

Quanlify with Ease

Quantity and Quality

Andreas Blaette

April 13, 2020

Problem Statement

There is no lack of algorithms! But ...

Problem Statement

There is no lack of algorithms! But ...

- Acquisition of NLP techniques in ther social sciences & humanities

Problem Statement

There is no lack of algorithms! But ...

- Acquisition of NLP techniques in ther social sciences & humanities
- Tools of processing that scale well

Problem Statement

There is no lack of algorithms! But ...

- Acquisition of NLP techniques in ther social sciences & humanities
- Tools of processing that scale well
- Availability of data for replication

Problem Statement

There is no lack of algorithms! But ...

- Acquisition of NLP techniques in ther social sciences & humanities
- Tools of processing that scale well
- Availability of data for replication
- Reproducibility of data (getting FAIRER)

Problem Statement

There is no lack of algorithms! But ...

- Acquisition of NLP techniques in ther social sciences & humanities
- Tools of processing that scale well
- Availability of data for replication
- Reproducibility of data (getting FAIRER)
- Integration of quantitative and qualitative approaches to text

Focus of the presentation

- Combining close and distant and close reading [Moretti2013] is an unfulfilled promise: Software often inhibits combining both perspectives. How to implement workflows for coding and annotating textual data?
- Scenarios:
 - flexdashboards:
 - Shiny Modules:
 - Gadgets: Interactive graph annotation as an approach to generate intersubjectively shared interpretations/understandings of discourse patterns.



A methodological divide?

Quantity

- Natural Language Processing (NLP)
- Big Data
- Data Mining
- Text Mining
- Machine Learning (ML)
- Artificial Intelligence (KI)
- Text as Data

A methodological divide?

Quantity

- Natural Language Processing (NLP)
- Big Data
- Data Mining
- Text Mining
- Machine Learning (ML)
- Artificial Intelligence (KI)
- Text as Data

Quality

- eHumanities / Digital Humanities
- Corpus Linguistics
- Computational Linguistics
- Interpretation

A methodological divide?

Quantity

- Natural Language Processing (NLP)
- Big Data
- Data Mining
- Text Mining
- Machine Learning (ML)
- Artificial Intelligence (KI)
- Text as Data

Quality

- eHumanities / Digital Humanities
- Corpus Linguistics
- Computational Linguistics
- Interpretation

We Shall Overcome ... By Quantification

PolMine Project

Data and Code for Corpus Analysis

The PolMine Project | www.polmine.de

- **Research**

On migration & integration policy: MigTex, MIDEM, PopParl

The PolMine Project | www.polmine.de

- **Research**

On migration & integration policy: MigTex, MIDEM, PopParl

- **Data**

Corpora of plenary protocols, Newspaper articles, ...

The PolMine Project | www.polmine.de

- **Research**

On migration & integration policy: MigTex, MIDEM, PopParl

- **Data**

Corpora of plenary protocols, Newspaper articles, ...

- **Code**

open source R packages for text analysis, at CRAN & GitHub

The PolMine Project | www.polmine.de

- **Research**

On migration & integration policy: MigTex, MIDEM, PopParl

- **Data**

Corpora of plenary protocols, Newspaper articles, ...

- **Code**

open source R packages for text analysis, at CRAN & GitHub

- **Tutorials**

Using Corpora in Social Science Research / UCSSR

The PolMine Project | www.polmine.de

- **Research**

On migration & integration policy: MigTex, MIDEM, PopParl

- **Data**

Corpora of plenary protocols, Newspaper articles, ...

- **Code**

open source R packages for text analysis, at CRAN & GitHub

- **Tutorials**

Using Corpora in Social Science Research / UCSSR

- **Centre**

CLARIN Centre category C, prospectively part of NFDI

The PolMine Project | www.polmine.de

- **Research**

On migration & integration policy: MigTex, MIDEM, PopParl

- **Data**

Corpora of plenary protocols, Newspaper articles, ...

- **Code**

open source R packages for text analysis, at CRAN & GitHub

- **Tutorials**

Using Corpora in Social Science Research / UCSSR

- **Centre**




CLARIN Centre category C, prospectively part of NFDI

Learn more: www.polmine.de

The PolMine Project R Packages




The PolMine Project R Packages

The core package family living at CRAN

- *polmineR*: elementary vocabulary for corpus analysis 
- *cwbtools*: tools to create and manage CWB indexed corpora 
- *RcppCWB*: wrapper for the Corpus Workbench (using C++/Rcpp) 

The PolMine Project R Packages

The core package family living at CRAN

- *polmineR*: elementary vocabulary for corpus analysis 
- *cwbtools*: tools to create and manage CWB indexed corpora 
- *RcppCWB*: wrapper for the Corpus Workbench (using C++/Rcpp) 

A toolchain for corpus preparation

- *frapp*: Framework for Parsing Plenary Protocols
- *bignlp*: Fast NLP processing for big corpora
- *biglda*: Fast LDA topic modelling
- *ctk*: corpus toolkit (misc functionality for corpus preparation)

Data

Corpora of Plenary Protocols

- *GermaParl*: German Bundestag, regional parliaments DOI 10.5281/zenodo.3742113
- *UNGA* DOI 10.5281/zenodo.3748858
- *ParisParl* / *AustroParl* / *TweedeKamer*:
- *MigParl*

