Homework 2

We are going to perform text message spam detection using the simple Bag of Words and Naïve Bayes models we talked about in class.

You will be setting up a simple experimental design that uses k-fold cross validation and gathers accuracy, precision, recall, and F-score.

* Create a Bag of Words representation of the text messages.
* Encode the categories into numerical values.
* Instead of a single train-test split, use 5-fold cross-validation (from sklearn.model\_selection import KFold)
  + Use a Naïve Bayes classifier to train the data
  + Evaluate the results using Accuracy, Precision, Recall, and F-score (all exist in sklearn!)

Answer the following questions:

1. Which of the four metrics has the lowest average score across the 5 folds?
   1. What does this mean in terms of spam detection?
   2. What could be a possible cause of this? (HINT: look at how many instances there are of each class.)
2. Which of the four metrics has the highest average score across the 5 folds?
   1. How do we interpret these results?
3. Which of the two discussed metrics (answers to Q1 & Q2), do you think is more important for spam detection and why?