ML Pipeline Preparation

Follow the instructions below to help you create your ML pipeline.

1. Import libraries and load data from database.

- Import Python libraries
- Load dataset from database with `read_sql_table` (https://pandas.pydata.org/pandas-docs/stable/generated/pandas.read_sql_table.html)
- Define feature and target variables X and Y

In [ ]:

```

```

In [34]:

```
# import libraries
import nltk
nltk.download('punkt')
nltk.download('wordnet')
#nltk.download()
import pandas as pd
from sqlalchemy import create_engine
from nltk.tokenize import word_tokenize
from nltk.stem.porter import PorterStemmer
from nltk.stem.wordnet import WordNetLemmatizer
from sklearn.feature_extraction.text import TfidfVectorizer
from nltk.corpus import stopwords
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
import re
import numpy as np
from sklearn.metrics import confusion_matrix
from sklearn.pipeline import Pipeline, FeatureUnion
from sklearn.feature_extraction.text import CountVectorizer, TfidfTransformer
from sklearn import multioutput
from sklearn.metrics import classification_report
from sklearn.model_selection import GridSearchCV
from sklearn.linear_model import SGDClassifier
from sklearn.svm import SVC
nltk.download('stopwords')
from self_transformers import StartingVerbExtractor
from sklearn.metrics import fbeta_score, make_scorer
import pickle

import matplotlib.pyplot as plt
%matplotlib inline
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]     /root/nltk_data...
[nltk_data]   Unzipping taggers/averaged_perceptron_tagger.zip.
```

In [2]:

```
# load data from database
engine = create_engine('sqlite:///DisasterResponse.db')
df = pd.read_sql ('SELECT * FROM MessagesCategories', engine)
#display (df.head (n=10))
X = df ['message']
y = df.iloc[:,4:]
y.head (n=3)
```

Out[2]:

	related	request	offer	aid_related	medical_help	medical_products	search_and_rescue	se
0	1	0	0	0	0	0	0	
1	1	0	0	1	0	0	0	
2	1	0	0	0	0	0	0	

3 rows × 36 columns

2. Write a tokenization function to process your text data

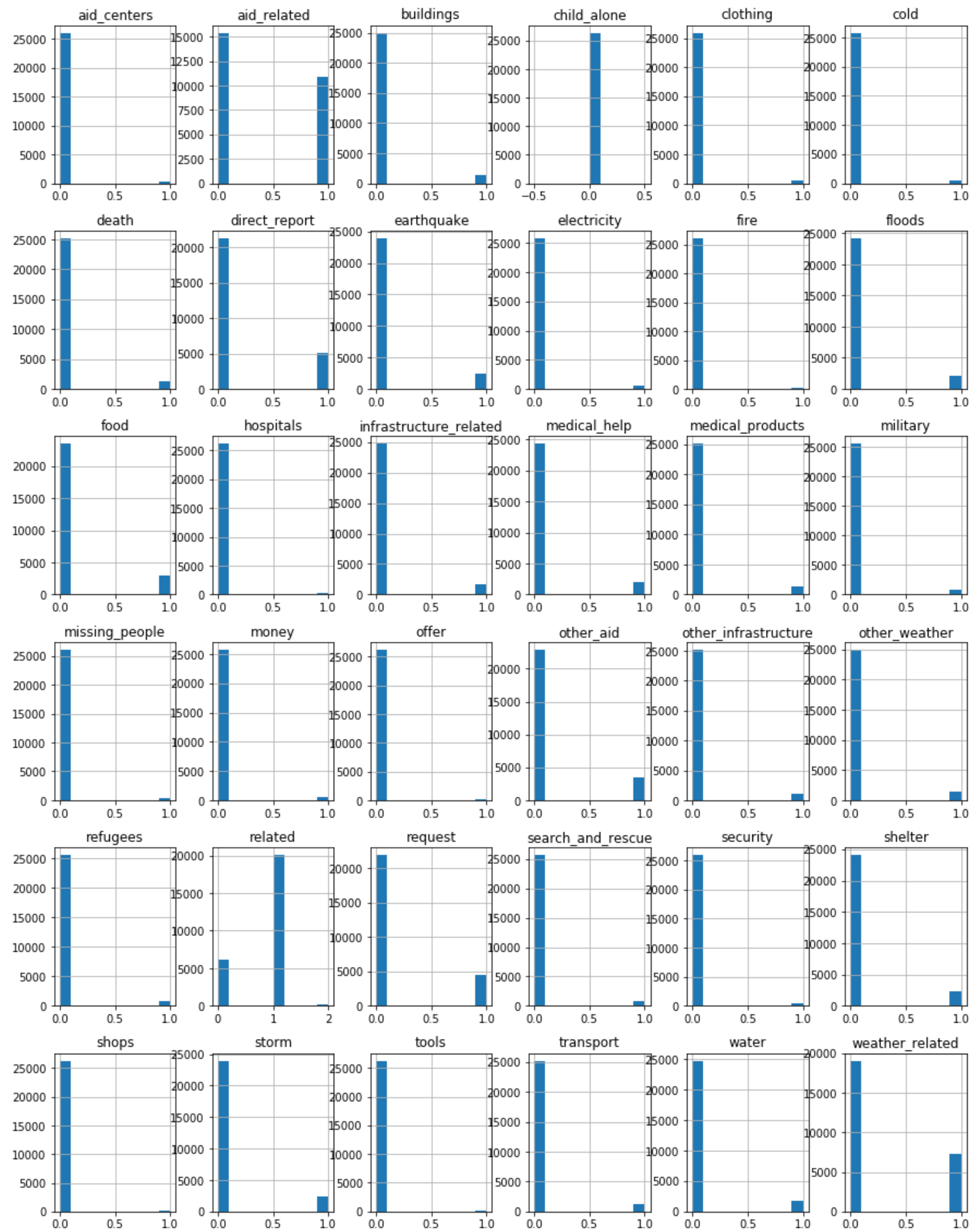
In [3]:

```
def tokenize(text):
    """Tokenization function. Receives as input raw text which afterwards normalized, s
top words removed, stemmed and lemmatized.
    Returns tokenized text"""

    # Normalize text
    text = re.sub(r"[^a-zA-Z0-9]", " ", text.lower())

    stop_words = stopwords.words("english")

    #tokenize
    words = word_tokenize (text)

    #stemming
    stemmed = [PorterStemmer().stem(w) for w in words]

    #lemmatizing
    words_lemmed = [WordNetLemmatizer().lemmatize(w) for w in stemmed if w not in stop_
words]

    return words_lemmed
```

