Model AI Assignment
Party Affiliation Classification from State of the Union Addresses

Project Time: 2 weeks

Instructions:
You will be using data from the Presidential State of the Union (SOTU) Addresses available as a zip archive. The speeches are available in text files labeled by their year, e.g., s1941.txt, s1973.txt. The text files are formatted such that there is one word per line and most punctuation has been removed (note, there are still hyphens or dashes left in the text files).

Review the lecture slides introducing the Naïve Bayes classifier, specifically take note of the slides discussing the multinomial and Bernoulli models for text classification.

Submission Instructions:
Most high-level programming languages can be used for this assignment. The use of Matlab, Python, or R will be fully supported.

Submit the code used to generate the answers to each question and the written responses to questions when asking for a description or report of values.

1. (8 points) Read in the SOTU addresses. You will need to create a vector listing the party affiliation of each president to match their speech.[1]

2. (8 points) Remove *stopwords* from consideration for the method. The stopwords are available at `stopwords.txt`.

3. Predict the party affiliations (Democrat / Republican) for the following speeches:

   - Barack Obama, 2014
   - George W. Bush, 2006
   - William Clinton, 1995
   - John F. Kennedy, 1962

   The training set will be the remaining speeches that can be associated with the Democratic or Republican presidents. You will need to complete the following steps:

   (a) (8 points) Describe how the probabilities $P(x_i|c_j)$ can be estimated from the training data using the Bernoulli model.

   (b) (24 points) For the 4 speeches listed above determine the party affiliations of the president. Calculate and report $P(C = Dems|\mathbf{X})$ and $P(C = Rep|\mathbf{X})$ under the Bernoulli model of Naïve Bayes. In order to avoid underflow errors, use the log probabilities and Laplace smoothing.

   (c) (8 points) Describe how the probabilities $P(x_i|c_j)$ can be estimated from the training data using the multinomial model.

   (d) (24 points) For the 4 speeches listed above determine the party affiliations of the president. Calculate and report $P(C = Dems|\mathbf{X})$ and $P(C = Rep|\mathbf{X})$ under the Multinomial model of Naïve Bayes. In order to avoid underflow errors, use the log probabilities and Laplace smoothing.

[1] Information on party affiliation is available at: `http://en.wikipedia.org/wiki/List_of_Presidents_of_the_United_States`

Bonus Problems, Extensions, and Alternative Problem Formulations

1. *Pre-processing:* The SOTU addresses can be provided to the students already pre-processed, ready to be analyzed for the classification problem. Or, student's may also be required to prepare the addresses from the source material.
2. *Alternative Analysis:* Throughout US history, party affiliations have changes with respect to their political view. Repeat the prediction of the party affiliation for the four speeches listed using only the other speeches since 1913 as the training data. Do your predictions change?
3. *Experimental Design and Analysis:* The size of the training data set is small. Rather than predict the party for the four selected presidential speeches, perform leave-one-out cross-validation (LOOCV) over the entire data set. Report the predicted party for each SOTU address.
4. *Feature Selection:* Have the students read the McCallum and Nigam AAAI, 1998 paper, and follow their description to perform feature selection on the SOTU data set. For each word in the vocabulary of the SOTU data set, calculate the mutual information with the class variable (party affiliation). Rank the words in the vocabulary by this metric and use this to form subsets of the data with the 100, 500, 1000, ... top-ranked words. Determine the classification accuracy on these data sets and report the results as learning curves.
5. *Alternative Data Sets:* The problem of text classification can be used for many other sets of texts that may be more topically relevant or fit a class interest. Examples include other political speeches (e.g., inaugural addresses during presidential election years), poems, novels, and plays. Another example would be determine whether a play or scene by Shakespeare is a comedy/tradegy/history.