Programming and Analyzing in R: Visualizing (and) Text

Wouter van Atteveldt http://vanatteveldt.com

June 2015

Topics

- Visualization
- Text analysis:
  - Querying AmCAT
  - Corpus analysis
  - Topic modeling

Recap

- Every variable has a type
  - Vectors of numeric, character, factor, ...
  - Data frame, list, matrix
- Select and assign data to columns:
  - data[rows, columns]
  - data$column
  - data$column[rows] = othervalue[rows]
- Merge and transform data
  - rbind, cbind
  - merge
  - cast and melt
- R as a programming language
  - Control execution of code
  - for, if, function

Built-in tools

- plot(.) plots data depending on object
  - plot(x=.., y=..)
  - plot(lm(d$y ~ d$x))
  - plot(d$y~d$x)
  - plot(d)
- hist(.) plots a histogram
- add extra information
  - lines, abline
  - abline(lm(..))

ggplot2

- Plots are composed of layers:
  - Mapping of data to aesthetics
    - x, y, colour, ..
  - geometry layers (line, point)
  - Add layers with +
- `ggplot(data, aes(.)) + geom_line(.) + ...`
- Layers 'inherit' data and mappings from base
  - can override (e.g. different y, colour, etc)

Interactive graphs

- qplot: a ggplot wrapper
- googleVis: interactive plots using Google's API
- rcharts: interactive plots with d3js
  - `http://rcharts.io/gallery/`
- dygraphs: interactive time series

Connecting to AmCAT

- `amcat.connect`
- `amcat.getarticlemeta`
- `amcat.hits`, `amcat.aggregate`

Term frequencies

- Document-term matrix
  - each row is document
  - each column is a term
  - cell is frequency of term
- Always stored as sparse matrix

Preprocessing and tokens

- Words contain lot of variation, noise
- Stemming, word selection
- NLP Processing can be useful
  - `amcat.gettokens`
  - `dtm.create`

Corpus comparison

- `term.statistics`
- `corpora.compare`
  - Compare speakers, sources, periods
  - Overrepresented words

Topic Modeling

- Cluster analysis of words
- Which words form 'topic'
- `lda.fit`
- `terms`
- `topics.per.document`

Conclusions

- R is a toolkit
- Most important are basics
  - How is data represented?
  - How do I get my data from A to B?
  - How do I figure out what's wrong?
- Powerful packages
  - Modeling
  - Visualizing
  - Text Analysis
- There's more to learn...