In [4]:

```
#let's take a look to the possible values distribution within classes

#making size of figure bigger
fig = plt.figure(figsize = (15,20))
ax = fig.gca()
y.hist(ax = ax)
plt.show()
```

```
/opt/conda/lib/python3.6/site-packages/IPython/core/interactiveshell.py:29
61: UserWarning: To output multiple subplots, the figure containing the pa
ssed axes is being cleared
  exec(code_obj, self.user_global_ns, self.user_ns)
```



3. Build a machine learning pipeline

This machine pipeline should take in the `message` column as input and output classification results on the other 36 categories in the dataset. You may find the [MultiOutputClassifier (http://scikit-learn.org/stable/modules/generated/sklearn.multioutput.MultiOutputClassifier.html)](http://scikit-learn.org/stable/modules/generated/sklearn.multioutput.MultiOutputClassifier.html) helpful for predicting multiple target variables.

In [5]:

```
#setting pipeline
pipeline = Pipeline([
        ('vect', CountVectorizer(tokenizer=tokenize)),
        ('tfidf', TfidfTransformer()),
#        ('clf', multioutput.MultiOutputClassifier (RandomForestClassifier(), n_jobs = 
35))
        ('clf', multioutput.MultiOutputClassifier (RandomForestClassifier()))
        ])
```

In [ ]:

```

```

4. Train pipeline

- Split data into train and test sets
- Train pipeline

In [6]:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state = 22)
```

In [7]:

```
# train classifier
pipeline.fit(X_train, y_train)
```

Out[7]:

```
Pipeline(memory=None,
     steps=[('vect', CountVectorizer(analyzer='word', binary=False, decode
_error='strict',
        dtype=<class 'numpy.int64'>, encoding='utf-8', input='content',
        lowercase=True, max_df=1.0, max_features=None, min_df=1,
        ngram_range=(1, 1), preprocessor=None, stop_words=None,
        strip...oob_score=False, random_state=None, verbose=0,
            warm_start=False),
           n_jobs=1))])
```

5. Test your model

Report the f1 score, precision and recall for each output category of the dataset. You can do this by iterating through the columns and calling sklearn's `classification_report` on each.

In [8]:

```
!pip install -U scikit-learn
```

```
Collecting scikit-learn
  Downloading https://files.pythonhosted.org/packages/f5/ef/bcd79e8d59250d
6e8478eb1290dc6e05be42b3be8a86e3954146adbc171a/scikit_learn-0.24.2-cp36-cp
36m-manylinux1_x86_64.whl (20.0MB)
    100% |████████████████████████████████| 20.0MB 1.6MB/s eta 0:00:01
Requirement already satisfied, skipping upgrade: scipy>=0.19.1 in /opt/con
da/lib/python3.6/site-packages (from scikit-learn) (1.2.1)
Collecting threadpoolctl>=2.0.0 (from scikit-learn)
  Downloading https://files.pythonhosted.org/packages/61/cf/6e354304bcb9c6
413c4e02a747b600061c21d38ba51e7e544ac7bc66aecc/threadpoolctl-3.1.0-py3-non
e-any.whl
Collecting numpy>=1.13.3 (from scikit-learn)
  Downloading https://files.pythonhosted.org/packages/45/b2/6c7545bb7a3875
4d63048c7696804a0d947328125d81bf12beaa692c3ae3/numpy-1.19.5-cp36-cp36m-man
ylinux1_x86_64.whl (13.4MB)
    100% |████████████████████████████████| 13.4MB 1.7MB/s eta 0:00:01
Requirement already satisfied, skipping upgrade: joblib>=0.11 in /opt/cond
a/lib/python3.6/site-packages (from scikit-learn) (0.11)
tensorflow 1.3.0 requires tensorflow-tensorboard<0.2.0,>=0.1.0, which is n
ot installed.
Installing collected packages: threadpoolctl, numpy, scikit-learn
  Found existing installation: numpy 1.12.1
    Uninstalling numpy-1.12.1:
      Successfully uninstalled numpy-1.12.1
  Found existing installation: scikit-learn 0.19.1
    Uninstalling scikit-learn-0.19.1:
      Successfully uninstalled scikit-learn-0.19.1
Successfully installed numpy-1.19.5 scikit-learn-0.24.2 threadpoolctl-3.1.
0
```

In [20]:

```
y_pred = pipeline.predict(X_test)
    # print the metrics
category_names = list(df.columns[4:])
for i, col in enumerate(category_names):
    print('{} category metrics: '.format(col))
    print(classification_report(y_test.iloc[:,i], y_pred[:,i]))
```

```
related category metrics:
             precision    recall  f1-score   support

