1. The file `4newgroups.csv` contains the text of 3874 messages originating from four newgroups,

   `rec.autos`, `rec.motorcycles`, `rec.sport.hockey`, `rec.sport.baseball`,

   labelled by the newsgroup to which it belongs. Our goal is to identify the newsgroup in which a message was posted, given the message text. You'll need to R or Python for this. Use whatever packages you want.

   We generated a list of words occuring in these messages and removed commonly occuring *stop words* (e.g., the, a, an, in, of), yielding a vocabulary of size 3392:

   aaa, ab, abc, abilities, ability, . . . , zero, zhitnik, zombo, zone, zx

   The file `4newsgroups-binary.csv` contains a 0/1 matrix indicating whether word j occurs in a message i. The file `4newsgroups-multinomial.csv` contains a matrix counting the number of occurences of word j of message i. Messages are listed in the same order as in `4newgroups.csv`.

   (a) For each of the two data files, construct logistic regression classifiers with your choice of softmax loss or binary cross-entropy together combined with a one-versus-rest stragegy like in HW2. Approximate the predictive accuracy of your classifiers using 5-fold cross validation. Should you normalize your features?

   (b) Construct a Naïve bayes classifiers for each of the two data files. Approximate their predictive accuracies using 5-fold cross validation. Should you normalize your features?

   (c) Construct a random forest classifiers for each of the two data files. Explain how you chose the number of trees in your forests. Use out-of-box data to estimate prediction accuracy. Which ten words have the highest feature importances?

   (d) Compare your results from the two data files the various classifiers you constructed. Report on any interesting phenomena you observe.

   (e) [**Bonus**] Can you improve on any of the above results using more sophisticated techniques, either at the preprocessing or the classification stage?

   Remark: The dataset used in this problem is a subset of the `20newsgroups` dataset avaiable at `https://archive.ics.uci.edu/ml/datasets/Twenty+Newsgroups`.