Making Text Analysis Accessible with R and Shiny

David Allbritton http://davidallbritton.com

DePaul University

Purpose

Text analysis and text mining are powerful tools for automatically extracting patterns from unstructured texts, such as texts from open-ended survey questions. The open-source statistical programming language R includes modules for text analysis such as tm (Text Mining) and NLP/openNLP (Natural Language Processing) (http://cran.r-project.org/web/views/NaturalLanguageProcessing.html).

Shiny (http://shiny.rstudio.com/) is an application framework for creating interactive web interfaces to data analysis programs in R. Analysis outputs are wrapped in html markup for viewing in a web browser. Input forms allow the user to change analysis parameters interactively and immediately view the results.

A Shiny application is presented for exploratory analysis of open-ended text data, such as course evaluation comments. Texts are analyzed based on lexical n-gram frequency, and the results are presented as frequency plots or as word clouds. The user can change analysis parameters (such as minimum frequency, whether to use stemming, n-gram size, etc.) interactively while viewing the results. This tool could be useful for generating coding categories prior to doing a traditional content analysis, or for identifying useful search terms for finding representative examples from a large set of open-ended responses. Future extensions could include the use of cluster analysis and latent semantic analysis to group similar responses. The exploration of actual course evaluation data is presented as an example of how R and Shiny can make text mining tools accessible to a wider set of (non-technical) users.

Method

All student responses to open-ended questions about (1) what they most enjoyed, and (2) what they would like to see improved were collected for my online courses taught from Fall 2009 to Spring 2014 (5 sections of Cognitive Psychology, 13 sections of Intro Psych I, and 5 sections of Intro Psych II). Each student response (295 for "enjoy" and 271 for "improve") was treated as one "text" and the results presented here are for all classes combined.

Corpora were stored as comma separated value (.csv) files, which the user could select among using the Shiny User Interface web page.

High-frequency English "stop" words were removed from the corpus, as were a small set of additional user-defined "stop" words in an Excel file.

Results

Aspects of the courses that students most liked or loathed became readily apparent through a bit of experimentation with the n-gram size setting. With n-gram size = 2 it would appear that "discussion questions" are *simultaneously* both the most loved *and* most hated aspect. But with n-gram size = 3, it becomes apparent that students are actually complaining about the workload (suggesting "less discussion questions") rather than complaining about the nature of the discussion assignments. With stemming, it also becomes evident that students enjoyed the "Open Discussion" assignment each week in particular.

Because the optimal n-gram size (and other parameters) may vary from corpus to corpus, it is helpful to be able to interactively change them and immediately see the results. For the corpora used in the shinyaps.io demo site for this project (candidate utterances from the second 2015 Republican Primary Debate), for example, an n-gram size of 2 seems most revealing.

Conclusions

Text mining can be a useful tool for exploring open-ended survey responses. R has packages that make text mining available to non-experts, and Shiny is a useful platform for making those capabilities accessible to a wider audience.

