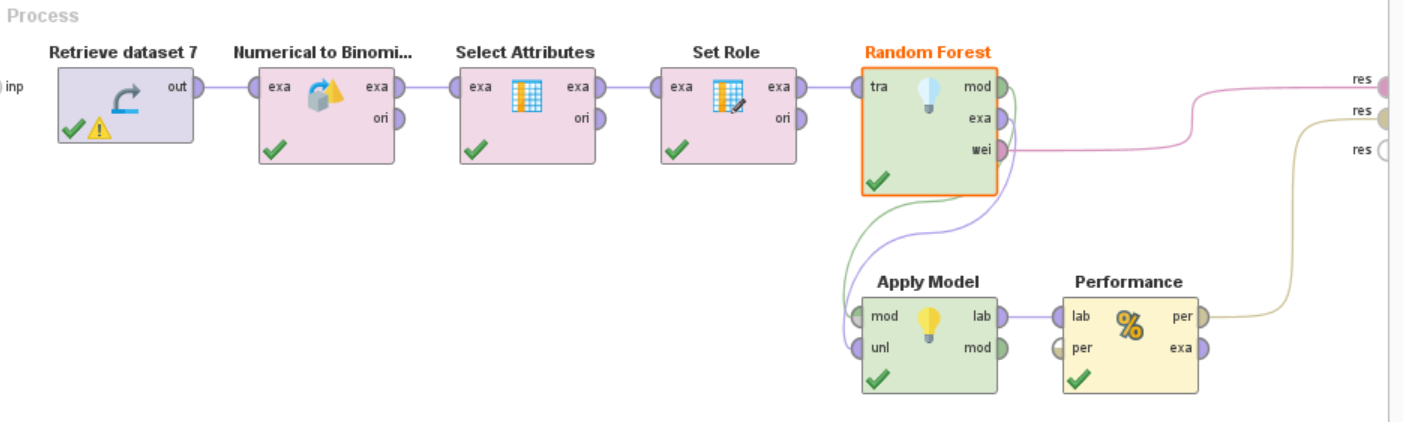
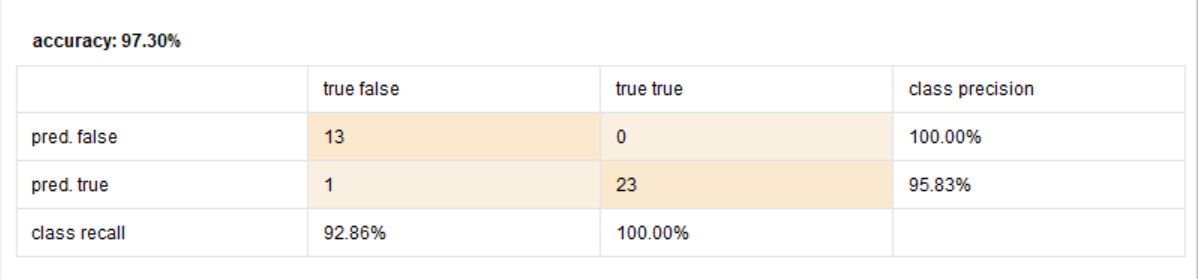
**MDS 560 Week 6 Hands-On Accelerator**

Your deliverables and hands-on activities for this week are:

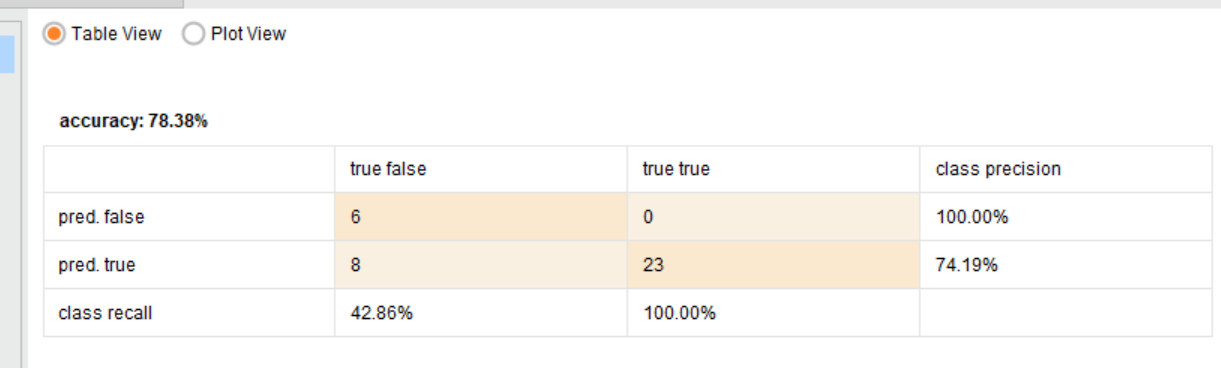
1. Import Dataset 7 (Wordstat Crosstab Export of project risk dataset) into Rapidminer. Perform an analysis and validation using Random Forest. Which variables appear to be most important?