          0       0.64      0.47      0.54      1541
          1       0.85      0.92      0.88      4991
          2       0.36      0.35      0.36        65

avg / total       0.79      0.81      0.80      6597

request category metrics:
             precision    recall  f1-score   support

          0       0.90      0.98      0.93      5480
          1       0.78      0.44      0.56      1117

avg / total       0.88      0.88      0.87      6597

offer category metrics:
             precision    recall  f1-score   support

          0       0.99      1.00      1.00      6561
          1       1.00      0.06      0.11        36

avg / total       0.99      0.99      0.99      6597

aid_related category metrics:
             precision    recall  f1-score   support

          0       0.75      0.86      0.80      3857
          1       0.75      0.59      0.66      2740

avg / total       0.75      0.75      0.74      6597

medical_help category metrics:
             precision    recall  f1-score   support

          0       0.93      0.99      0.96      6083
          1       0.55      0.10      0.17       514

avg / total       0.90      0.92      0.90      6597

medical_products category metrics:
             precision    recall  f1-score   support

          0       0.96      1.00      0.98      6280
          1       0.79      0.10      0.17       317

avg / total       0.95      0.96      0.94      6597

search_and_rescue category metrics:
             precision    recall  f1-score   support

          0       0.97      1.00      0.99      6415
          1       0.50      0.04      0.07       182

avg / total       0.96      0.97      0.96      6597

security category metrics:
             precision    recall  f1-score   support

          0       0.98      1.00      0.99      6474
```

```
          1       0.00      0.00      0.00       123

avg / total       0.96      0.98      0.97      6597

military category metrics:
             precision    recall  f1-score   support

          0       0.97      1.00      0.98      6374
          1       0.68      0.08      0.14       223

avg / total       0.96      0.97      0.95      6597

child_alone category metrics:
             precision    recall  f1-score   support

          0       1.00      1.00      1.00      6597

avg / total       1.00      1.00      1.00      6597

water category metrics:
             precision    recall  f1-score   support

          0       0.96      1.00      0.98      6214
          1       0.80      0.31      0.45       383

avg / total       0.95      0.96      0.95      6597

food category metrics:
             precision    recall  f1-score   support

          0       0.94      0.99      0.96      5868
          1       0.82      0.47      0.59       729

avg / total       0.92      0.93      0.92      6597

shelter category metrics:
             precision    recall  f1-score   support

          0       0.94      0.99      0.96      6011
          1       0.79      0.34      0.48       586

avg / total       0.93      0.93      0.92      6597

clothing category metrics:
             precision    recall  f1-score   support

          0       0.99      1.00      0.99      6486
          1       0.84      0.19      0.31       111

avg / total       0.98      0.99      0.98      6597

money category metrics:
             precision    recall  f1-score   support

          0       0.98      1.00      0.99      6452
          1       0.69      0.06      0.11       145

avg / total       0.97      0.98      0.97      6597

missing_people category metrics:
             precision    recall  f1-score   support
```

```
           0       0.99      1.00      1.00      6533
           1       0.00      0.00      0.00        64

 avg / total       0.98      0.99      0.99      6597

 refugees category metrics:
              precision    recall  f1-score   support

           0       0.97      1.00      0.98      6377
           1       0.62      0.08      0.14       220

 avg / total       0.96      0.97      0.96      6597

 death category metrics:
              precision    recall  f1-score   support

           0       0.96      1.00      0.98      6289
           1       0.79      0.16      0.26       308

 avg / total       0.95      0.96      0.95      6597

 other_aid category metrics:
              precision    recall  f1-score   support

           0       0.88      0.99      0.93      5744
           1       0.47      0.06      0.10       853

 avg / total       0.82      0.87      0.82      6597

 infrastructure_related category metrics:
              precision    recall  f1-score   support

           0       0.94      1.00      0.97      6207
           1       0.12      0.01      0.01       390

 avg / total       0.89      0.94      0.91      6597

 transport category metrics:
              precision    recall  f1-score   support

           0       0.96      1.00      0.98      6292
           1       0.67      0.07      0.12       305

 avg / total       0.94      0.96      0.94      6597

 buildings category metrics:
              precision    recall  f1-score   support

           0       0.95      1.00      0.98      6266
           1       0.75      0.11      0.19       331

 avg / total       0.94      0.95      0.94      6597

 electricity category metrics:
              precision    recall  f1-score   support

           0       0.98      1.00      0.99      6478
           1       0.73      0.07      0.12       119

 avg / total       0.98      0.98      0.98      6597
```

```
tools category metrics:
              precision    recall  f1-score   support

          0       0.99      1.00      1.00      6562
          1       0.00      0.00      0.00        35

avg / total       0.99      0.99      0.99      6597

hospitals category metrics:
              precision    recall  f1-score   support

          0       0.99      1.00      0.99      6530
          1       0.00      0.00      0.00        67

avg / total       0.98      0.99      0.98      6597

shops category metrics:
              precision    recall  f1-score   support

          0       1.00      1.00      1.00      6572
          1       0.00      0.00      0.00        25

avg / total       0.99      1.00      0.99      6597

aid_centers category metrics:
              precision    recall  f1-score   support

          0       0.99      1.00      0.99      6526
          1       0.00      0.00      0.00        71

avg / total       0.98      0.99      0.98      6597

other_infrastructure category metrics:
              precision    recall  f1-score   support

          0       0.96      1.00      0.98      6331
          1       0.00      0.00      0.00       266

avg / total       0.92      0.96      0.94      6597

weather_related category metrics:
              precision    recall  f1-score   support

