# BUS212f: Analyzing Big Data II Fall 2018

## Case # 4: Text Mining Airbnb Reviews

### Introduction

Ted Kwartler of Liberty Mutual was your instructor in two DataCamp courses on Text Mining. In the second course, the concluding case study was an analysis of customer reviews on Airbnb for Boston-area properties. In this assignment, your team will revisit that analysis using a more recent set of reviews for the properties we’ve analyzed before.

Consider this case as both an extension of the DataCamp course assignment (that is, you should be able to re-use some of the code from that activity) and a preparation for our final project.

Technical Note: Package **qdap**, which was featured in DataCamp, may be difficult to install and run, depending on your operating system. You are advised to use packages **tm, tidytext**, **RWeka**, and the various graphing and other support packages shown in our text and in DataCamp, as alternatives to **qdap**. Although it does have some attractive features, everything in this assignment can be done with other packages.

### The Task

Using the reviews provided in the csv file called “reviews.csv”, which is available in GitHub in compressed form. This file contains approximately 178,000 reviews of properties in the Boston area. In this file, we have very little information about the properties—we really just have a date and a review. Essentially, you will use **qdap, tm** , **tidytext** and **RWeka** to perform an analysis very similar to those performed in the readings (our text and DataCamp). Specifically, your analysis should do the following:

1. Set a random seed before Step 1 so that your results are fully reproducible. This is especially important in collaborative coding.
2. Because the number of reviews is so enormous, and because we will use these reviews later in the final project, select a random subset of 1,000 reviews for analysis in this case.
3. Perform the usual cleaning steps of removing numbers, punctuation, and stop words. Your team will need to make decisions about customizing your stop words list. In your document, briefly discuss the team decisions about adding or dropping stop words.
4. Use TfIdf weighting to create a TermDocument Matrix. In a sentence, report on the dimensions of the resulting matrix (#rows, #columns).
5. Create and display a well-labeled bar chart of the 15-20 most frequent terms in the reviews. Add your comments about the frequent words. Do they make sense to you?
6. Identify the most common bigrams in your sample. Comment on what these short phrases reveal (or not) in comparison to the single-word analysis.
7. Find word associations (findAssocs() in tm) and create a word network plot to find words associated with the term “location” (similar to example shown in the DataCamp Bag of Words, Chapter 2). Comment on what the plot reveals to your team.
8. Use the BING lexicon to assign positive or negative sentiments to terms, and split your sample into positive and negative comments. Make a pyramid plot to compare the most common terms in positive vs. negative reviews.
9. Repeat the prior step using the FINN lexicon.
10. Consider and discuss the merits of the two lexicons in this use case. Which lexicon does your team prefer to use for these reviews, and why? Using the lexicon of your choice, create a comparison cloud that contrasts the positive and negative terms in your sample.
11. Some reviews talk about the hosts, while others talk about aspects of the property. Following the methods discussed in Chapter 6 (“Topic Modeling”—sections 6.0 and 6.1) of Silge and Robinson’s book *Tidy Text Mining*, use the LDA() function to identify k = 4 topics in these reviews. Report on the most common terms for the 4 topics, and explain what your team thinks the topics represent. (e.g. one topic might be about financial value, and another about location, etc.). Produce and include a graph similar to Figure 6.4 in Silge and Robinson, showing the most frequent terms in the four topics you’ve extracted.
12. Finally, draw a conclusion: Are these differences between positive and negative Airbnb review? Are positive and negative reviews equally common? Are they comparable in intensity?

### Variable Definitions:

This file is simple in structure, and you will focus on Column 6 “comments”. Here are the columns:

listing\_id: a code number for the property

id: a unique ID for a specific rental

date: of the review

reviewer\_id: Unique ID for the reviewer

reviewer\_name: Reviewer’s first name

comments: Review comments. Each entry is a document.

### Data Preparation:

Described above, as typical for text mining.

### Deliverable:

Your team should prepare a pdf document created with R Notebook (markdown) reporting on your analysis and discussing your conclusions. In addition, prepare a four-slide PowerPoint presentation showing:

1. A graph (your choice of type) of most common bigrams
2. Word Association network graph
3. Your “favorite” comparison cloud (based either on BING or FINN)
4. A graph similar to Figure 6.4 in Silge and Robinson, showing the most frequent terms in the four topics you’ve extracted.

Please save the PowerPoint as a pdf as well.

**In all you will upload two pdfs: one from the knitted notebook, and one from PowerPoint.**

The R Markdown document must include all the graphs, analyses, and commentary mentioned in the “Task” section above. If you wish to add commentary, feel free to do so.