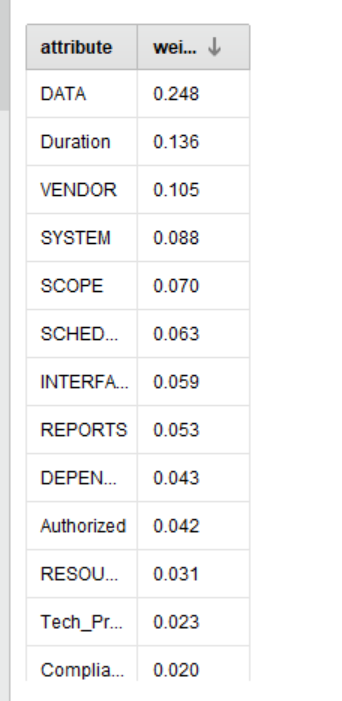
Results: I didn’t do too much work on the variables since I know that a decision tree-based model will weed out unimportant variables. It looks like the overall model performed well, perhaps too well with an overall accuracy of 97.3%. This high accuracy level inspired me to review the default presets on the random forest model. The first model was built with 100 trees and a max depth of 10. Of course it had a high accuracy, it was overfitting the model! I changed this number to 10 trees and a max depth of 3. This model performed less accurate, as expected. This is a result of learning the general patterns of the training data instead of just memorizing the training data which will perform better in the long run. What is interesting about this model is that it is able to predict the outcome of success well, but found it a bit more difficult to identify project failure.

It appears that the mention of Data, Duration, Vendor, System, and Scope are among the top predictors of project outcome. These major concepts of project success are certainly key factors in analyzing risk!



Max 10 trees and 3 dept:



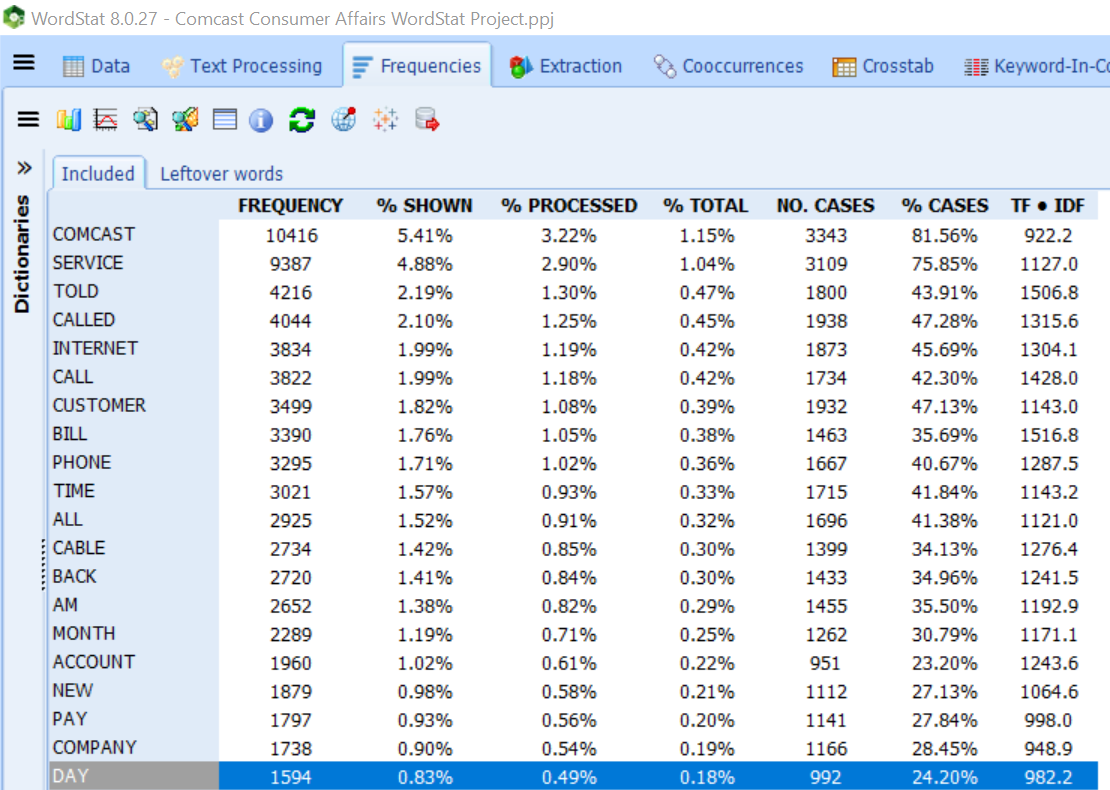


2. Import the Comcast Dataset (Comcast\_consumeraffairs\_happy2) into Wordstat. Cut and paste the following screen shots:

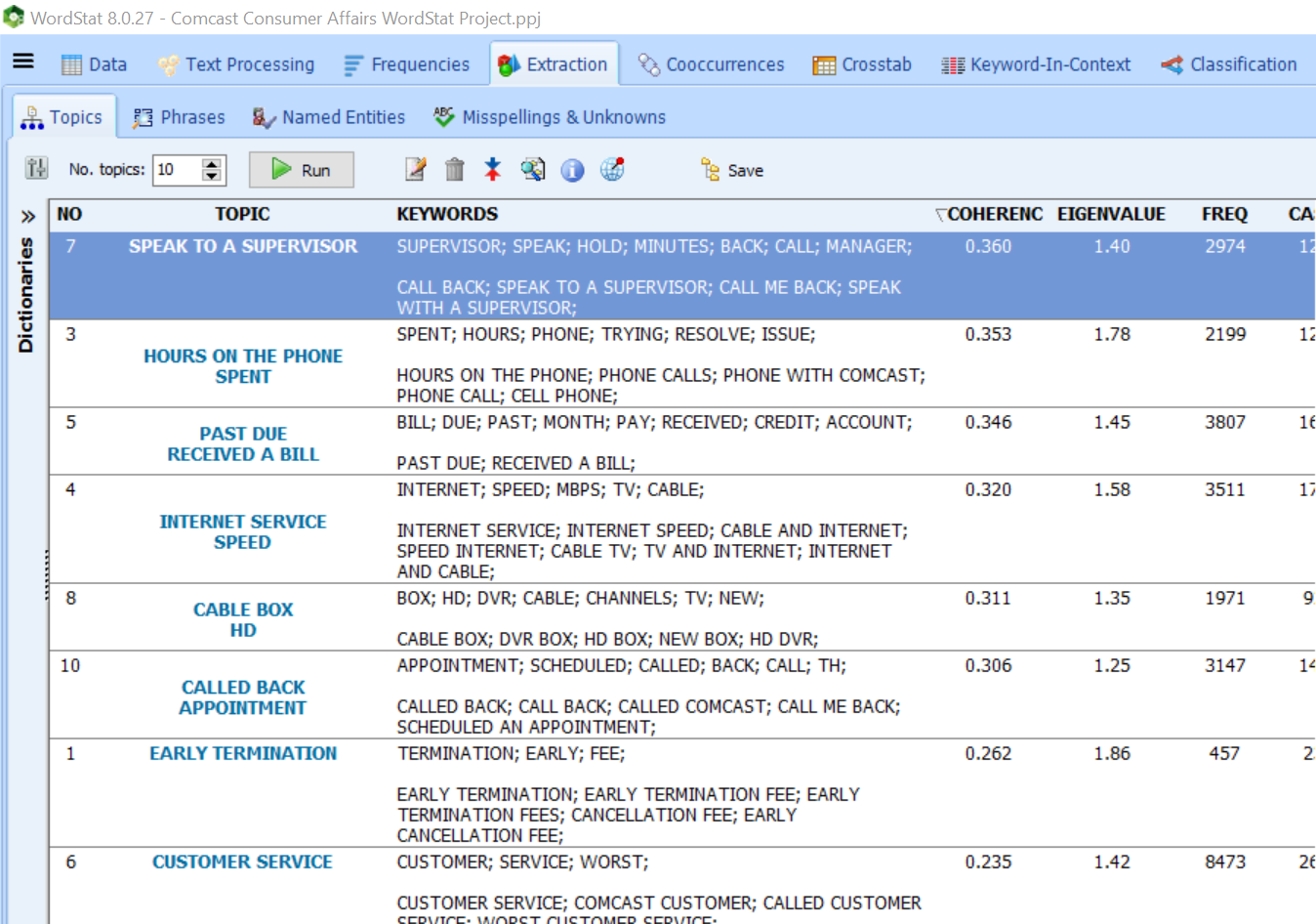
* The top 20 bag of words by frequency
* The top 10 topics extracted
* The top 10 phrases extracted
* The mapping of concurrences (bubble chart)
* The proximity plot of words in proximity to the word “Cancel”