Resources and Downloads

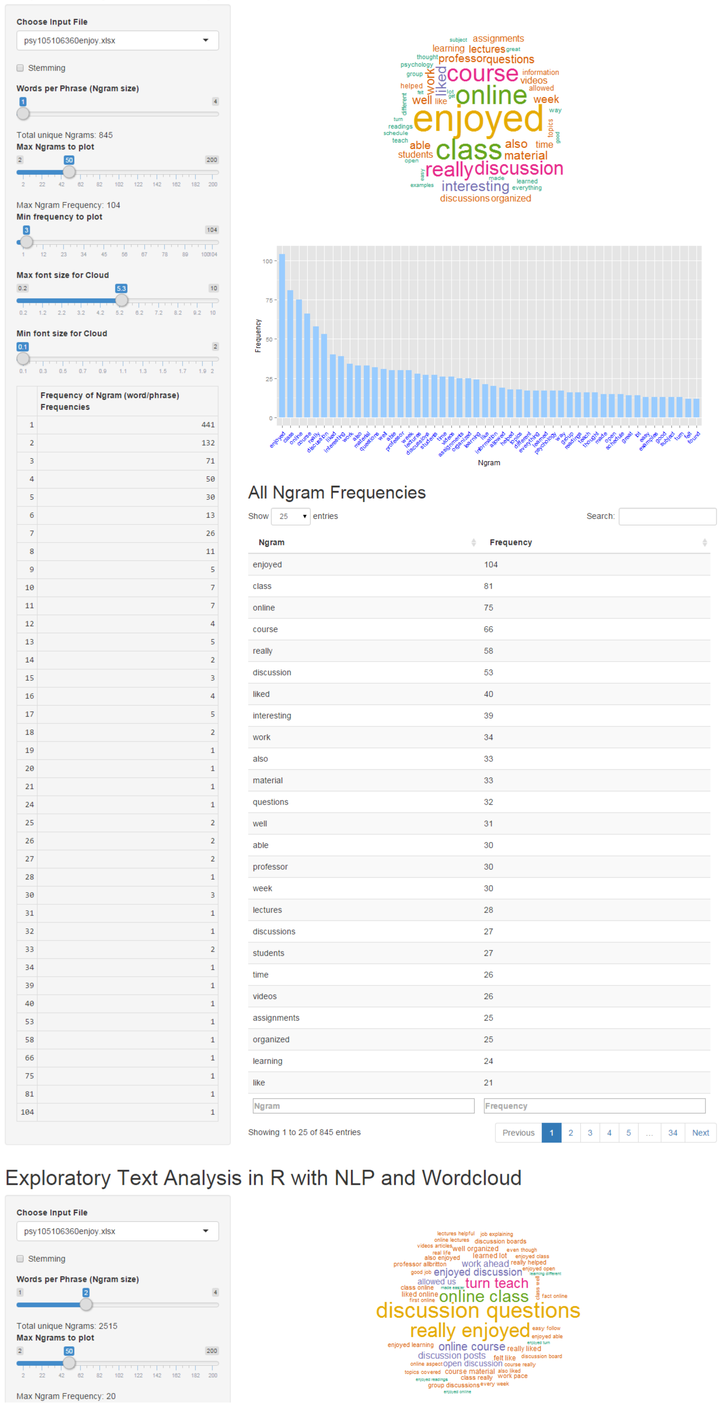
A working demo of the Shiny app is available at shinyapps.io. Source code, sample input files, and this poster are available on github. Links to both can be found at:

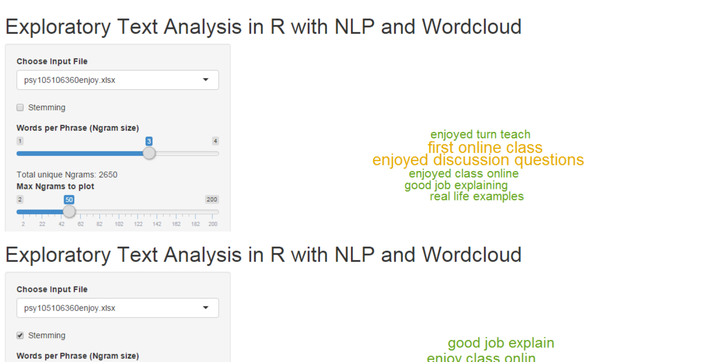
http://davidallbritton.com

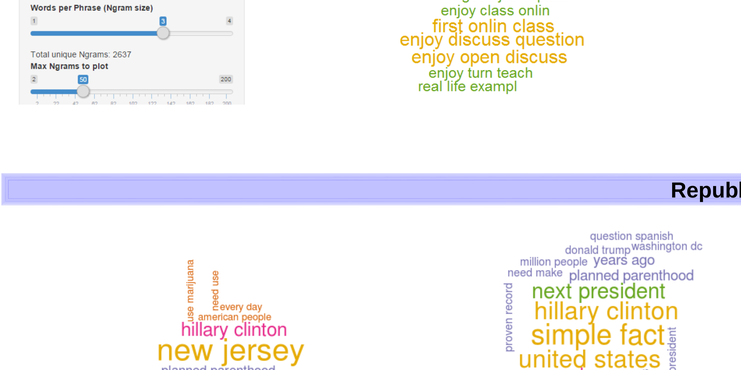
To install and use the app on your own computer:

1. Download and install R
   http://www.r-project.org/
2. Download and install the Desktop version of Rstudio
   http://www.rstudio.com/
3. Download the R script and sample input file (click "Download Zip" if you do not use GitHub)
   https://github.com/davidallbritton/
4. Place the R script and sample input files in a folder on your computer (after extracting from the zip archive)
5. Open the ui.R script in RStudio and click "run app"
6. Select "Run External" in the drop-down menu to the right of "Run App" for best results

