CODING SKILL TEST

Natural Language Processing and Machine Learning

Task:

You have been provided both Python code and a dataset. The Python code trains a model and outputs the performance metrics to the variable called *model_metrics*. Your task is to improve the performance of the model (current code yields ~40% precision, recall and fscore) through pre-processing steps and any code refactoring you feel is necessary.

Data:

The data consists of 237 corpuses belonging to any one of three topics: fly fishing, ice hockey and machine learning. The corpus breakout for each topic is below:

topic	count
fly_fishing	86
ice_hockey	70
machine_learning	81

The code (my_test.py and utils.py) and the data (data.pkl) resides on a github repository called coding_test: https://github.com/Gamelands/coding_test. The data has the following columns:

column	description
body_basic	crawled content
label	topic

Please go ahead and clone the directory, load the pickle object into python and refactor the code required to solve for the ask.

The code works by simply changing a path that points to the data.pkl object.

Please commit your .py files to a public github repository you create and send the github repository link to your HR contact.

Note:

Data Science is both an art and a science, there is no one specific solution, numerous solutions will suffice for the above ask, so be as creative as you want and most of all have fun with the exercise!