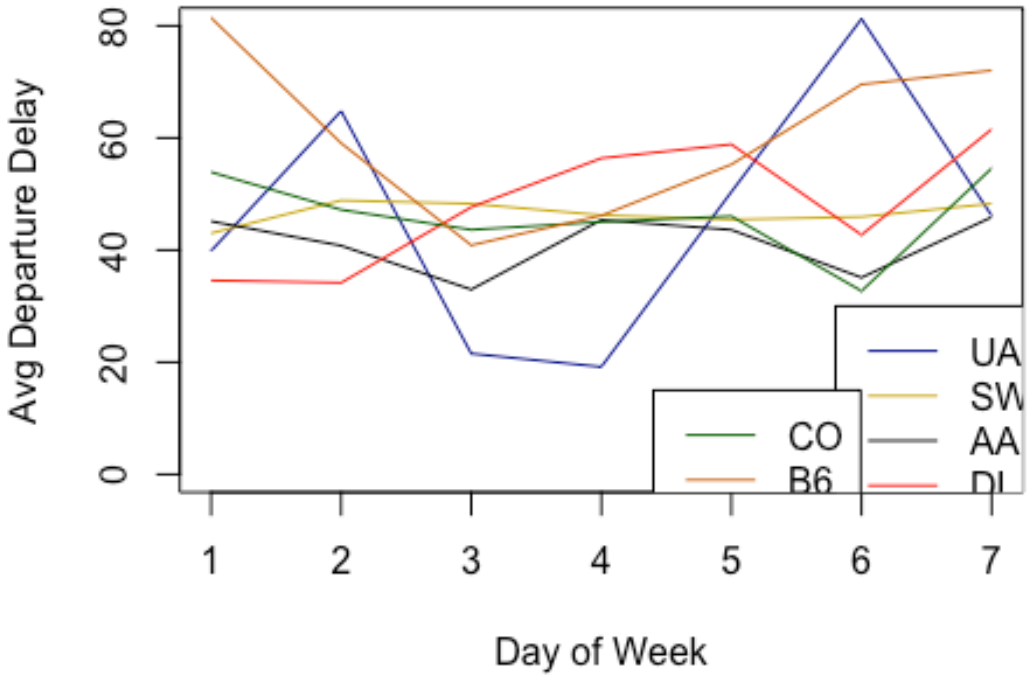
Homework2

Andrew Bohl

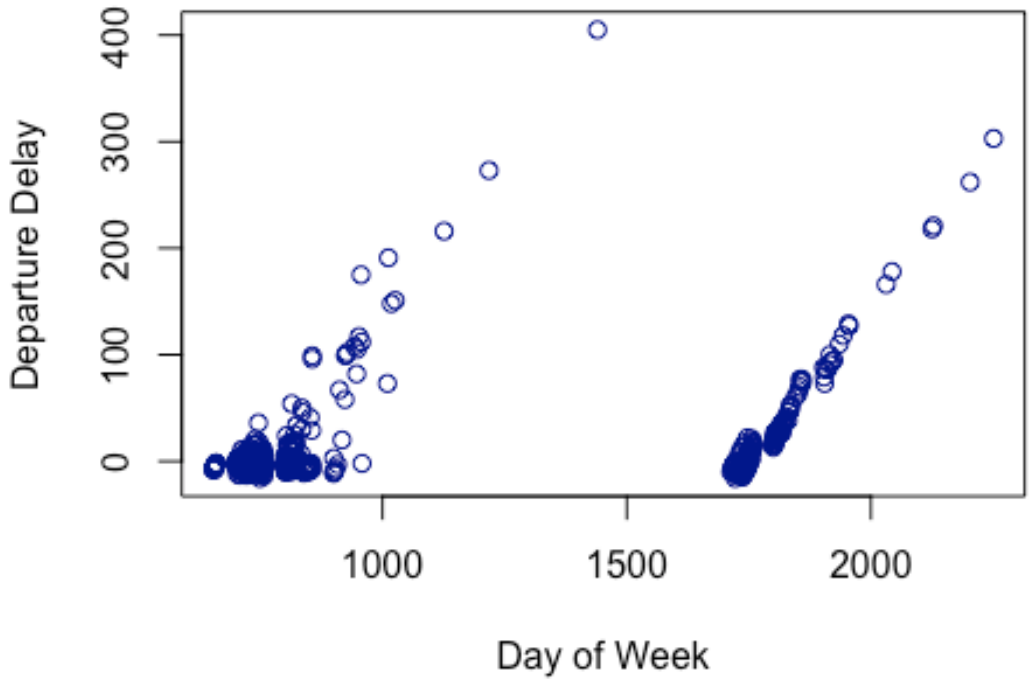
8/15/2018

ABIA 2008 Outbound Delays

Austin Bergstrom hosts hundreds of flights per day. However, unless you final destination is too a large city, you will most likely have to connect through another airport and unless you have elite status on an airline, the plane at your connection will not wait if you are late. So when is the best time to depart from Austin to guarantee that you can make your connecting flight?



The graph above shows the average delay time for the most popular airlines flying out of Austin Bergstrom. Along the X axis we have day of the week plotted numerically. 1 in this case corresponds to Monday with 7 as Sunday. We can see from the graph that midweek is the best time to travel with the lowest average delay time for most airlines.



We saw earlier that on Wednesdays and Thursdays, United Airlines will have the shortest delay time of all airlines flying out of Austin. Now when we look at only United's flights out of Austin, we can see a general trend that morning flights are delayed less often than evening flights. This is most likely due to a chain reaction in delayed flights throughout the day. Once a single flight is delayed, the flights following will most likely be delayed as well. While we only see the data for United Airlines in this graph, we see similar trends throughout all airlines provided in the data.

Author Attribution

For both models I decided to use a NaiveBayes algorithm, however, the differences in the two models come from the inputs. In general, processing the articles into vectors of numbers was similar. All words were turned to lower case, all numbers were dropped and punctuation stripped. One of the key differences however came in the second model when we kept the stem of the word rather than the full word itself.

Model 1

