MEMORANDUM – ASSIGNMENT
To: BDA 2018

Subject: BDA Assign: Text Mining #2

From: R Bohn
Date: May 6, 2018. Filename error fixed May 8.

For class on Wednesday May 9, we will go through the exercise in Chapter 20.6. We will also use the class to discuss specific problems you are running into for your projects.
Assigned reading: 1)Read chapter 20, especially 20.6.
2) Study one of the advanced references discussed anywhere on BDA2020.wordpress.com. Choose something relevant to your project. (See Option B, below)

Deliverable by 11PM Tuesday May 8. This is an experimental assignment. Do Option A or B, not both!

Option A: Submit a memo asking for advice on a technical difficulty you have encountered in your project. It can be about R, Data Mining algorithms, strategy, or something else. Whenever possible, include specific error messages, examples of R code, screen shots, etc. What resources have you consulted about it, so far?

Option B: Submit a short description of the advanced material you read, from the web site or something that you encountered yourself. Write it to be useful to *current and future students* in the BDA course, not to faculty. Don't try to summarize the content. Instead, you may want to explain:
What technique, issue, industry, situation, etc. does it discuss?
What kind of student projects would it be useful for?
How does it help you with your project?

Due Friday May 11 at noon: Textbook problem 20.3. (Online classified ads.)
Don't answer questions a, b, or d in the book unless you wish to.
Instead, solve question b. *Use a random forest* instead of (or in addition to) logistic regression. Write the usual memorandum, to the general manager of the website, explaining what they should do to filter ads.

Discuss how well it will work. Assume that the cost of false positives (accepting an irrelevant ad) is lower than the cost of false negatives. Use that to choose a cutoff. (See textbook page 127ff, especially page 133.)

(Errata in the textbook. There are 3 different names for the zip file! On the book's web site, it is http://www.dataminingbook.com/system/files/AutoAndElectronics.zip)

For the week of May 14:

We will look at another algorithm called LASSO (least absolute shrinkage and selection operator). **Read**: Gareth James et al, ISLR *An Introduction to Statistical Learning with Applications in R*. Chapter 6 introduction, and Chapter 6.2 and 6.4. Also look at Lab 2.