ASSIGNMENT
To: BDA 2018

Subject: BDA Assignment Week 6: Text Mining
From: R Bohn
Date: April 30, 2018

We are going to look at text mining, which is important to many of the projects. Text is becoming widely available and contains lots of information that people care about. Because it is written for people and not machines, making use of it can be tricky.

- Read Chapter 20 in main textbook.
- Read all of Chapter 10 in book by Provost and Fawcett. "Representing and Mining Text" in * Data Science for Business* by Foster Provost and Tom Fawcett. UCSD library http://melvyl.worldcat.org/title/data-science-for-business/oclc/858807267?ht=edition&referer=br
- If your project involves any use of text, you will need to go beyond what is in these textbooks. There are many ways to use text. I put together a memo with more resources on the subject. Get familiar with the resources in the memo that may be relevant to your project. If you plan to work with Twitter, get familiar with the student paper mentioned on this page: https://irgn452.wordpress.com/how-to-do-projects/text-mining/ If you are going to make no use of text (aside from categorical information) you can skip this.
- Assignment: Take about 60 words from a general news article in the NY Times or similar publication. Start at the beginning of the online text, including any information such as headline, dates and name of the author.
- By hand, tokenize the text. Submit it as two versions of the same text, one the original and the other one stemmed. The purpose of this is to think about where stemming works well, and where it does not.
- Circle or highlight words where you think the stemming algorithm makes an error, e.g. does not capture the true root word.
- This assignment sounds easy, but in the past students have had a lot of trouble with it. Remember that the computer *just follows the rules*. It does not "understand" that predicate and predict are completely different words, so when you stem them, they will be identical or very close. That is not what you *want*, but it is what we get.

Process the text according to the rules in Provost and Fawcett: lower case, stem, and

remove stop words. Hint: make it easy to read. Print the text in alternating lines. Odd numbered lines show the original text, even numbered lines are your processed version of the same text. Use a large font, alternating colors, and any other tricks. For example, you could put all the changed words in italics or a different color.

Example: write writer writing writhe writes writers written writhing writings wrong . Will these all be stemmed the same? Should they be?