Commentary on the Task

1. Data Analysis
    a. I first plotted label frequency distributions, which showed that almost all labels except one had the same number of instances, and even the label in minority had sufficient number of instances in the dataset. Therefore, I didn't consider the dataset to have label imbalance issue, which if it was true might have required other measures, such over-sampling the minority label, or collecting more instances for it, etc. File: label_count_bar_chart.png

    b. Then, I plotted a word cloud of names belonging to each of the labels. It gave an idea on what kind of words are contained in which labels, and if there are certain labels that can be more easily identified due to the presence of some indicator words without requiring a sophisticated (pre-trained) language model. I observed that for *educational institution*, there were dominant words like *university*, *college*, *school, institute, academy etc.* If I had more time then I would have tried a simple classifier for such labels as they seem to be a lot easier to classify without requiring external knowledge, in comparison to a label, say "*artist"*, which as a matter of fact cannot be classified without prior knowledge of the artist's name via pre-training. These word clouds also showed that different words were often coming from different languages. Directory: word_cloud_plots.png

    c. Afterwards, I wanted to find out how many of those labels belong to which language, so I plot the frequency of each language in a bar chart. I observe that even though an overwhelming number of names are in English, but there are plenty of other languages present in the dataset ,which could combined account for significant number of instances. Therefore, if I had more time I would have first tried classification with a multilingual BERT, and also perhaps one BERT for every language, such that we would first identify the language of the name, and then use the specific instance of BERT pre-trained on text of that language to classify that name. File: label_count_bar_chart.png

    d. Finally, I observed that there was no negative/zero label in the dataset for the case when an instance doesn't belong to any of the 14 labels, but is in fact out of domain. To ensure the classifier can identify such names, I collected a list of random nouns from the internet and assigned them label 0, increasing the total number of labels in the dataset to 15 from the original 14. File: nounlist.txt

2. BERT based classification
    a. I implemented a code that does the following things:
        i. Divides the data into train/validation/test
        ii. Uses a pre-trained BERT model for sequence
        iii. Plots train/validation loss curve

iv. Prints out the accuracy on the test set. File: loss_plots.png

b. As I mentioned above, If I had more time I would have experimented with multilingual BERT, or multiple BERTs pre-trained in the language of the instances.
c. Unfortunately, **I didn't have easy access to a GPU**, therefore, I was only able to experiment with the code on an extremely small (toy) subsample of the original dataset.

3. Classical machine learning
    a. I implemented a code to perform 5-fold cross validation using some classical machine learning models, namely naïve_bayes, linear_classifier, ridge_classifier. These were the few classifiers that could train in reasonable time over high dimensional bag-of-words representations of the documents (names).
    b. Code computes the accuracies of these models over 5-fold cross-validation
    c. Code also plots a bar-chart depicting which models perform better in comparison. File: model_comparison_bar_chart.png
    d. I'm fully aware that some of the labels would require pre-trained language models that encode external information, such as artist, album or film, since there is no way to train a classifier to identify a new artist, film or album name without having seen the names before in the correct context. But I was still curious to find out what kind of accuracies we could obtain with these classical models. File: model_comparison_bar_chart.png