          0       0.86      0.95      0.91      4755
          1       0.84      0.61      0.70      1842

avg / total       0.86      0.86      0.85      6597

floods category metrics:
              precision    recall  f1-score   support

          0       0.95      1.00      0.97      6079
          1       0.90      0.37      0.53       518

avg / total       0.95      0.95      0.94      6597

storm category metrics:
              precision    recall  f1-score   support

          0       0.95      0.98      0.97      5998
```

```
          1       0.73      0.46      0.57       599

avg / total       0.93      0.94      0.93      6597

fire category metrics:
             precision    recall  f1-score   support

          0       0.99      1.00      0.99      6526
          1       1.00      0.01      0.03        71

avg / total       0.99      0.99      0.98      6597

earthquake category metrics:
             precision    recall  f1-score   support

          0       0.97      0.99      0.98      5990
          1       0.89      0.74      0.81       607

avg / total       0.97      0.97      0.97      6597

cold category metrics:
             precision    recall  f1-score   support

          0       0.98      1.00      0.99      6458
          1       0.80      0.06      0.11       139

avg / total       0.98      0.98      0.97      6597

other_weather category metrics:
             precision    recall  f1-score   support

          0       0.95      1.00      0.97      6252
          1       0.42      0.03      0.05       345

avg / total       0.92      0.95      0.92      6597

direct_report category metrics:
             precision    recall  f1-score   support

          0       0.85      0.97      0.91      5302
          1       0.75      0.31      0.44      1295

avg / total       0.83      0.84      0.82      6597

/opt/conda/lib/python3.6/site-packages/sklearn/metrics/classification.py:1
135: UndefinedMetricWarning: Precision and F-score are ill-defined and bei
ng set to 0.0 in labels with no predicted samples.
```

6. Improve your model

Use grid search to find better parameters.

In [24]:

```
# fbeta_score scoring object using make_scorer()
#scorer = make_scorer (f1_scorer_evaluate)

parameters = { 'vect__max_df': (0.75, 1.0),
             #  'vect__stop_words': ('english', None),
                'clf__estimator__n_estimators': [10, 20],
                'clf__estimator__min_samples_split': [2, 5]
              }

cv = GridSearchCV (pipeline, param_grid= parameters, verbose =3 )
```

In [25]:

```
model = cv
```

In [26]:

```
model.fit(X_train, y_train)
```

```
Fitting 3 folds for each of 8 candidates, totalling 24 fits
[CV] clf__estimator__min_samples_split=2, clf__estimator__n_estimators=10, 
vect__max_df=0.75 
[CV]  clf__estimator__min_samples_split=2, clf__estimator__n_estimators=1
0, vect__max_df=0.75, score=0.23040776110353192, total= 1.1min
[CV] clf__estimator__min_samples_split=2, clf__estimator__n_estimators=10, 
vect__max_df=0.75 

[Parallel(n_jobs=1)]: Done   1 out of   1 | elapsed:  1.5min remaining:
0.0s

[CV]  clf__estimator__min_samples_split=2, clf__estimator__n_estimators=1
0, vect__max_df=0.75, score=0.24514857489387507, total= 1.1min
[CV] clf__estimator__min_samples_split=2, clf__estimator__n_estimators=10, 
vect__max_df=0.75 

[Parallel(n_jobs=1)]: Done   2 out of   2 | elapsed:  3.0min remaining:
0.0s

[CV]  clf__estimator__min_samples_split=2, clf__estimator__n_estimators=1
0, vect__max_df=0.75, score=0.2427228623408126, total= 1.1min
[CV] clf__estimator__min_samples_split=2, clf__estimator__n_estimators=10, 
vect__max_df=1.0 

[Parallel(n_jobs=1)]: Done   3 out of   3 | elapsed:  4.4min remaining:
0.0s

[CV]  clf__estimator__min_samples_split=2, clf__estimator__n_estimators=1
0, vect__max_df=1.0, score=0.23010459299681674, total= 1.0min
[CV] clf__estimator__min_samples_split=2, clf__estimator__n_estimators=10, 
vect__max_df=1.0 

[Parallel(n_jobs=1)]: Done   4 out of   4 | elapsed:  5.8min remaining:
0.0s

[CV]  clf__estimator__min_samples_split=2, clf__estimator__n_estimators=1
0, vect__max_df=1.0, score=0.24620982413583992, total= 1.1min
[CV] clf__estimator__min_samples_split=2, clf__estimator__n_estimators=10, 
vect__max_df=1.0 

[Parallel(n_jobs=1)]: Done   5 out of   5 | elapsed:  7.3min remaining:
0.0s

[CV]  clf__estimator__min_samples_split=2, clf__estimator__n_estimators=1
0, vect__max_df=1.0, score=0.24408732565191024, total= 1.1min
[CV] clf__estimator__min_samples_split=2, clf__estimator__n_estimators=20, 
vect__max_df=0.75 