Model one looked at TFIDF weights for every word in the testing articles. The inputs were vectors of TFIDF weights which were smoothed out to account for words in the testing set not seen in the original training set. By adding this small count we could assure the probability of any word was not zero.

```
sum(y == nb.pred) / 2500
```

```
## [1] 0.3684
```

```
AuthorPert1/50
```

```
##  [1] 0.46 0.22 0.52 0.04 0.32 0.62 0.12 0.04 0.12 0.50 0.68 0.52
0.24 0.26
## [15] 0.30 0.50 0.48 0.44 0.24 0.36 0.34 0.38 0.26 0.28 0.58 0.26
0.56 0.66
## [29] 0.74 0.34 0.28 0.22 0.64 0.32 0.16 0.44 0.32 0.26 0.26 0.64
0.32 0.48
## [43] 0.54 0.18 0.46 0.34 0.36 0.28 0.26 0.28
```

When we look at the accuracy of our model, it correctly predicts the author 37% of the time. While this is not great, it is still better than randomly guessing at which author wrote each article. In addition, when we look at the percentage of correct attributions by author we can see that it will at least get one article correct for each author. The author that we are most accurate with is Lynnley Browning(29) with close to 3/4 of articles being correctly attributed to her.

Model 2

For model two, we took a look at the strict frequency counts rather than the TFIDF weights. In this case we restricted the dictionary to words only seen in the training set. This inherently has its disadvantages which are shown in the overall accuracy but this was chosen to simplify the computation time. In an attempt to counteract the fewer amount of words, we looked at the stems of words instead of full words to possibly provide more similarity between articles.

```
sum(y == nb.pred2) /2500
```

```
## [1] 0.2828
```

```
AuthorPert2/50
```

```
##  [1] 0.66 0.72 0.08 0.50 0.28 0.08 0.00 0.42 0.04 0.20 0.52 0.32
0.08 0.14
## [15] 0.00 0.66 0.42 0.12 0.18 0.22 0.26 0.24 0.06 0.12 0.12 0.76
0.78 0.64
## [29] 0.22 0.16 0.04 0.14 0.58 0.08 0.00 0.24 0.06 0.90 0.30 0.66
0.88 0.04
## [43] 0.02 0.00 0.22 0.00 0.50 0.44 0.04 0.00
```