Results:

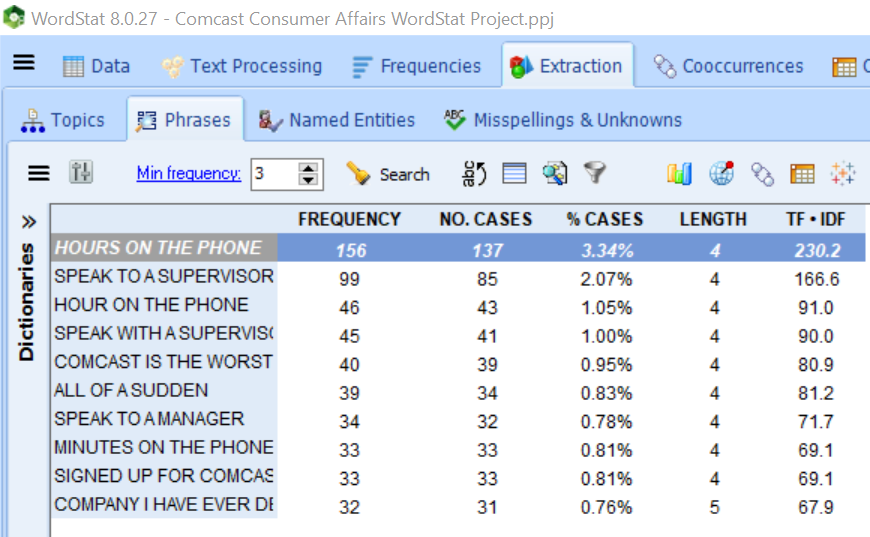
Top 20 BOW by Frequency



Top 10 Topics Extracted



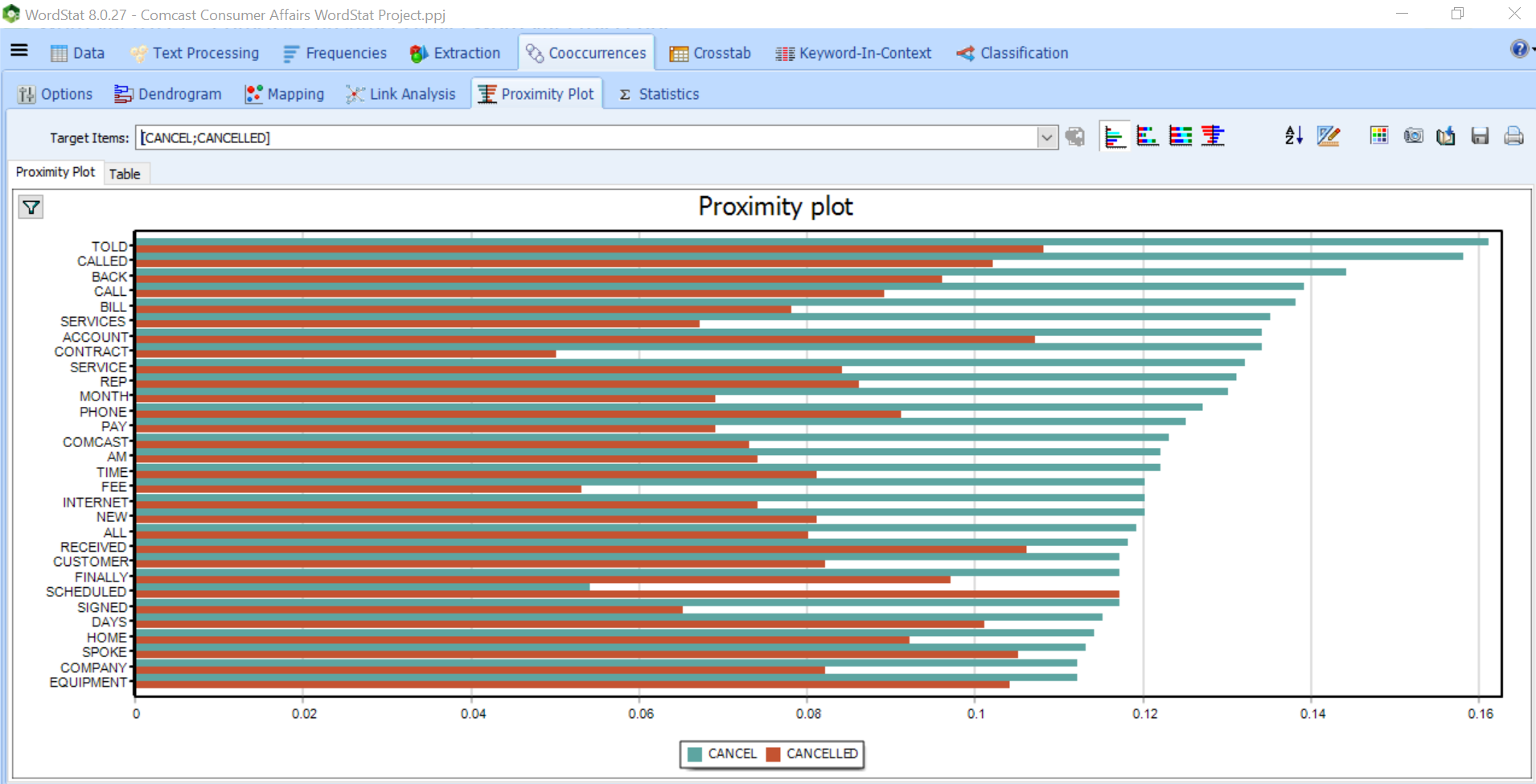
Top 10 Phrases Extracted:



Mapping of Concurrency



Proximity Plot of Words in Relation to ‘Cancel’

