Text Mining in R
Word Clouds

Manuel S. González Canché & Cecilia Rios Aguilar
msgc@email.arizona.edu
Cecilia.Rios-Aguilar@cgu.edu

University of Arizona
Claremont Graduate University

INCHER, University of Kassel
June 25, 2012

Notes

Motivation

- Many times qualitative and quantitative researchers face a big problem: the impressive amount of text files (interviews, speeches, discourses, Facebook posts) that need to be coded or summarized.
- This is problematic because finding structure in unstructured data is a challenging process and summarizing thousands of texts can be a task that is nearly impossible for the human being brain to be handled or that would take months of work.
- We can rely on computers to ask them to find that hidden structure and report back to us the outcome.
- The challenge is to detect that structure, because by definition we have eliminate waste or noise in the data.
- The process to do this is called Text Mining (TM)

Notes

What is Text Mining (TM)

- TM is an interdisciplinary set of procedures involving linguistics, computational statistics, and computer science
- It has been used in used in statistical and learning machine methods (Feinerer, Hornik, & Meyer, 2008), but its use has been almost inexisting in the social sciences
- Its main characteristic is that unstructured texts are the source of information.
- After removing the unnecessary information and ensuring that the messages hidden in the texts are preserved we have a dataset that perfectly resembles traditional datasets used in statistics
- A word cloud is one of the techniques used to represent TM methods.
- It accommodates hundreds or thousands of variables that are sized according to their frequency in the Texts

Notes

Word Cloud summarizing text from 1,272 texts

Figure 6. Community College #6, Arizona

Notes

- The data used from the word cloud can then be used as the input for a two-mode or affiliation network
- It also can guide researches to deep into categories or topics that are theoretically relevant, (*i.e.* asking questions about Financial Aid or asking help in math).
- The procedures used in our second session (Affiliation networks) can be used to identify actors who are posting about specific important categories, using these categories as communities.

Purpose To understand and replicate the Text Mining mechanisms required to successfully plot a word cloud in R.

Notes

Example of Text Mining and SNA

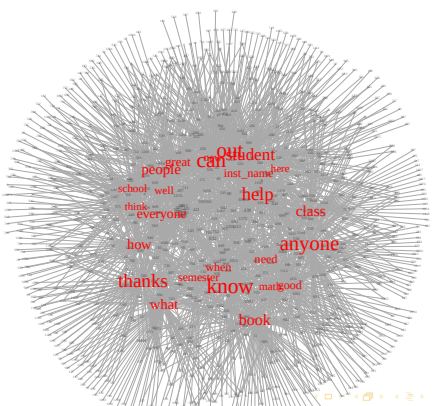

Figure 1: Example linking SNA with Text Mining

CGCC, Common words mentioned by users

Notes

Starting R...

Thank you!

Contact information

msgc@email.arizona.edu
Cecilia.Rios-Aguilar@cgu.edu

Notes

Notes

Notes