Project: Sentiment Analysis of Intra-Platform Ties on Uberpeople.net
Members: Amaris Huang, Ryan Huang, Ziling Cheng

#1 Problem Statement

In this project, we would like to adopt the original research of Dr. Xiao Chen, and Dr. Wei Chi, on exploring the phenomenon of informal intra-platform ties, among on-demand ride-hailing drivers in North America. We base our model on data scraped from www.uberpeople.net, a widely used online platform for uber users-- drivers and customers. We would like to apply machine learning models on analyzing the wordings of individual posts to detect whether a post gets emotional.

#2 Data Preprocessing

We use data.scraped by Dr. Xiao Chen and Dr. Wei Chi, who were kind enough to share the data with us. There are 14986 sample posts in our dataset and each contains the url where it is scraped from, the title of the threads, the contents of the post, the name of the author. We choose the column of the contents of the post out of the whole dataset and label each positive post to be 1, negative post to be -1 and a post completely informational to be 0 using SentimentIntensityAnalyzer from nltk.sentiment. And now we are ready to clean the data by lower-casing all words and only keeping letters and spaces, then turn each post into a vector. By doing so, each data point in the training set is in its binary bag of word representation rather than its string representation.

#3 Machine Learning Model
We allocate 70%,15%,15% of data for training, validation and testing respectively so that the model sees enough examples to be more likely to find a better solution, helps us make a better decision when choosing the final model to be used in the test set and gives us a fair sense of how well our model generalizes to unseen posts.

We decide to train data using three traditional machine learning models, namely, Naive Bayes, Support Vector Machines and Random Forests and compare their performance on the validation set. We haven't finished all the coding part at this point, so we cannot choose the final model for now unfortunately, but we will later use classification accuracy, which is the proportion of the correct predictions, to evaluate performance of our models helping us to decide on which model to be used for testing and the have a feel of the performance on the unseen data.

The packages that we use are:

```
import pandas as pd
from matplotlib import pyplot as plt
from pprint import pprint
import nltk
from nltk.sentiment import SentimentIntensityAnalyzer
```

```
import csv
import random
from datetime import datetime
import string
import re
from collections import Counter
from sklearn.metrics import accuracy_score
from sklearn.svm import LinearSVC
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import AdaBoostClassifier
from sklearn.model_selection import GridSearchCV
from sklearn.naive_bayes import GaussianNB
```

#4 Preliminary Results
See our GitHub page: https://github.com/butterfly002/uberpeople

#5 Next Steps
After we choose the final model, we will tune the hyper-parameter as we plan to use default values when choosing from those three traditional machine learning models.

Although traditional machine learning models are easier to implement, they are usually not the models with the greatest performance so we consider implementing RNNs or more advanced language models.

Our classification task is simply binary for now, to make the classification task a bit more complex, we are also considering scaling the emotion range when a post is classified with emotions.