This is the readme by Piervincenzo Ventrella for the project of the course CS412 Introduction to Machine Learning.

For the project I choose Task #3 which goal is to develop a model able to predict where in a question-anwsering plattoform some questions posted by the user should be removed or not from the platform by using the Quora Insincere Question Dataset.

The project follows this structure:

-**Dataset Inspection and preprocessing**
The .py files for this step are the ones named **PreprocessingX.py** in which I explore the dataset and produce the BagOfWords models to finally save them as **BagOfWordDataSetX.csv.** You are free to run the files to check how the code is working but they will require a lot of time to terminate because of the performed computations. I suggest running them for the first part to visualize the extracted vocabulary but then quit when it starts to create the new preprocessed Datasets.

-**First Baseline and Model's Comparison.**
The related files are the ones named **ModelsComparisonX.py.**
You are free to take a look at them.
Note: these files load datasets produced from the corresponding **PreprocessinX.py** file.
I provide these Preprocessed Datasets, so it will be possible to run them without running the preprocessing.

-Final Comparison and Development.
In **Implementation.py** I compare once again the 2 most promising models and finally I implement the final model and save it as **FinalModel.joblib.**

-**Testing**
In this step (**Testing.py** file), for comparison purpose I load **FinalModel.joblib** and **NaiveBayesBaseline.joblib**.
The 2 models are trained on a different preprocessed Dataset, so I load the 2 corresponding test sets (**testSet.csv** and **testSet0.csv**, both already preprocessed respectively in **Implementtion.py** and **ModelComparison0.py**).

Important: in loading the datasets, the programs use a relative path starting from their current directory so make sure to keep everything in the same folder or alternatively change the path directly from the code.