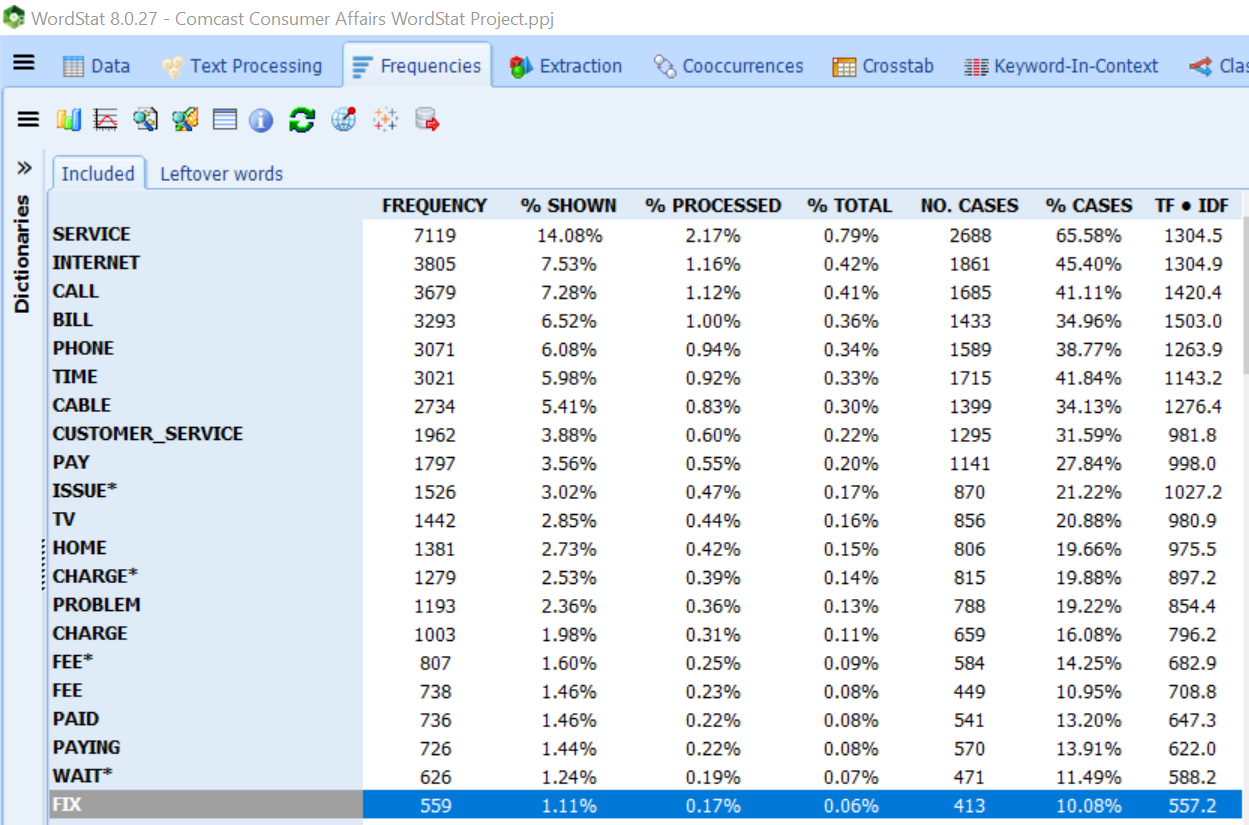


3. Next, while in Wordstat with the Comcast Dataset go to the Dictionary tab and pull in the Dictionary file “Threat Master 2020”. (Do not check “Categories Only”). Now re-perform the following steps and cut and paste the following screen shots:

* The top 20 bag of words by frequency
* The top 10 topics extracted
* The top 10 phrases extracted
* The mapping of concurrences (bubble chart)
* The proximity plot of words in proximity to the word “Cancel”

Results:

