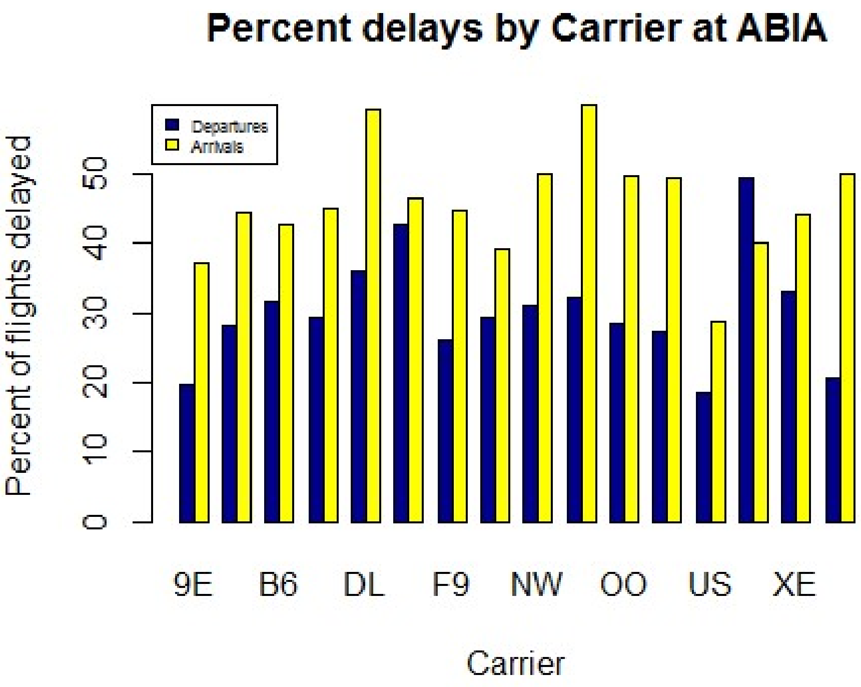
Exercises02_Slocum

Chase Slocum

August 12, 2016

Flights at ABIA



Overwhelmingly, flights are more often delayed in arriving to ABIA than upon departing regardless of carrier with the notable exception of WN (Southwest).

Author Attribution

I used two different models:

- A Naive Bayes model using term frequencies
- A distance vector space model using TFIDF

Both models were built using the tm package in r. The vector space model compared the distance between each test set document and the vector for each author of the average TFIDF for each word. The accuracies of the models are below:

```
                            Accuracy
Naive Bayes: Term Frequency   0.6036
Vector Space: TFIDF           0.5400
```

Clearly, the Naive Bayes peformed better, so that is my model of choice. The most common confused authors are:

```
 [1] "HeatherScoffield & DarrenSchuettler"
 [2] "JohnMastrini & JanLopatka"
 [3] "JoeOrtiz & AlexanderSmith"
 [4] "ToddNissen & DavidLawder"
 [5] "JohnMastrini & AlanCrosby"
 [6] "JaneMacartney & ScottHillis"
 [7] "PeterHumphrey & TanEeLyn"
 [8] "MarcelMichelson & PierreTran"
 [9] "ScottHillis & JaneMacartney"
[10] "DavidLawder & ToddNissen"
```

Association Rule Mining

In building association rules for the grocery baskets, to use .3 as my confidence threshold and 2.75 as my lift threshold because I was interested in finding a few very effective rules that had enough confidence to suggest the relationship was more than just coincidence. It was clear that many of the rules were being driven by consistently purchased items like milk and other vegetables. The rules I found are below:

```
  lhs                      rhs                   support confidence     lift
1 {beef}                => {root vegetables}  0.01738688  0.3313953 3.040367
2 {curd,
   whole milk}          => {yogurt}           0.01006609  0.3852140 2.761356
3 {citrus fruit,
   root vegetables}     => {other vegetables} 0.01037112  0.5862069 3.029608
4 {citrus fruit,
   other vegetables}    => {root vegetables}  0.01037112  0.3591549 3.295045
5 {root vegetables,
   tropical fruit}      => {other vegetables} 0.01230300  0.5845411 3.020999
6 {other vegetables,
   tropical fruit}      => {root vegetables}  0.01230300  0.3427762 3.144780
7 {other vegetables,
   whole milk}          => {root vegetables}  0.02318251  0.3097826 2.842082
```

Because many of the fruit and vegetable items are multi-item groups, it is difficult to decipher how exactly some of the item sets are connected, but others are clear. For instance, beef -> root vegetables makes sense. People might be making a beef stew or having steak and potatoes. The general overlap of fruits and vegetables is not surprising either as they are generally located in the same part of the store, so a customer buying one is going to spend time near the other items. The second rule with curds, milk, and yogurt is potentially evidence of the same concept.