Exploratory Text Analysis with R and Shiny

David Allbritton

http://davidallbritton.com

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 grouped into "topics" using Latent Dirichlet Allocation (LDA). The user can change analysis parameters (such as minimum frequency, whether to use stemming, n-gram size, number of topics, etc.) interactively while viewing the results. Model results and parameters can be downloaded to an Excel file.

This tool could be useful for generating coding categories prior to doing a traditional content analysis, or automated exploration of a large set of open-ended responses. Source code is provided for use as a teaching tool or further development.

Method

The texts explored consisted of all student responses to open-ended questions about what they "most enjoyed" about my online courses 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 total) was treated as one "text."

Corpora are input as Excel files, one text per row. 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.

R packages *tm* and *NLP* were used to process a corpus and create a term-document matrix. The *wordcloud* package was used to generate word clouds, *SnowballC* was used for stemming, and LDA topic models were created with the *topicmodels* package.

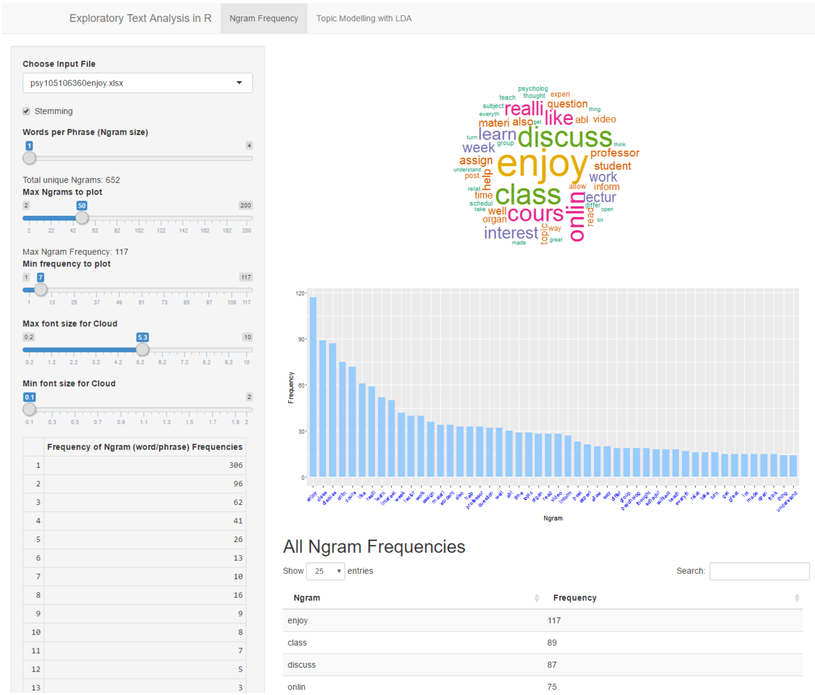
Results

Because the optimal n-gram size, number of topics, and other parameters may vary from corpus to corpus, it is helpful to be able to interactively change them and immediately see the results. In the example shown here, one of the topics appears to focus on flexibility (although in general this corpus did not produce any set of obviously interpretable and distinct topics).

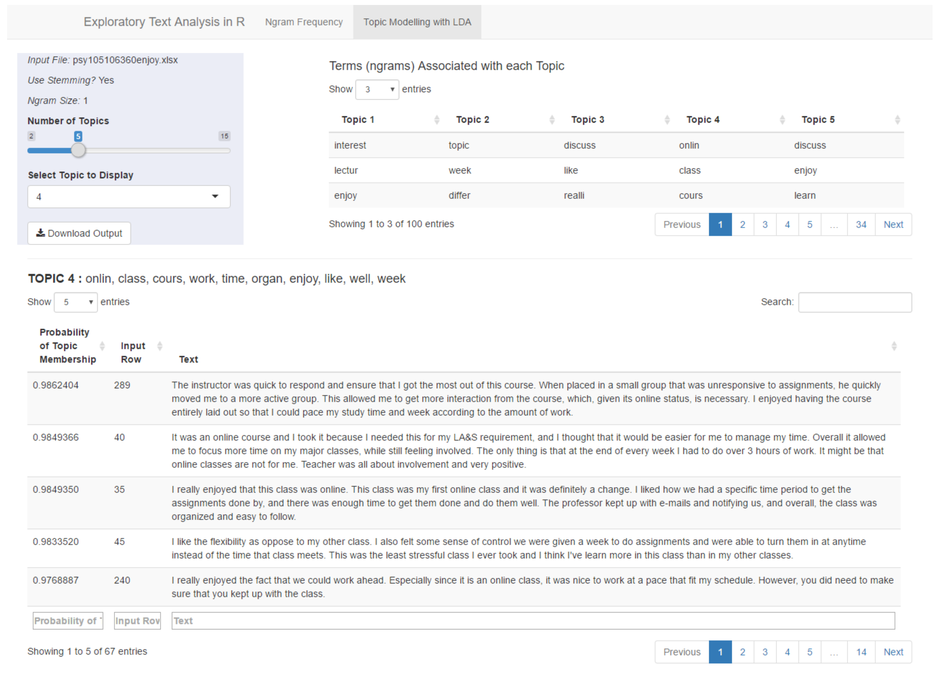
Conclusions

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.

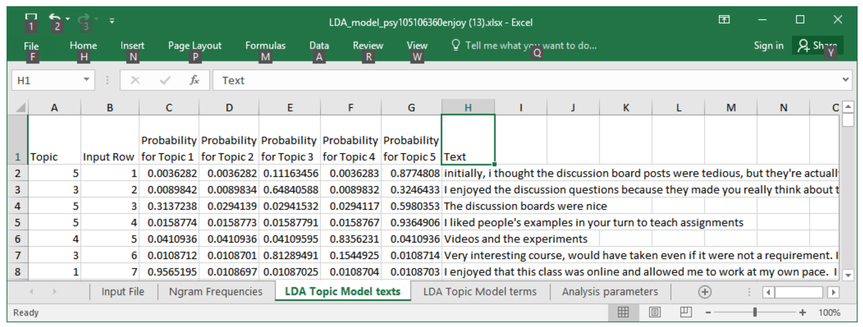
Ngram Frequency Analysis Tab

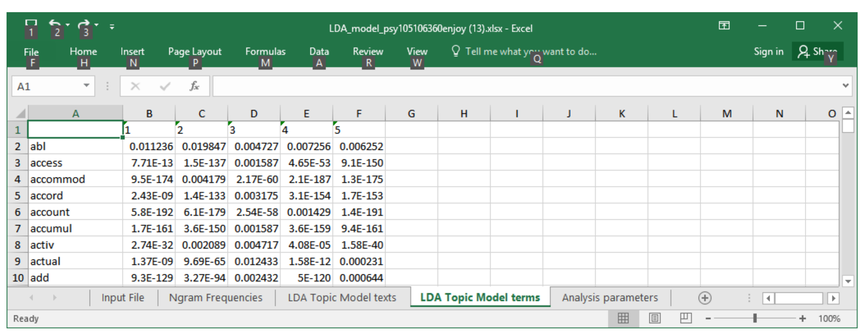


LDA Topic Modeling Tab



Downloaded Results File





Resources and Downloads

A working demo of the Shiny app is at https://shiny.is.depaul.edu

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