Homework 3 - Supervised Learning

1. In this question, you will develop supervised learning methods to classify the party of a speaker. The text of the congressional record for 2006 is contained in the file "CongressionalRecord.zip" on canvas. Separate the text into a series of snippets by speaker (i.e. a given document could be a snippet of text that Senator Snowe said during a debate. After parsing the congressional record, you should have two vectors of length about 20,000--one containing the text (i.e. "Mr. President, I further wish to honor the...") and one containing the speaker (i.e. "DAYTON").

You can use the following code to find the party of the speaker of each snippet of text:

```
# read the dwnominate file for getting the name and party of the senators
dwnom <- read.xlsx("dwnom.xlsx")
nameparty <-
subset(cbind(dwnom$Name,dwnom$Party),dwnom$CONG==109&dwnom$Cong..Dist.==0)
name <- nameparty[,1]
party <- nameparty[,2]

# code the party of the speaker
speakerparty <- rep(NA,n)
for(i in 1:n)
{
      partynum <- party[which(name == speakers2[i])]
      if(length(partynum) == 1) {
                    if(partynum == "100") speakerparty[i] <- "D"
                else if(partynum == "200") speakerparty[i] <- "R"
      }
}
```

Drop all snippets of text that do not correspond to a Democratic or Republican speaker. Create a set of features using both 1-grams and 2-grams (you can choose how to pre-process the text, dropping stopwords, dropping very common or very uncommon terms, etc.). Divide the set of texts randomly into a training and test set.

Apply the following supervised learning methods to the training set and measure performance on the test set: naive Bayes, logistic regression with the LASSO, and support vector machines with the Gaussian kernel. Report the performance (using both 1-grams and 2-grams as features) for each of these methods. If the method allows you to interpret the model, interpret the model.