Modules:
preprocessing.py contain functions that used in corpus preprocessing and query preprocessing, including tokenization, stemming and stop words handling. It also contains functions to store the invert index into persistent shelve file for further retrieval.
nbrun.py: give benchmark for Naïve Bayes method.
knnrun.py give benchmark for KNN and use KNN to predict the origin of covid-19
xgbrun.py: run the same work as preprocessing.py plus give benchmarks for XGBoost and use XGBoost to predict the origin of COVID-19

Build Instructions:
If the pickle folder is empty, run preprocessing.py first to pre proceed RNA data in fasta format and dump them in the folder pickle. However, this would take at least 30 mins to complete. The dumped data files are already handed in in the pickle folder. Hence, **you do not need to run preprocessing.py for running machine learning model.**

Run Instructions:
Run knn.py or nbrun.py directly to see benchmark and/or prediction result.
Open xgbrun.ipynb in Jupyter Notebook to run it.