[Parallel(n_jobs=1)]: Done   6 out of   6 | elapsed:  8.8min remaining:
0.0s
```

```
[CV]  clf__estimator__min_samples_split=2, clf__estimator__n_estimators=2
0, vect__max_df=0.75, score=0.24344398969228437, total= 1.5min
[CV] clf__estimator__min_samples_split=2, clf__estimator__n_estimators=20,
vect__max_df=0.75
[CV]  clf__estimator__min_samples_split=2, clf__estimator__n_estimators=2
0, vect__max_df=0.75, score=0.25651910248635534, total= 1.6min
[CV] clf__estimator__min_samples_split=2, clf__estimator__n_estimators=20,
vect__max_df=0.75
[CV]  clf__estimator__min_samples_split=2, clf__estimator__n_estimators=2
0, vect__max_df=0.75, score=0.25272892662219526, total= 1.6min
[CV] clf__estimator__min_samples_split=2, clf__estimator__n_estimators=20,
vect__max_df=1.0
[CV]  clf__estimator__min_samples_split=2, clf__estimator__n_estimators=2
0, vect__max_df=1.0, score=0.24283765347885403, total= 1.6min
[CV] clf__estimator__min_samples_split=2, clf__estimator__n_estimators=20,
vect__max_df=1.0
[CV]  clf__estimator__min_samples_split=2, clf__estimator__n_estimators=2
0, vect__max_df=1.0, score=0.2548514251061249, total= 1.6min
[CV] clf__estimator__min_samples_split=2, clf__estimator__n_estimators=20,
vect__max_df=1.0
[CV]  clf__estimator__min_samples_split=2, clf__estimator__n_estimators=2
0, vect__max_df=1.0, score=0.26303820497271074, total= 1.6min
[CV] clf__estimator__min_samples_split=5, clf__estimator__n_estimators=10,
vect__max_df=0.75
[CV]  clf__estimator__min_samples_split=5, clf__estimator__n_estimators=1
0, vect__max_df=0.75, score=0.23025617705017432, total=  58.1s
[CV] clf__estimator__min_samples_split=5, clf__estimator__n_estimators=10,
vect__max_df=0.75
[CV]  clf__estimator__min_samples_split=5, clf__estimator__n_estimators=1
0, vect__max_df=0.75, score=0.2348392965433596, total= 1.0min
[CV] clf__estimator__min_samples_split=5, clf__estimator__n_estimators=10,
vect__max_df=0.75
[CV]  clf__estimator__min_samples_split=5, clf__estimator__n_estimators=1
0, vect__max_df=0.75, score=0.23544572468162522, total= 1.0min
[CV] clf__estimator__min_samples_split=5, clf__estimator__n_estimators=10,
vect__max_df=1.0
[CV]  clf__estimator__min_samples_split=5, clf__estimator__n_estimators=1
0, vect__max_df=1.0, score=0.23177201758375018, total=  59.4s
[CV] clf__estimator__min_samples_split=5, clf__estimator__n_estimators=10,
vect__max_df=1.0
[CV]  clf__estimator__min_samples_split=5, clf__estimator__n_estimators=1
0, vect__max_df=1.0, score=0.2340812613705276, total=  59.7s
[CV] clf__estimator__min_samples_split=5, clf__estimator__n_estimators=10,
vect__max_df=1.0
[CV]  clf__estimator__min_samples_split=5, clf__estimator__n_estimators=1
0, vect__max_df=1.0, score=0.23832625833838691, total=  59.5s
[CV] clf__estimator__min_samples_split=5, clf__estimator__n_estimators=20,
vect__max_df=0.75
[CV]  clf__estimator__min_samples_split=5, clf__estimator__n_estimators=2
0, vect__max_df=0.75, score=0.23904805214491434, total= 1.4min
[CV] clf__estimator__min_samples_split=5, clf__estimator__n_estimators=20,
vect__max_df=0.75
[CV]  clf__estimator__min_samples_split=5, clf__estimator__n_estimators=2
0, vect__max_df=0.75, score=0.25227410551849605, total= 1.4min
[CV] clf__estimator__min_samples_split=5, clf__estimator__n_estimators=20,
vect__max_df=0.75
[CV]  clf__estimator__min_samples_split=5, clf__estimator__n_estimators=2
0, vect__max_df=0.75, score=0.25667070952092175, total= 1.4min
[CV] clf__estimator__min_samples_split=5, clf__estimator__n_estimators=20,
vect__max_df=1.0
[CV]  clf__estimator__min_samples_split=5, clf__estimator__n_estimators=2
```

```
0, vect__max_df=1.0, score=0.23707745945126574, total= 1.5min
[CV] clf__estimator__min_samples_split=5, clf__estimator__n_estimators=20,
vect__max_df=1.0
[CV]  clf__estimator__min_samples_split=5, clf__estimator__n_estimators=2
0, vect__max_df=1.0, score=0.25667070952092175, total= 1.4min
[CV] clf__estimator__min_samples_split=5, clf__estimator__n_estimators=20,
vect__max_df=1.0
[CV]  clf__estimator__min_samples_split=5, clf__estimator__n_estimators=2
0, vect__max_df=1.0, score=0.24787750151607035, total= 1.4min

[Parallel(n_jobs=1)]: Done  24 out of  24 | elapsed: 40.6min finished
```

Out[26]:

```
GridSearchCV(cv=None, error_score='raise',
       estimator=Pipeline(memory=None,
     steps=[('vect', CountVectorizer(analyzer='word', binary=False, decode
_error='strict',
        dtype=<class 'numpy.int64'>, encoding='utf-8', input='content',
        lowercase=True, max_df=1.0, max_features=None, min_df=1,
        ngram_range=(1, 1), preprocessor=None, stop_words=None,
        strip...oob_score=False, random_state=None, verbose=0,
            warm_start=False),
           n_jobs=1))]),
       fit_params=None, iid=True, n_jobs=1,
       param_grid={'vect__max_df': (0.75, 1.0), 'clf__estimator__n_estimat
ors': [10, 20], 'clf__estimator__min_samples_split': [2, 5]},
       pre_dispatch='2*n_jobs', refit=True, return_train_score='warn',
       scoring=None, verbose=7)
```

In [ ]:

```

```

7. Test your model

Show the accuracy, precision, and recall of the tuned model.

Since this project focuses on code quality, process, and pipelines, there is no minimum performance metric needed to pass. However, make sure to fine tune your models for accuracy, precision and recall to make your project stand out - especially for your portfolio!

In [27]:

```
y_pred_tuned = model.predict(X_test)
#converting to a dataframe
#y_pred_tuned = pd.DataFrame(y_pred_tuned, columns = y_test.columns)

category_names = list(df.columns[4:])
for i, col in enumerate(category_names):
    print('{} category metrics: '.format(col))
    print(classification_report(y_test.iloc[:,i], y_pred_tuned[:,i]))
```

```
related category metrics:
             precision    recall  f1-score   support

          0       0.71      0.44      0.55      1541
          1       0.84      0.94      0.89      4991
          2       0.38      0.51      0.43        65

avg / total       0.81      0.82      0.80      6597

request category metrics:
             precision    recall  f1-score   support

          0       0.90      0.98      0.94      5480
          1       0.81      0.46      0.59      1117

avg / total       0.88      0.89      0.88      6597

offer category metrics:
             precision    recall  f1-score   support

          0       0.99      1.00      1.00      6561
          1       1.00      0.06      0.11        36

avg / total       0.99      0.99      0.99      6597

aid_related category metrics:
             precision    recall  f1-score   support

          0       0.78      0.85      0.81      3857
          1       0.76      0.66      0.71      2740

avg / total       0.77      0.77      0.77      6597

medical_help category metrics:
             precision    recall  f1-score   support

          0       0.93      0.99      0.96      6083
          1       0.59      0.09      0.15       514

avg / total       0.90      0.92      0.90      6597

medical_products category metrics:
             precision    recall  f1-score   support

          0       0.96      1.00      0.98      6280
          1       0.69      0.09      0.15       317

avg / total       0.94      0.95      0.94      6597

search_and_rescue category metrics:
             precision    recall  f1-score   support

          0       0.97      1.00      0.99      6415
          1       0.55      0.07      0.12       182

avg / total       0.96      0.97      0.96      6597

security category metrics:
             precision    recall  f1-score   support

          0       0.98      1.00      0.99      6474
```

```
          1       0.00      0.00      0.00       123

avg / total       0.96      0.98      0.97      6597

military category metrics:
             precision    recall  f1-score   support

          0       0.97      1.00      0.98      6374
          1       0.73      0.09      0.15       223

avg / total       0.96      0.97      0.96      6597

child_alone category metrics:
             precision    recall  f1-score   support

          0       1.00      1.00      1.00      6597

avg / total       1.00      1.00      1.00      6597

water category metrics:
             precision    recall  f1-score   support

          0       0.96      1.00      0.98      6214
          1       0.88      0.32      0.47       383

avg / total       0.96      0.96      0.95      6597

food category metrics:
             precision    recall  f1-score   support

          0       0.94      0.99      0.96      5868
          1       0.81      0.53      0.64       729

avg / total       0.93      0.93      0.93      6597

shelter category metrics:
             precision    recall  f1-score   support

          0       0.94      0.99      0.97      6011
          1       0.83      0.41      0.55       586

avg / total       0.93      0.94      0.93      6597

clothing category metrics:
             precision    recall  f1-score   support

          0       0.99      1.00      0.99      6486
          1       0.88      0.13      0.22       111

avg / total       0.98      0.98      0.98      6597

money category metrics:
             precision    recall  f1-score   support

          0       0.98      1.00      0.99      6452
          1       0.73      0.06      0.10       145

avg / total       0.97      0.98      0.97      6597

missing_people category metrics:
             precision    recall  f1-score   support
```

```
          0       0.99      1.00      1.00      6533
          1       0.00      0.00      0.00        64

avg / total       0.98      0.99      0.99      6597

refugees category metrics:
             precision    recall  f1-score   support

          0       0.97      1.00      0.98      6377
          1       0.62      0.07      0.12       220

avg / total       0.96      0.97      0.95      6597

death category metrics:
             precision    recall  f1-score   support

          0       0.96      1.00      0.98      6289
          1       0.78      0.15      0.26       308

avg / total       0.95      0.96      0.94      6597

other_aid category metrics:
             precision    recall  f1-score   support

          0       0.88      0.99      0.93      5744
          1       0.56      0.04      0.08       853

avg / total       0.83      0.87      0.82      6597

infrastructure_related category metrics:
             precision    recall  f1-score   support

          0       0.94      1.00      0.97      6207
          1       0.08      0.00      0.00       390

avg / total       0.89      0.94      0.91      6597

transport category metrics:
             precision    recall  f1-score   support

          0       0.96      1.00      0.98      6292
          1       0.70      0.10      0.18       305

avg / total       0.95      0.96      0.94      6597

buildings category metrics:
             precision    recall  f1-score   support

          0       0.96      1.00      0.98      6266
          1       0.82      0.17      0.28       331

avg / total       0.95      0.96      0.94      6597

electricity category metrics:
             precision    recall  f1-score   support

          0       0.98      1.00      0.99      6478
          1       0.73      0.07      0.12       119

avg / total       0.98      0.98      0.98      6597
```

```
tools category metrics:
             precision    recall  f1-score   support

          0       0.99      1.00      1.00      6562
          1       0.00      0.00      0.00        35

avg / total       0.99      0.99      0.99      6597

hospitals category metrics:
             precision    recall  f1-score   support

          0       0.99      1.00      0.99      6530
          1       0.00      0.00      0.00        67

avg / total       0.98      0.99      0.98      6597

shops category metrics:
             precision    recall  f1-score   support

          0       1.00      1.00      1.00      6572
          1       0.00      0.00      0.00        25

avg / total       0.99      1.00      0.99      6597

aid_centers category metrics:
             precision    recall  f1-score   support

          0       0.99      1.00      0.99      6526
          1       0.00      0.00      0.00        71

avg / total       0.98      0.99      0.98      6597

other_infrastructure category metrics:
             precision    recall  f1-score   support

          0       0.96      1.00      0.98      6331
          1       0.00      0.00      0.00       266

avg / total       0.92      0.96      0.94      6597

weather_related category metrics:
             precision    recall  f1-score   support

          0       0.88      0.96      0.92      4755
          1       0.86      0.66      0.74      1842

avg / total       0.87      0.87      0.87      6597

floods category metrics:
             precision    recall  f1-score   support

          0       0.95      1.00      0.97      6079
          1       0.91      0.44      0.59       518

avg / total       0.95      0.95      0.94      6597

storm category metrics:
             precision    recall  f1-score   support

          0       0.94      0.99      0.96      5998
```

```
          1       0.74      0.38      0.50       599

avg / total       0.92      0.93      0.92      6597

fire category metrics:
             precision    recall  f1-score   support

          0       0.99      1.00      0.99      6526
          1       0.67      0.03      0.05        71

avg / total       0.99      0.99      0.98      6597

earthquake category metrics:
             precision    recall  f1-score   support

          0       0.98      0.99      0.98      5990
          1       0.90      0.76      0.82       607

avg / total       0.97      0.97      0.97      6597

cold category metrics:
             precision    recall  f1-score   support

          0       0.98      1.00      0.99      6458
          1       0.83      0.07      0.13       139

avg / total       0.98      0.98      0.97      6597

other_weather category metrics:
             precision    recall  f1-score   support

          0       0.95      1.00      0.97      6252
          1       0.50      0.03      0.05       345

avg / total       0.93      0.95      0.93      6597

direct_report category metrics:
             precision    recall  f1-score   support

          0       0.86      0.98      0.91      5302
          1       0.78      0.34      0.47      1295

avg / total       0.84      0.85      0.83      6597

/opt/conda/lib/python3.6/site-packages/sklearn/metrics/classification.py:1
135: UndefinedMetricWarning: Precision and F-score are ill-defined and bei
ng set to 0.0 in labels with no predicted samples.
```

8. Try improving your model further. Here are a few ideas:

- try other machine learning algorithms
- add other features besides the TF-IDF

In [35]:

```
#trying to add another feature.

upd_pipeline = Pipeline([
        ('features', FeatureUnion ([

            ('text_pipeline', Pipeline ([
                ('vect', CountVectorizer(tokenizer=tokenize)),
                ('tfidf', TfidfTransformer())
                                        ])),
             ('starting_verb', StartingVerbExtractor ())
                                    ])),

        ('clf', multioutput.MultiOutputClassifier (RandomForestClassifier ()))
        ])

# train SVM classifier
upd_pipeline.fit(X_train, y_train)
```

Out[35]:

```
Pipeline(memory=None,
     steps=[('features', FeatureUnion(n_jobs=1,
       transformer_list=[('text_pipeline', Pipeline(memory=None,
     steps=[('vect', CountVectorizer(analyzer='word', binary=False, decode
_error='strict',
        dtype=<class 'numpy.int64'>, encoding='utf-8', input='content',
        lowercase=True, max_d...oob_score=False, random_state=None, verbos
e=0,
            warm_start=False),
           n_jobs=1))])
```

In [36]:

```
y_pred_upd = upd_pipeline.predict(X_test)
#converting to a dataframe
#y_pred_tuned = pd.DataFrame(y_pred_tuned, columns = y_test.columns)

category_names = list(df.columns[4:])
for i, col in enumerate(category_names):
    print('{} category metrics: '.format(col))
    print(classification_report(y_test.iloc[:,i], y_pred_upd[:,i]))
```

```
related category metrics:
             precision    recall  f1-score   support

          0       0.65      0.46      0.54      1541
          1       0.85      0.92      0.88      4991
          2       0.37      0.45      0.40        65

avg / total       0.79      0.81      0.80      6597

request category metrics:
             precision    recall  f1-score   support

          0       0.90      0.97      0.93      5480
          1       0.78      0.45      0.57      1117

avg / total       0.88      0.89      0.87      6597

offer category metrics:
             precision    recall  f1-score   support

          0       0.99      1.00      1.00      6561
          1       1.00      0.06      0.11        36

avg / total       0.99      0.99      0.99      6597

aid_related category metrics:
             precision    recall  f1-score   support

          0       0.75      0.86      0.80      3857
          1       0.75      0.60      0.66      2740

avg / total       0.75      0.75      0.74      6597

medical_help category metrics:
             precision    recall  f1-score   support

          0       0.93      0.99      0.96      6083
          1       0.56      0.12      0.19       514

avg / total       0.90      0.92      0.90      6597

medical_products category metrics:
             precision    recall  f1-score   support

          0       0.96      1.00      0.98      6280
          1       0.67      0.10      0.18       317

avg / total       0.94      0.95      0.94      6597

search_and_rescue category metrics:
             precision    recall  f1-score   support

          0       0.97      1.00      0.99      6415
          1       0.50      0.03      0.06       182

avg / total       0.96      0.97      0.96      6597

security category metrics:
             precision    recall  f1-score   support

          0       0.98      1.00      0.99      6474
```

```
          1       0.00      0.00      0.00       123

avg / total       0.96      0.98      0.97      6597

military category metrics:
             precision    recall  f1-score   support

          0       0.97      1.00      0.98      6374
          1       0.58      0.07      0.12       223

avg / total       0.96      0.97      0.95      6597

child_alone category metrics:
             precision    recall  f1-score   support

          0       1.00      1.00      1.00      6597

avg / total       1.00      1.00      1.00      6597

water category metrics:
             precision    recall  f1-score   support

          0       0.96      1.00      0.98      6214
          1       0.82      0.27      0.40       383

avg / total       0.95      0.95      0.94      6597

food category metrics:
             precision    recall  f1-score   support

          0       0.94      0.99      0.96      5868
          1       0.85      0.47      0.60       729

avg / total       0.93      0.93      0.92      6597

shelter category metrics:
             precision    recall  f1-score   support

          0       0.94      0.99      0.96      6011
          1       0.80      0.34      0.47       586

avg / total       0.93      0.93      0.92      6597

clothing category metrics:
             precision    recall  f1-score   support

          0       0.99      1.00      0.99      6486
          1       0.90      0.17      0.29       111

avg / total       0.98      0.99      0.98      6597

money category metrics:
             precision    recall  f1-score   support

          0       0.98      1.00      0.99      6452
          1       0.70      0.05      0.09       145

avg / total       0.97      0.98      0.97      6597

missing_people category metrics:
             precision    recall  f1-score   support
```

```
          0       0.99      1.00      1.00      6533
          1       0.00      0.00      0.00        64

avg / total       0.98      0.99      0.99      6597

refugees category metrics:
             precision    recall  f1-score   support

          0       0.97      1.00      0.98      6377
          1       0.78      0.06      0.12       220

avg / total       0.96      0.97      0.95      6597

death category metrics:
             precision    recall  f1-score   support

          0       0.96      1.00      0.98      6289
          1       0.93      0.16      0.28       308

avg / total       0.96      0.96      0.95      6597

other_aid category metrics:
             precision    recall  f1-score   support

          0       0.88      0.99      0.93      5744
          1       0.56      0.08      0.13       853

avg / total       0.84      0.87      0.83      6597

infrastructure_related category metrics:
             precision    recall  f1-score   support

          0       0.94      1.00      0.97      6207
          1       0.07      0.00      0.00       390

avg / total       0.89      0.94      0.91      6597

transport category metrics:
             precision    recall  f1-score   support

          0       0.96      1.00      0.98      6292
          1       0.70      0.09      0.15       305

avg / total       0.95      0.96      0.94      6597

buildings category metrics:
             precision    recall  f1-score   support

          0       0.96      1.00      0.98      6266
          1       0.71      0.12      0.20       331

avg / total       0.94      0.95      0.94      6597

electricity category metrics:
             precision    recall  f1-score   support

          0       0.98      1.00      0.99      6478
          1       0.67      0.05      0.09       119

avg / total       0.98      0.98      0.97      6597
```

```
tools category metrics:
             precision    recall  f1-score   support

          0       0.99      1.00      1.00      6562
          1       0.00      0.00      0.00        35

avg / total       0.99      0.99      0.99      6597

hospitals category metrics:
             precision    recall  f1-score   support

          0       0.99      1.00      0.99      6530
          1       0.00      0.00      0.00        67

avg / total       0.98      0.99      0.98      6597

shops category metrics:
             precision    recall  f1-score   support

          0       1.00      1.00      1.00      6572
          1       0.00      0.00      0.00        25

avg / total       0.99      1.00      0.99      6597

aid_centers category metrics:
             precision    recall  f1-score   support

          0       0.99      1.00      0.99      6526
          1       0.00      0.00      0.00        71

avg / total       0.98      0.99      0.98      6597

other_infrastructure category metrics:
             precision    recall  f1-score   support

          0       0.96      1.00      0.98      6331
          1       0.17      0.00      0.01       266

avg / total       0.93      0.96      0.94      6597

weather_related category metrics:
             precision    recall  f1-score   support

          0       0.86      0.95      0.90      4755
          1       0.83      0.60      0.70      1842

avg / total       0.85      0.85      0.85      6597

floods category metrics:
             precision    recall  f1-score   support

          0       0.95      1.00      0.97      6079
          1       0.90      0.39      0.55       518

avg / total       0.95      0.95      0.94      6597

storm category metrics:
             precision    recall  f1-score   support

          0       0.94      0.99      0.96      5998
```

```
          1       0.73      0.36      0.48       599

avg / total       0.92      0.93      0.92      6597

fire category metrics:
             precision    recall  f1-score   support

          0       0.99      1.00      0.99      6526
          1       0.75      0.04      0.08        71

avg / total       0.99      0.99      0.98      6597

earthquake category metrics:
             precision    recall  f1-score   support

          0       0.97      0.99      0.98      5990
          1       0.89      0.69      0.77       607

avg / total       0.96      0.96      0.96      6597

cold category metrics:
             precision    recall  f1-score   support

          0       0.98      1.00      0.99      6458
          1       0.77      0.12      0.21       139

avg / total       0.98      0.98      0.97      6597

other_weather category metrics:
             precision    recall  f1-score   support

          0       0.95      1.00      0.97      6252
          1       0.44      0.06      0.10       345

avg / total       0.92      0.95      0.93      6597

direct_report category metrics:
             precision    recall  f1-score   support

          0       0.85      0.97      0.91      5302
          1       0.73      0.32      0.45      1295

avg / total       0.83      0.84      0.82      6597


/opt/conda/lib/python3.6/site-packages/sklearn/metrics/classification.py:1
135: UndefinedMetricWarning: Precision and F-score are ill-defined and bei
ng set to 0.0 in labels with no predicted samples.
```

9. Export your model as a pickle file

In [37]:

```
pickle.dump(model, open('final_model.sav', 'wb'))
```

10. Use this notebook to complete `train.py`

Use the template file attached in the Resources folder to write a script that runs the steps above to create a database and export a model based on a new dataset specified by the user.

In [ ]:

```

```