R for text analysis

Wouter van Atteveldt

April 2019

Introduction

- Wouter van Atteveldt
- Background in liberal arts, AI; now full-time social scientists
- Computational Communication Science & Political Communication
- VU Amsterdam
- I love teaching, computers, research, cooking, sailing, biking

How will this course work?

- Each session: short introduction, work on handout
- Work on
- Please interrupt me!
  - I'm an informal guy
  - But I sometimes speak too fast (and too Dutch!)

Schedule

`http://vanatteveldt.com/vienna-r-text-analysis/`

- Monday: Introduction to R & Rstudio
- Tuesday: R for Data Analysis
- Wednesday: Quantitative Text Analysis
- Thursday: Scraping and Cleaning Text
- Friday: Topic Models and Machine Learning

(but I'm flexible!)

Assessment

- Wednesday: Small exercise and in-class test (30%)
- Final paper (70%)
    - Solve a problem from your own research in R
    - Submit code, results, interpretation

Mid-week Exercise

- Data: US Election results at county level
  - `https://github.com/houstondatavis/data-jam-august-2016/tree/master/csv`
- Do county features (e.g. gender, education) predict election outcomes?
- More details will follow

What is R?

- Statistics package & programming language
- Open source, multi-platform, community driven
- Syntax mode only!
- Very easy to be part of the 'creators'
    - Do something useful, write package, submit to CRAN

Cathedral vs. Bazaar

R Packages

- Base R: data handling, simple statistics
- Tidyverse (Mon-Tue)
  - Uniform data handling
- Quanteda (Wed-Fri)
  - Quantitative text analysis
- Resources:
  - `https://www.rstudio.com/resources/cheatsheets/`
  - All packages have help functions
  - google "task view" + your task, e.g. time series

Rstudio

- Simple shell around R
- Makes it easier to use R
- Use projects to organize your files
- Use files and control+enter for writing code
- Use 'tab' for completion and options

Morning Session

- Work through hand-outs
  - Fun with Text
  - Getting started with R
- Bored?
  - Play around!
  - Try to get your own data into R
  - Start working on the Assignment

Tidyverse

- Collection of packages
- Today: `dplyr`
  - every function alters data in a small way
  - filter, select, rename, arrange, mutate
- Data cleaning as a series of function calls
  - Can use %>% pipe symbol to chain functions

Afternoon Session

- Work through hand-outs
  - Importing data (external tutorial)
  - Cleaning and filtering data
- Bored?
  - Load your own data into R
  - Play around with the tidyverse operations
  - Start working on the Assignment