Project Title: News Category Classification
Team 1 Names: Benjamin Barnett and Yaniv Bronshtein

Dataset: https://www.kaggle.com/rmisra/news-category-dataset

This dataset contains around 200k news headlines (with 41 different categories) from the years 2012 to 2018 obtained from HuffPost. We would like to extract the first 1000 records for each category and split them into training (80%) and testing (20%) data. For modeling purposes, we intend to try two approaches: 1) Use the headline and brief description from the data and 2) Use the beautiful soup library to parse the text for each HuffPost news article.

For each approach, we plan to compare our models to the Naive Bayes classifier, and plot the results on confusion matrix heatmaps. We would like to use word embeddings, but may decide to also use TF-IDF for the headline and brief description approach. (Since TF-IDF would require more time to run than word embeddings, we do not plan to use it for the text parsing approach.) By comparing the performance of different models, we can determine which one is the best, with the Naive Bayes classifier as a benchmark.

The models we would like to use include single and multilayer RNNs with 100 units, along with LSTMs and potentially BERT. While the classification performance will likely be similar for the first approach, we predict that the LSTMs will perform better than the traditional RNNs under the second approach. We will import the keras package and more to run these models on a Google Colaboratory notebook, so that we can work on the same file.

So far, we downloaded the data as a JSON document, converted it to a .csv file, and have it uploaded to our shared Google Colaboratory notebook. We also grouped the data by category, taking only the first 1000 records each for a total of 41,000 news articles. Splitting the data into training and testing data, our next step is to train our models (under the two approaches) and to compare their performances to the Naive Bayes model. One thing we may need to keep in mind is to remove stop words and/or punctuation before proceeding with TF-IDF or word embedding. For the embeddings in particular, there may be words that are not in the embedding vocabulary such as proper nouns. By assigning random embeddings to out-of-vocabulary words when we see them for the first time, we can obtain better results than just simply ignoring them.

With our evaluation of different models, we hope that we will be able to find a model with high classification performance (using micro averaged and macro averaged F1 scores), particularly one that does better than the Naive Bayes model. The task of labeling news article categories is important, as it can be used to identify tags for untracked news articles. Another use would be to review language differences by category. The way crime articles are written, for instance, might be different from the way political articles are written. There are many applications for category classification, and we desire to improve such modeling for news articles.