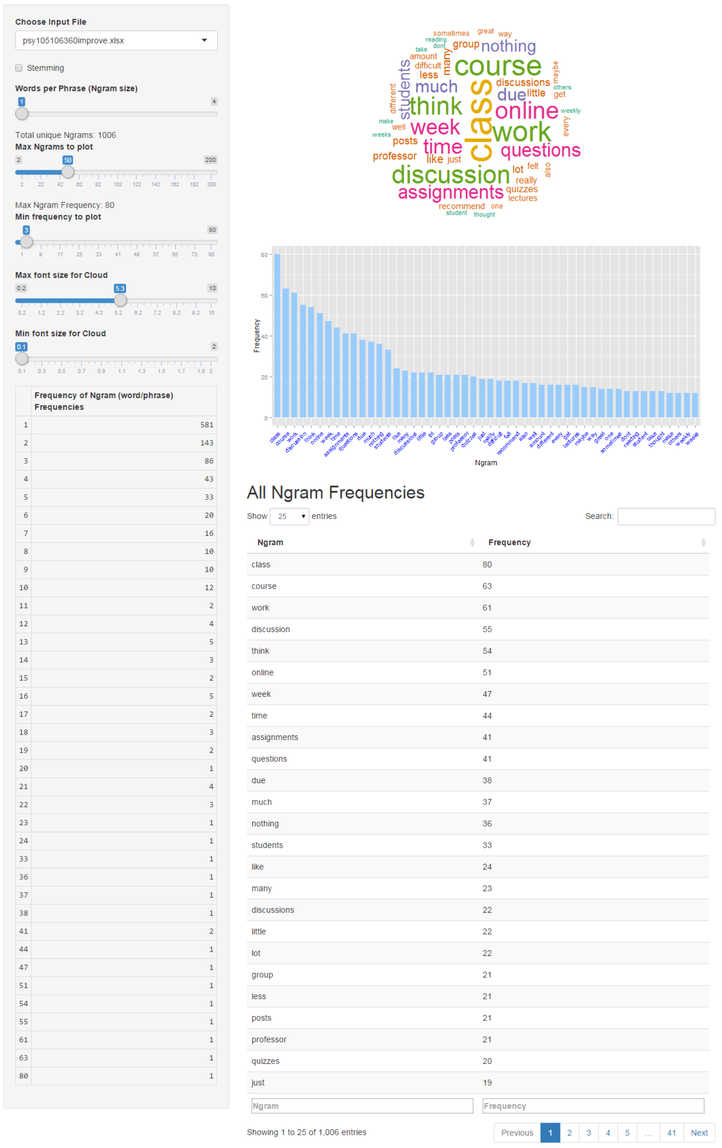
Course Evaluation Question: "Most Enjoyed"

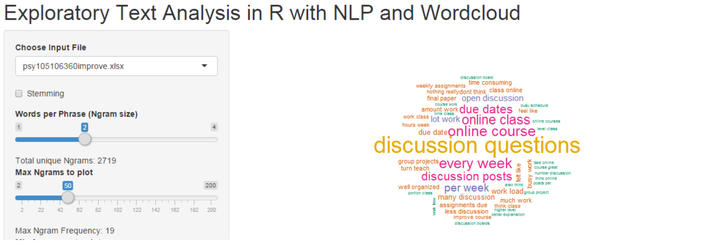


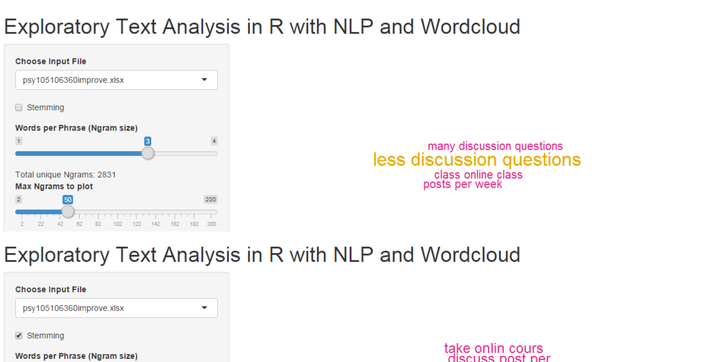


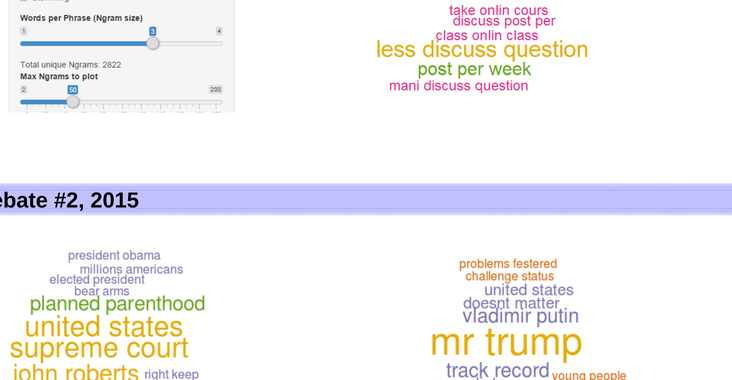


Course Evaluation Question: "Could Improve"









Republican Primary Debate #2, 2015