In the end, we only achieve 28% accuracy, which is much less than the first model. However, we see something interesting now looking at the author percentages compared to model 1. We see more extreme values in this case than before. In two cases we can correctly attribute an article to the right author with 90% accuracy. On the other hand, multiple authors are never predicted correctly, with many others at an accuracy of less than 10%.

```
##       [,1] [,2]
##  [1,] 0.46 0.66
##  [2,] 0.22 0.72
##  [3,] 0.52 0.08
##  [4,] 0.04 0.50
##  [5,] 0.32 0.28
##  [6,] 0.62 0.08
##  [7,] 0.12 0.00
##  [8,] 0.04 0.42
##  [9,] 0.12 0.04
## [10,] 0.50 0.20
## [11,] 0.68 0.52
## [12,] 0.52 0.32
## [13,] 0.24 0.08
## [14,] 0.26 0.14
## [15,] 0.30 0.00
## [16,] 0.50 0.66
## [17,] 0.48 0.42
## [18,] 0.44 0.12
## [19,] 0.24 0.18
## [20,] 0.36 0.22
## [21,] 0.34 0.26
## [22,] 0.38 0.24
## [23,] 0.26 0.06
## [24,] 0.28 0.12
## [25,] 0.58 0.12
## [26,] 0.26 0.76
## [27,] 0.56 0.78
## [28,] 0.66 0.64
## [29,] 0.74 0.22
## [30,] 0.34 0.16
## [31,] 0.28 0.04
## [32,] 0.22 0.14
## [33,] 0.64 0.58
## [34,] 0.32 0.08
## [35,] 0.16 0.00
## [36,] 0.44 0.24
## [37,] 0.32 0.06
## [38,] 0.26 0.90
## [39,] 0.26 0.30
## [40,] 0.64 0.66
## [41,] 0.32 0.88
## [42,] 0.48 0.04
```

```
## [43,] 0.54 0.02
## [44,] 0.18 0.00
## [45,] 0.46 0.22
## [46,] 0.34 0.00
## [47,] 0.36 0.50
## [48,] 0.28 0.44
## [49,] 0.26 0.04
## [50,] 0.28 0.00
```

Conclusion

In the end, we can reasonably assume that Darren Schuettler(7), Edna Fernandes(9), and Mure Dickie(35) are all authors that are hard to attribute an article too. While neither model is good at prediction overall, they both have relatively good attributions for some authors. However, neither model can accurately predict these authors, leading use to believe that they have a general writing style or write articles similar to others.

After comparison model 1 is a much better model to use for overall prediction. While model 2 can predict articles by Roger Fillion(41) and Peter Humphrey(38) incredibly accurately, unless we look at just those two, model 1 will provide better results.

After comparing the two models here, a third model may be necessary to fit the data better. After some trial and error with PCA we can get low out-of-sample training error rates, however, in practice I was unable create a model using PCA on the full data set. This would definitely be an area to research into more in the future.

Association Rule Mining

After looking at all of the association rules that the apriori algorithm produced, I decided to trim the rules based on thresholds for lift and confidence. To begin, the maximum lift of all rules was a little over 4, so in my final selection I chose to cut all rules that had lift less than 3. In addition, for confidence, I found that a cut of .35 coupled with the lift cut provided a healthy amount of rules to gain valuable insights without being overloaded with rules.

After applying our thresholds and looking at the newly provided rules, we can see that most of our rules revolve around vegetables and fruit purchases. The most common association occurs with purchasing vegetables together. When one type of vegetable, for example, root vegetables, are purchased, it is very likely that other vegetables will be purchased along with them. In addition, the purchase of one type of fruit, being tropical or citrus will be associated with purchasing the other in addition to some type of vegetable. All of the rules that we have seen in this case make sense as they are staples to a balanced diet. Those who are looking at one type of fruit/vegetable will not generally stick to only eating that type, but all kinds of the specific food category.

One of the last associations we see with our threshold cut association rules is dairy products being purchased together. Whole milk, yogurt, whipped cream are all often purchased together, which could be due to the close proximity to one another in a grocery store. Often times all three of these items are stocked next to each other leading to consumers buying these items together.