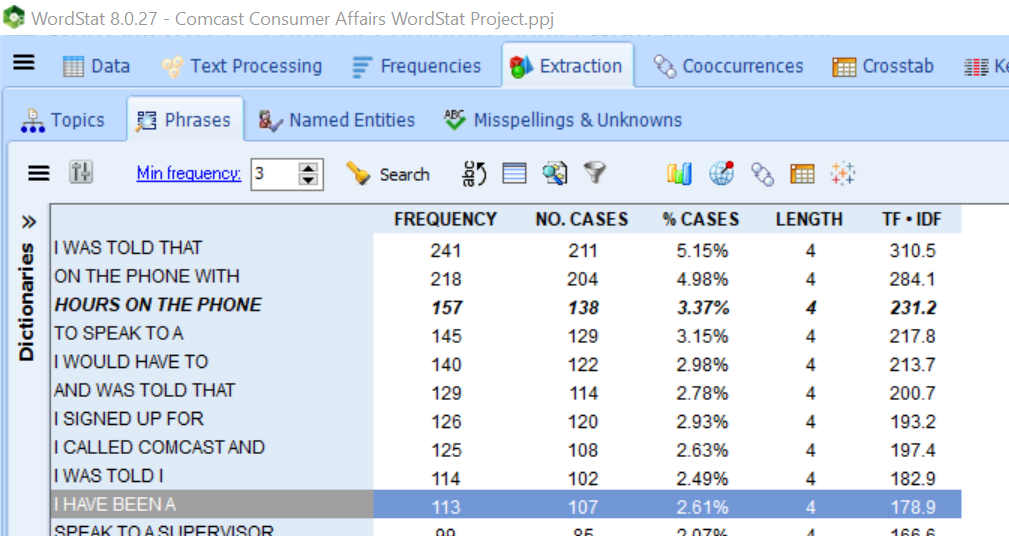
Top 20 BOW by Frequency



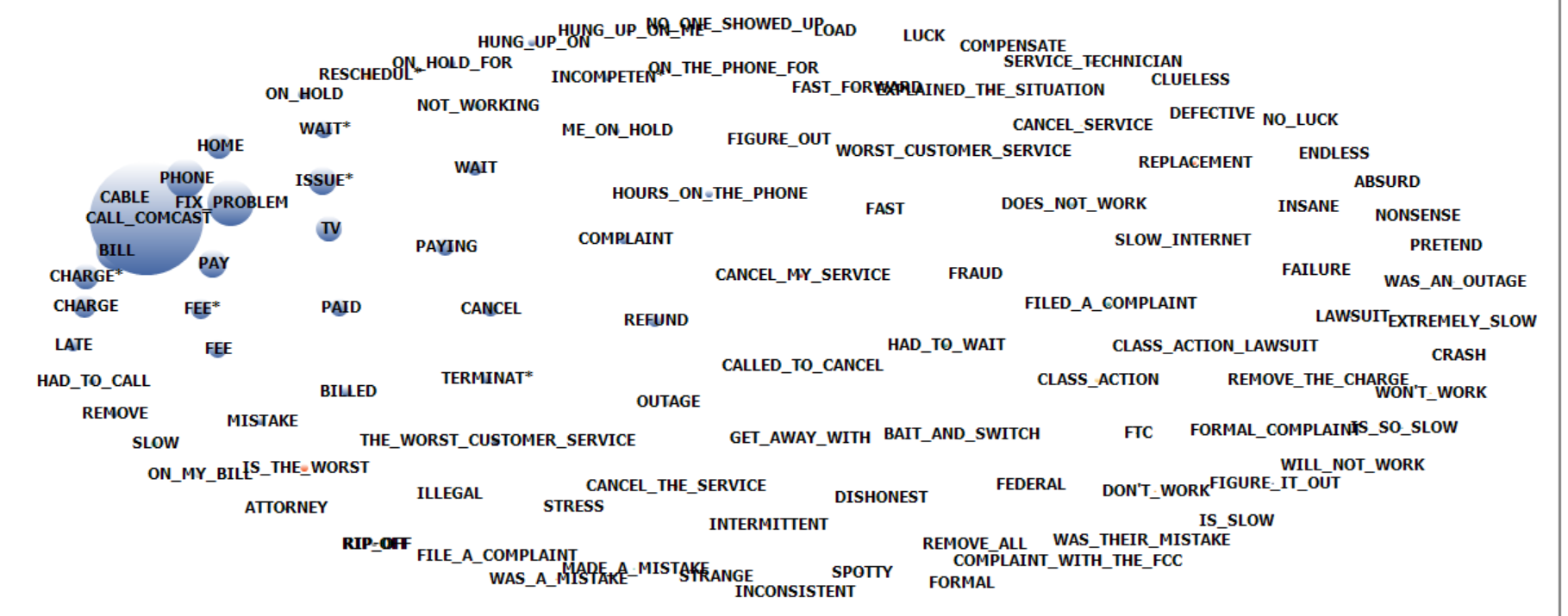
Top 10 Topics Extracted



Top 10 Phrases Extracted:



Mapping of Concurrency



Proximity Plot of Words in Relation to ‘Cancel’



4. Next, while in Wordstat with the Comcast go back to the frequency view and temporarily remove the word Cancel by hovering over the word, right clicking and removing. Then under the Classification menu run a Naïve Bayes Classification for the label “Kancel”. Show the confusion matrix and list the top 20 predictor variables.

Results: The overall precision/recall were near 50% so it does not seem like the classifier did much better than guessing. It was able to identify people who do not cancel better than those who did, but it is also the larger class and easier to identify.

The top features do give us some insight on what customers who are more likely to cancel are discussing. For example, the customers are talking about the service that is being provided to them such as their phone or their bill. Honestly, I do not think any of this information would be surprising to hear from a company such as comcast. I would be interested to know if there is some correlation of what is being discussed on these calls in certain regions of their areas. Perhaps a more rural area gets calls regarding internet service or certain call centers/representatives are providing bad customer service.

