Task 1 - Data acquisition and cleaning

Help Center

Large databases comprising of text in a target language are commonly used when generating language models for various purposes. In this exercise, you will use the **English database** but may consider three other databases in German, Russian and Finnish.

The goal of this task is to get familiar with the databases and do the necessary cleaning. After this exercise, you should understand what real data looks like and how much effort you need to put into cleaning the data. When you commence on developing a new language, the first thing is to understand the language and its peculiarities with respect to your target. You can learn to read, speak and write the language. Alternatively, you can study data and learn from existing information about the language through literature and the internet. At the very least, you need to understand how the language is written: writing script, existing input methods, some phonetic knowledge, etc.

Note that the data contain words of offensive and profane meaning. They are left there intentionally to highlight the fact that the developer has to work on them.

Tasks to accomplish

1. Tokenization - identifying appropriate tokens such as words, punctuation, and numbers. Writing a function that takes a file as input and returns a tokenized version of it.
2. Profanity filtering - removing profanity and other words you do not want to predict.

Tips, tricks, and hints

1. **Loading the data in**. This dataset is fairly large. We emphasize that you don't necessarily need to load the entire dataset in to build your algorithms (see point 2 below). At least initially, you might want to use a smaller subset of the data. Reading in chunks or lines using R's `readLines` or `scan` functions can be useful. You can also loop over each line of text by embedding `readLines` within a `for` / `while` loop, but this may be slower than reading in large chunks at a time. Reading pieces of the file at a time will require the use of a file connection in R. For example, the following code could be used to read the first few lines of the English Twitter dataset:

   con <- file("en_US.twitter.txt", "r") readLines(con, 1) ## Read the first line of text readLines(con, 1) ## Read the next line of text readLines(con, 5) ## Read in the next 5 lines of text close(con) ## It's important to close the connection when you are done

See the `?connections` help page for more information.

1. **Sampling**. To reiterate, to build models you don't need to load in and use all of the data. Often relatively few randomly selected rows or chunks need to be included to get an accurate approximation to results that would be obtained using all the data. Remember your inference class and how a representative sample can be used to infer facts about a population. You might want to create a separate sub-sample dataset by reading in a random subset of the original data and writing it out to a separate file. That way, you can store the sample and not have to recreate it every time. You can use the `rbinom` function to "flip a biased coin" to determine whether you sample a line of text or not.

2. If you need a refresher on regular expressions, take a look at Jeff Leek's lectures from Getting and Cleaning Data: Part 1 Part 2

Created Mon 25 Aug 2014 10:15 AM PDT

Last Modified Fri 31 Oct 2014 8:35 AM PDT