*Document Classification* Final Project Report

The main objective of our project was to construct a method of classifying *New York Times* articles according to subject, by means of scraping data, using frequencies of words, and applying logistic regression techniques. We were given a set of data that consisted of archived news stories that were pre-tagged as either “Art” or “Music.” The articles were in XML format that also contained the document’s metadata. Specifically the problems that we had to resolve was to first describe the parameters of a function that mapped from article to label, train (optimize) the parameters for this function on training data, and assess the function on test data. Our implementation involved five key steps: the actual scraping of documents, dictionary vector construction, principle component analysis (PCA), logistic regression, and lastly document classification.

In order to begin working with the actual data, the words from the *New York Times* articles first had to be extracted from its XML format. We did this by going through each of the provided articles and looking for strings that matched the regular expression [[alpha]+[num]\*/’]. We figured that individual words would still be the most significant for the purpose of out project, but in our first step we aimed to make our dictionary as large as possible in order to take into account all of the data from each article. We did not feel it was necessary to distinguish variations of words, for instance, conjugations or pluralisms, because the additional functionality would not significantly improve performance. For our algorithm, we need to construct a single vector to describe each document. Inside each of the vectors the index of the elements corresponded with unique words, with the elements themselves containing the word occurrences. The first crude dictionary was very large (length of 8196 words) and our next objective was to shrink it to a more manageable dimensionality, so we only took the 150 most frequent words from either category. Our next step was to take out the words that appeared in both articles from the “Art” section and in articles from the “Music” section thus we just focused on the word that were unique to each category. After we removed the common words we were left with 60 words in our dictionary. In order for us to reduce the dictionary’s dimension even furthermore we conducted a principle component analysis (PCA) on the resulting data.

Our goal with the PCA was to examine the importance of each word, or simply, a “variable”. We began by taking out variables until we were left with the 30 variables that predicted the category most strongly (30 was an appropriate yet arbitrary number as it was less than the number of final articles that we were going to examine but enough to give us information as to which words were most relevant for each category). First we row-binded the word counts for both art and music articles so that the words match up. The prcomp command outputted us a matrix, showing the coefficients for each variables and each principal component. We then used the coefficients for principal component one as it accounted for the most variability in the data amongst the principal components. We looked at the absolute magnitudes of each coefficient, and sorted them from highest to lowest, in order to take the 30 highest ones. We then built a 30 variable data frame and added an extra column for each category which was then used for running our logistic regression. We also wrote code for drawing the top five words from the PCA, when we were tweaking our logistic regression, but in the end we stuck with our 30-explanatory variable model due to better predictive ability.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | |  | Coefficient | | (Intercept) | 16.9 | | Well | -3.2 | | Jazz | 0.7 | | Museums | -7.4 | | Paintings | 2.9 | | Center | -0.3 | | March | -1.2 | | Much | -12.2 | | Images | -0.2 | | Opera | 0.9 | | Studio | 1.2 | | Says | 8.0 | | Morris | 11.2 | | West | 5.2 | | No | -7.4 | | Performance | -0.3 | | Gifts | -14.6 | | Music | -8.3 | | Last | 2.0 | | Company | -18.4 | | Her | 4.7 | | Get | -15.9 | | Major | -1.4 | | Strike | 0.9 | | Stage | 8.5 | | Collection | -4.1 | | Life | 8.1 | | Building | -3.9 | | Barbera | -2.9 | | Me | 2.2 | | Artist | 1.8 |   *< Logistic Regression Summary>* |

Before training our logistic regression model, we realized the need of separating the articles into training and testing articles.  For the purpose of training, 30 articles from each folder were used in the logistic regression model and the remaining art’s 27 and music’s 15 articles were separated from the original file to be used as the testing articles to see how accurate our regression model is. Boolean values of “1” and “0” were given to art and music articles respectively for running the regression model. By using the general linear regression function, glm(…,family=”binomial”), we calculated each parameters’ estimated coefficients.

The reported table’s coefficient values were rounded for convenience. Running the logistic regression model, the numbers of variables that we used were 30 while the sample size was 60 (30 each).  The R output initially displayed odd features and we suspect this was due to the high sample vs. variable size ratio (n:p = 2:1).  When the parameter numbers were reduced to five, the regression output looked much better. The five words were “well”, “jazz”, “museums”, “paintings”,  and “center”, however, we weren’t able to get an accurate predicting ability due to such few parameters.  In order to assess the reasonableness of our predictions we tried a few other techniques, one of which was a step-wise regression method though ultimately not the most appropriate for our purposes. The final words that were reduced by the step-wise function were “museums”, “much”, “no”, “gifts”, and “get.”  Except for “museum”, other words were difficult to be associated with music or art.  The accuracy of the step-wise model was worse than the 5 variable regression model’s.  Knowing the fact that the regression model is over fitted, we concluded to use the model with 30 variables in categorizing the testing articles.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *<Classification Success Rate-*  *Logistic Regression Model>*   |  |  |  |  | | --- | --- | --- | --- | |  | Art | Music | Average | | Ratio | 70.37% | 33.33% | 57.14% |   -----------------------------------------------------  *<Classification Success Rate –*  *Naïve Base Classifier>*   |  |  |  |  | | --- | --- | --- | --- | |  | Art | Music | Average | | Ratio | 70.305% | 28.03% | 60.86% | |

Using the logistic transformation probabilty equation, the Pr( Y=1 | X=x), same as

Pr( Y=Art | X=vocabs), was calculated for each of the testing articles.  The threshold was set as 0.5 and any probabilty over than 0.5 was identified as an art related litarature, while lower than 0.5 was classified as a music article.

Testing were conducted on 27 art and 15 music articles.  The result showed that the logistic regression model that we have is better in classifying the art articles than those about music.  The classification success ratio was 70.37% and 33.33% respectively with a total accuracy of 57.14%.  We were worried about the low success classification rate for the music articles.  To verify that our regression model was working in a reasonable way, we also tried to use a Naïve Bayes Classifier (a probabilistic classifier based on Bayes’ theorem with strong assumptions about independence). The NBC classified art and music with 72% and 28 % accuracy for a combined 61.04% accuracy. These results were surprising similar to our regression accuracy.

For the logistic regression model that we have built, the predicting power of the model was not meeting the success ratio of what we have expected when we first started this project.  Reasons that might have caused this issue are the small numbers of articles that we have used for the training purpose and the numbers of words in the regression model.  In general, the document classification model takes in many training materials to build the initial dictionary.  In our case, we only used 60 articles, constructed a dictionary with 8,196 words, and used 30 variables for the logistic regression.  If we had more training articles, we might have been able to get more solid results than what we saw.