Name: Minh Ngoc Pham

Course: Data Science Career Track

Capstone Project # 2 Project Ideas

1. Toxic Comment Classification:

Dataset: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data

This dataset contains Wikipedia comments that have been rated for toxic behaviour. There are 6 classes for 6 types of toxic behaviour including: toxic, severe_toxic, obscene, threat, insult, identity_hate. There are 395,667 comments in the training set (which also includes comments that are rated as neutral, meaning being scored 0 for all classes). The test set include 381,580 comments.

This is a multi class text classification problem and we will make use of NLP (Natural Language Processing for this problem).

2. Belief about climate change

Dataset: https://data.world/crowdflower/sentiment-of-climate-change

The dataset contains tweets regarding belief in the existence of climate change. The data helps evaluate the general belief of the public regarding climate change. The classification for the tweets is “Yes” if there is suggestion that there is a belief that climate change is occurring. In addition, there is a confidence level (ranging from 0 to 1) regarding the confidence in the score for each tweets.

3. Predicting whether a Tweet is about a disaster or not

Dataset: https://www.kaggle.com/c/nlp-getting-started/data

This dataset contains tweets with 3,264 in the test set and 7,614 in the training set. In both sets, we have id regarding the unique id for the tweet, the text of the tweet, the location of the tweet (can be a black value), keyword of the text and in the training set, we have the target to determine whether the tweet is about a real disaster or not (taking the value of 0 or 1).