स', 'prefix-2': 'सफ', 'prefix-3': 'सफल', 'suffix-1': 'ल
{'word': 'रहे', 'is_first': False, 'is_last': True, 'prefix-1': 'र', 'prefix-2': 'रह', 'prefix-3': 'रहे', 'suffix-1': 'े', 'su
{'word': '२६९९.२०', 'is_first': True, 'is_last': False, 'prefix-1': '२', 'prefix-2': '२६', 'prefix-3': '२६९९', 'suffix-1': '
{'word': 'करोड़', 'is_first': False, 'is_last': True, 'prefix-1': 'क', 'prefix-2': 'कर', 'prefix-3': 'करोड़', 'suffix-1': 'ड़
{'word': 'केंद्र', 'is_first': True, 'is_last': False, 'prefix-1': 'क', 'prefix-2': 'के', 'prefix-3': 'केंद', 'suffix-1': 'र', 's
{'word': 'के', 'is_first': False, 'is_last': True, 'prefix-1': 'क', 'prefix-2': 'के', 'prefix-3': 'के', 'suffix-1': 'े', 'suf
{'word': 'भारी', 'is_first': True, 'is_last': False, 'prefix-1': 'भ', 'prefix-2': 'भा', 'prefix-3': 'भारी', 'suffix-1': 'ी', '
{'word': 'क्षति', 'is_first': False, 'is_last': True, 'prefix-1': 'क', 'prefix-2': 'क', 'prefix-3': 'क्षत', 'suffix-1': 'ि', 's
```

```
1 print(len(X_), len(y_))
```

```
22000 22000
```

```
print(y_[0: 10])
print(actual_tags[0: 10])
```

```
['nn', 'psp', 'psp', 'nnpc', 'psp', 'nnpc', 'vaux', 'vaux', 'nnp', 'psp']
['nn', 'psp', 'psp', 'nnpc', 'psp', 'nnpc', 'vaux', 'vaux', 'nnp', 'psp']
```

```
X_train, X_test, y_train, y_test = train_test_split(X_, y_, train_size=0.75)
```

```
dict_vectorizer = DictVectorizer(sparse=False)
dict_vectorizer.fit(X_)
X_train = dict_vectorizer.transform(X_train)
X_test = dict_vectorizer.transform(X_test)
```

```
X_ = dict_vectorizer.transform(X_)
```

```
model_ml = SVC()
```

```
model_ml.fit(X_train, y_train)
print("Accuracy: {}".format(model_ml.score(X_test, y_test)))
```

```
Accuracy: 81.81818181
```

```
final_predictions_ml = []
for i in X_test:
  final_predictions_ml.extend(model_ml.predict([i]))
```

```
ml_tags = []
for i in X_:
  ml_tags.extend(model_ml.predict([i]))
```

```
with open ('/content/drive/MyDrive/Cleaned_Sentences_Task/svm_predictions.txt', 'w+') as file:
    json.dump(ml_tags, file)
```