Other Corpora

- *MigPress*

aasdfasdf

- Prerequisites: Any kind of computer, installation of R/RStudio

```
install.packages("polmineR")  
install.packages("GermaParl") # the downloaded package includes a small sample dataset  
GermaParl::germaparl_download_corpus() # get the full corpus
```

asdf

aasdfasdf

- Prerequisites: Any kind of computer, installation of R/RStudio

```
install.packages("polmineR")  
install.packages("GermaParl") # the downloaded package includes a small sample dataset  
GermaParl::germaparl_download_corpus() # get the full corpus
```

asdf

```
library(polmineR)  
kwic("GERMAPARL", query = "Integration") # activate the corpora in the GermaParl package, i
```


aasdfasdf

- Prerequisites: Any kind of computer, installation of R/RStudio

```
install.packages("polmineR")  
install.packages("GermaParl") # the downloaded package includes a small sample dataset  
GermaParl::germaparl_download_corpus() # get the full corpus
```

asdf

```
library(polmineR)  
kwic("GERMAPARL", query = "Integration") # activate the corpora in the GermaParl package, i
```

```
drat::addRepo("polmine")  
install.packages("UNGA")  
UNGA::unga_download_corpus()
```

Theory is Code

Ideas behind "quantification"

From text to numbers {.smaller}

- **from computer-assisted content analysis to "text as data"**
scaling party positions as a driver (wordscore and wordfish)
- **joyful blasphemy against reading ...**
"[...]because it treats words simply as data rather than requiring any knowledge of their meaning as used in the text, our word scoring method works irrespective of the language in which the texts are written. In other words, while our method is designed to analyse the content of a text, it is not necessary for an analyst using the technique to understand, or even read, the texts to which the technique is applied. The primary advantage of this feature is that the technique can be applied to texts in any language." (Laver, Benoit & Garry 2003)
- **common methods and applications**
 - sentiment analyses
 - topic modelling (unsupervised learning)
 - classification (cp. Comparative Agendas Project / CAP)
- **"Validate, validate, validate" (Grimmer et al. 2013)**
An (almost) unheard plea

The idea of "distant reading" {.smaller}

„[...] the trouble with close reading [...] is that it necessarily depends on an extremely small canon. [...] you invest in individual texts so much only if you think that very few of them really matter. [...] if you want to look beyond the canon [...], close reading will not do it. [...] At the bottom it's a theological exercise – very solemn treatment of very few texts taken very seriously – whereas what we really need is a little pact with the devil: we know how to read texts, so now let's learn how not to read them. Distant reading, where distance, let me repeat is, is a condition of knowledge. It allows you to focus on units that are much smaller or much larger than the text: devices, themes, types – or genres and systems. And if, between the very small and the very large, the text itself disappears, well, this is one of the cases where one can justifiably say, Less is more. If we want to understand the system in its entirety, we must accept losing something. We always pay a price for theoretical knowledge; concepts are abstract, are poor. But it's precisely this poverty that makes it possible to handle them, and therefore to know. This is why less is actually more.“ (Moretti [2000] 2013: 49)

Why and how text matters {.smaller}

- **The social sciences and the "linguistic turn"**
 - An evolving theoretical movement
 - analysing discourse
 - analysing frames
 - analysing narratives
- **Methodological development**
 - persistence of paper & pencil-analyses
 - computer-assisted qualitative analysis (QDA, see MAXQDA, Atlas.ti)
 - digital humanities / eHumanities
 - Visual analytics
- **Varieties of "distant reading" (Moretti 2000)**
 - "blended reading" (Stulpe, Lemke 2015)
 - "scalable reading" (Weitin 2017)

polmineR - a basic vocabulary {.smaller}

- **Corpora and subcorpora**

- corpus objects: *corpus()*
- subsetting corpora: *partition()* / *subset()*

- **Quantification**

- counting: *hits()*, *count()*, *dispersion()* (and *size()*)
- cooccurrences: *cooccurrences()*, *Cooccurrences()*
- feature extraction: *features()*
- term-document-matrices: *as.sparseMatrix()*, *as.TermDocumentMatrix()*

- **Qualitative analysis**

- Keywords-in-context/concordances: *kwic()*
- full text (of a subcorpus): *get_token_stream()*, *as.markdown()*, *as.html()*, *read()*

The Quanlification Familiy

- *annolite*: light-weight full text display and annotation tool
 - *topicanalysis*: integrate quantitative/qualitative approaches to topic models
 - *gradget*: graph annotation widget
 - *fulltext*: htmlwideget
 - *quanlify*: ddd
-

class: inverse

Scenario I

Scenario II

Flexdashboards for Digging into LDA Topic Models

Problem Statement

asdfasdf

Scenario III

Graph Annotation Widgets: Gradgets

Problem: The elusive merit of cooccurrence graphs

- Popularity of cooccurrence graphs [@Rhizome2013; @2016TMid].
- Suggestive visualisations ... But are these interpretations sound and do they meet standards of intersubjectivity?
 - The graph layout depends heavily on filter decisions.
 - Filtering is necessary, but there are difficulties to justify filter decisions.
 - Graph visualisation implies many possibilities to provide extra information, but there are perils of information overload.
 - If we try to omit filter decisions, we run into the problem of overwhelming complexity of large graphs.
 - How to handle the complexity and create the foundations for intersubjectivity?

Gradgets

So 'gradgets' are the solution suggested here. The links to the following three gradgets offer a visualisation that is interactive in a double sense: