**READ ME File - Nano Degree Program-Data Scientist- Project 2-Disaster Response Pipeline**

Contents

[1. Installation 2](#_Toc23761404)

[2. Project Motivation 2](#_Toc23761405)

[3. File Descriptions 3](#_Toc23761406)

[4. Instructions: 3](#_Toc23761407)

[5. RESULTS 3](#_Toc23761408)

[6. Licensing, Authors, Acknowledgements 6](#_Toc23761409)

# Installation

The following additional packages needs to be installed. The code uses Python 3.7

* punkt
* wordnet
* stopwords

Example:

* #nltk.download('wordnet')
* #nltk.download('punkt')
* #nltk.download('stopwords')

# Project Motivation

In this project we have analyzed thousands of real messages provided by Figure 8, sent during natural disasters either via social media or directly to disaster response organizations. I have built an ETL pipeline that processes message and category data from csv files and load them into a SQLite database. Then the machine learning pipeline reads the data from SQLite database, create and save a multi-output supervised learning model. Finally the web app will extract data from this database to provide data visualizations and use the model to classify new messages for 36 categories.

Machine learning is critical to helping different organizations understand which messages are relevant to them and which messages to prioritize. In this project, I have used the skills related to ETL pipelines, natural language processing, and machine learning pipelines to create the Model to classify the messages that would have real world impact.

# File Descriptions

The project work is housed across three folders. All the data /scripts/models can be accessed in GitHub using following URL;

GitHub URL;

1. data
   * **disaster\_categories.csv**: categories dataset
   * **disaster\_messages.csv**: messages dataset
   * **process\_data.py**: ETL pipeline scripts to read, clean, and save data into a database
   * **DisasterResponse.db**: ETL pipeline output , i.e. SQLite DB containing cleaned combined data of messages and categories
2. models
   * **train\_classifier.py**: Machine learning pipeline script to create, improve and train a model and then export the model/classifier as pickle file
   * **classifier.pkl**: output of the machine learning pipeline, i.e. a trained classifier
3. app
   * **run.py**: Flask file to run the web application
   * **templates**: This contains html files needed to render the results for Web Application

# Instructions:

1. Run the following commands in the project's root directory to set up your database and model.
   * To run ETL pipeline that cleans data and stores in database: **python data/process\_data.py data/disaster\_messages.csv data/disaster\_categories.csv data/DisasterResponse.db**
   * To run ML pipeline that trains classifier and saves python :**models/train\_classifier.py data/DisasterResponse.db models/classifier.pkl**
2. Run the following command in the app's directory to run your web app. python run.py
3. Go to [http://*localhost*:3001/](http://localhost:3001/)

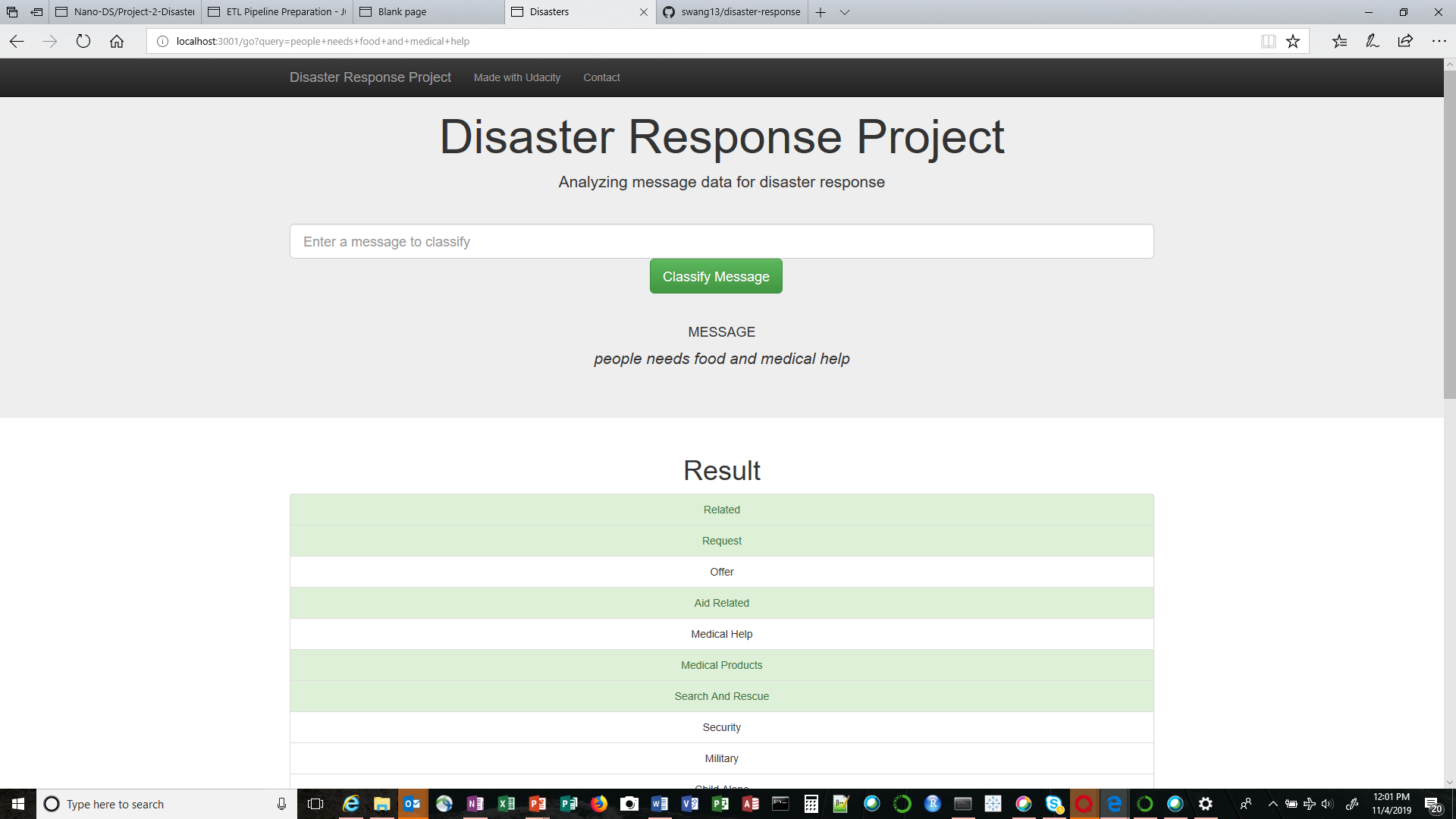
# RESULTS

I have selected a Supervised ML “Random Forest Classification” model to classify the disaster messages into one or more of 36 categories.

**ETL Pipeline output-** Creates the SQLite *DisasterReponse.DB*

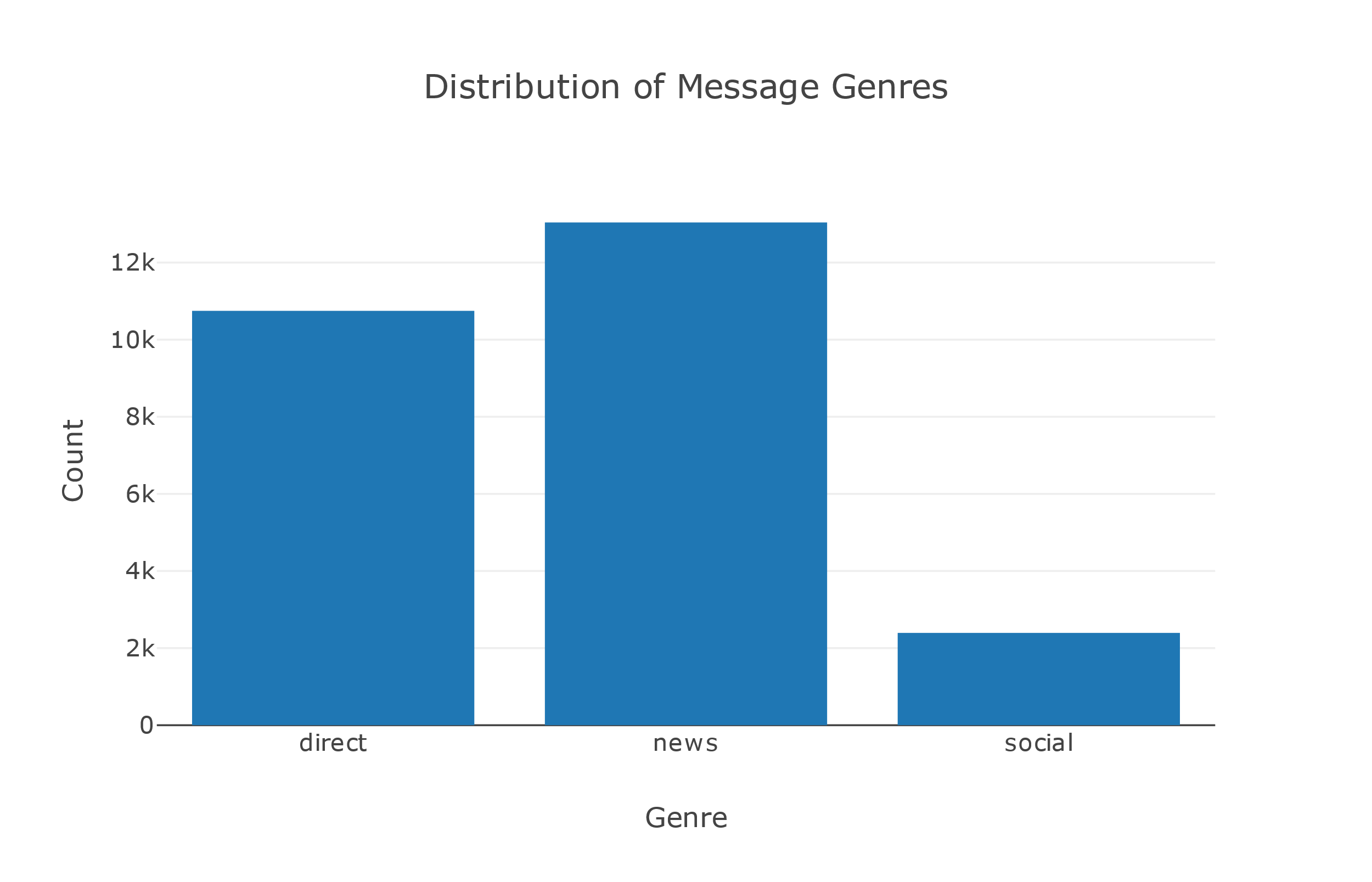
**Machine learning Pipeline –** Creates the *classifier.pkl*, which will be used by Web app for message classification

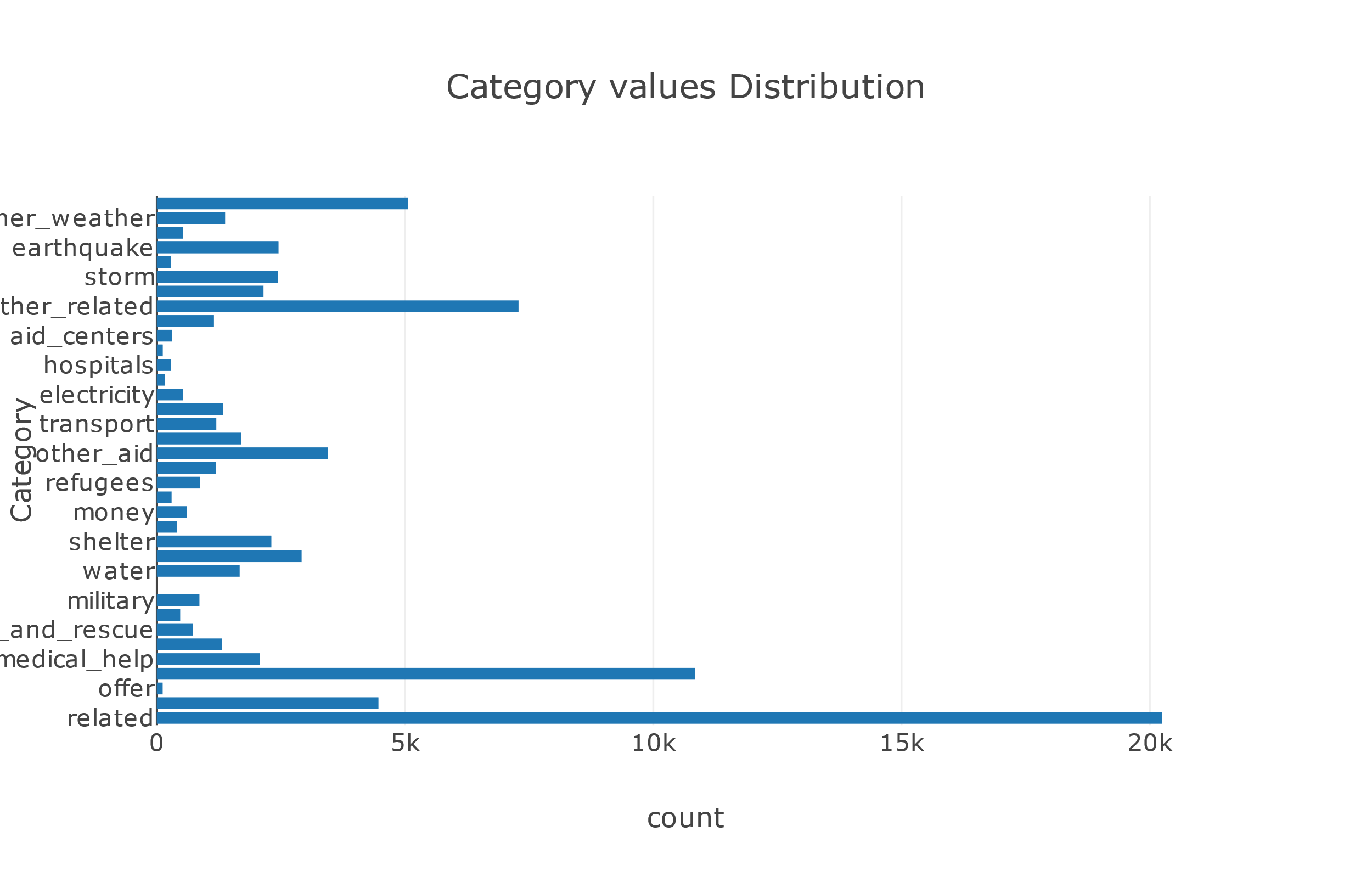
**Flask app-** This loads the data and model and start the web app @ <http://0.0.0.0:3001/> to accepts user query/message for classification. Below is the sample of message classification

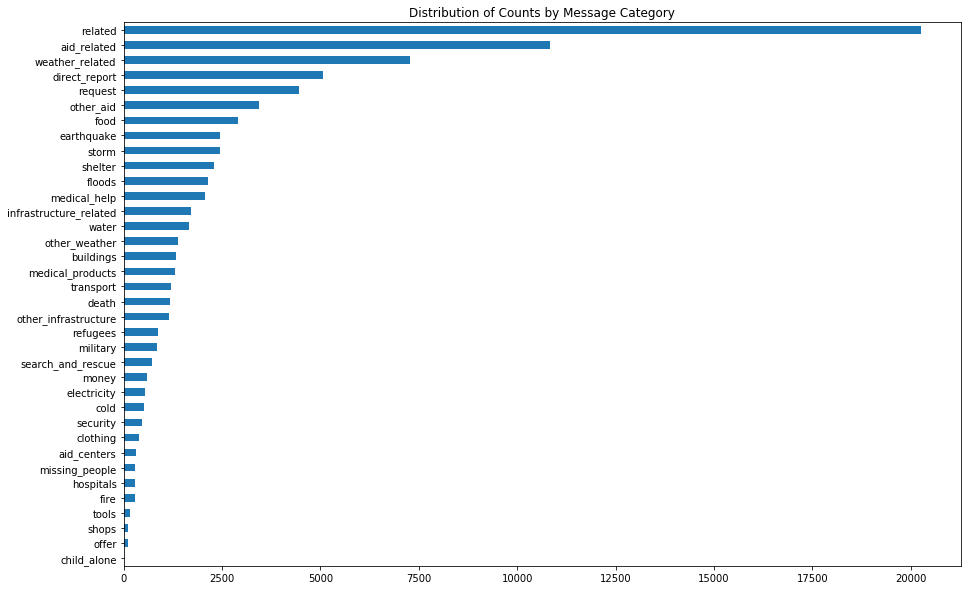


**Data Visualization**

* Distribution of Message Genre
* Message category counts







# Licensing, Authors, Acknowledgements

I would like to give credit to Udacity online courses and to Figure-Eight.com for providing the data.

Few additional links/resources are stated below that I have used as reference to complete my project.

1. Choosing the right metric [Choosing the Right Metric for Evaluating Machine Learning Models](https://www.kdnuggets.com/2018/04/right-metric-evaluating-machine-learning-models-1.html).
2. Article about [What metrics should be used for evaluating a model on an imbalanced data set](https://towardsdatascience.com/what-metrics-should-we-use-on-imbalanced-data-set-precision-recall-roc-e2e79252aeba)
3. Article on [various techniques of the data exploration process](https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/).
4. Data Mining for Business Analytics- *Galit Shmueli, pertr C Bruce ,Wiley Publications*