ERG2050: Introduction to Data Analytics

Assignment 3

March 17, 2021

Due: 11:59 pm, March 24, 2021

Note: Please use **Python 3.7** possibly with NumPy, other external packages are not allowed.

Exercise 1 (100 pts)

Description:

- There are 20 newsgroups with 18,846 posts and we want to explore the relationship between newsgroups posts and their topics. Therefore, you are required to construct a Naïve Bayes model to classify the posts.
- Totally, there are 11314 posts for training and 7532 posts for testing.
- We already convert texts to vectors using TF-IDF. For each post, it is represented as a vector with shape (130107,1).

Grading Standard:

- Xs and Ys are given (Xs is given in **{train/test}_feats.npy** and Ys can be otained from **{train/test}_labels.npy**)
  - Naïve Bayes model
  - Training process
  - Evaluation process
- We will use the ACCURACY over all given Xs in test set to evaluate your results.
- FILL IN the blank in NaiveBayes Class.

Describe how do you design it.