Lab

# Lab 9 – Rule Mining and Word Association

Chapter 17 introduces association rules mining – a very flexible and easy-to-understand form of unsupervised machine learning. Association rules mining is also sometimes called market basket analysis, and indeed the chapter works an example that focuses on groceries. Yet association rules mining can be applied to a variety of data types, basically any dataset where you have a list of “containers” and each container having a list of stuff inside it. Association rules mining looks for commonalities across these containers – what are the combinations of items that frequently occur together.

If you think about it for a minute, you might see that this idea applies to documents (e.g., emails or web pages) and the words that appear in them. Each document can be thought of as a container of words, and within each container, certain combinations of words may appear together. In a previous chapter, we explored this idea by creating a “term-document” matrix. In this exercise, we are going to apply association rules mining to a term-document matrix.

For this exercise there is no sample R code to help you along. Because you have had so much practice with R by now, you can create and/or find all of the codes you need to accomplish these steps:

1. There is a nice, manageable term-document matrix that Yanchang Zhao has created based on a set of tweets about data mining that he extracted: [www.rdatamining.com/data/termDocMatrix.rdata](http://www.rdatamining.com/data/termDocMatrix.rdata)

If for some reason that link doesn’t work, you should visit this page and download the file entitled “termDocMatrix.rdata” onto your computer:

[www.rdatamining.com/data](http://www.rdatamining.com/data)

As you have guessed from the file extension, this is a dataset that is already prepared for opening in R. Run RStudio on your laptop and use the Open File command to load this file. After answering the confirmation message, you will find that a data object called “termDocMatrix” appears in your environment window. Inspect this data object.

2. For association rules mining (and more specifically the apriori() command) to work properly, you want items in containers/baskets to be your columns and the rows to be your containers. You will find that this data object is set up the opposite way: It has the terms as rows and the documents as columns. You will need to transpose the dataset, and fortunately R has a command to do that very easily. Do some research to find that command and how to use it. Then transpose your dataset and place it in a new data object.

3. Next, apply all of the techniques you learnt from Chapter 17. This means that you will have to load the arules package, run apriori(), set the parameters correctly, inspect the results, visualize the results using the arulesViz package, and make sense out of what you find. You should set your parameters so you generate at least 20 rules.

4. At the end of your code file for this exercise, write a few sentences interpreting the results of this analysis and describing how this technique might be valuable in making sense out of large sets of documents (e.g., emails).