Introduction to Text Analysis

Computational Text Analysis

Theresa Gessler

University of Zurich | http://theresagessler.eu/ | @th_ges

2022-05-09

Program

- Who?
- Why?
  - text analysis
  - web data
- Principles of Computational Text Analysis
- What?
  - overview of the course
- Homework

Who

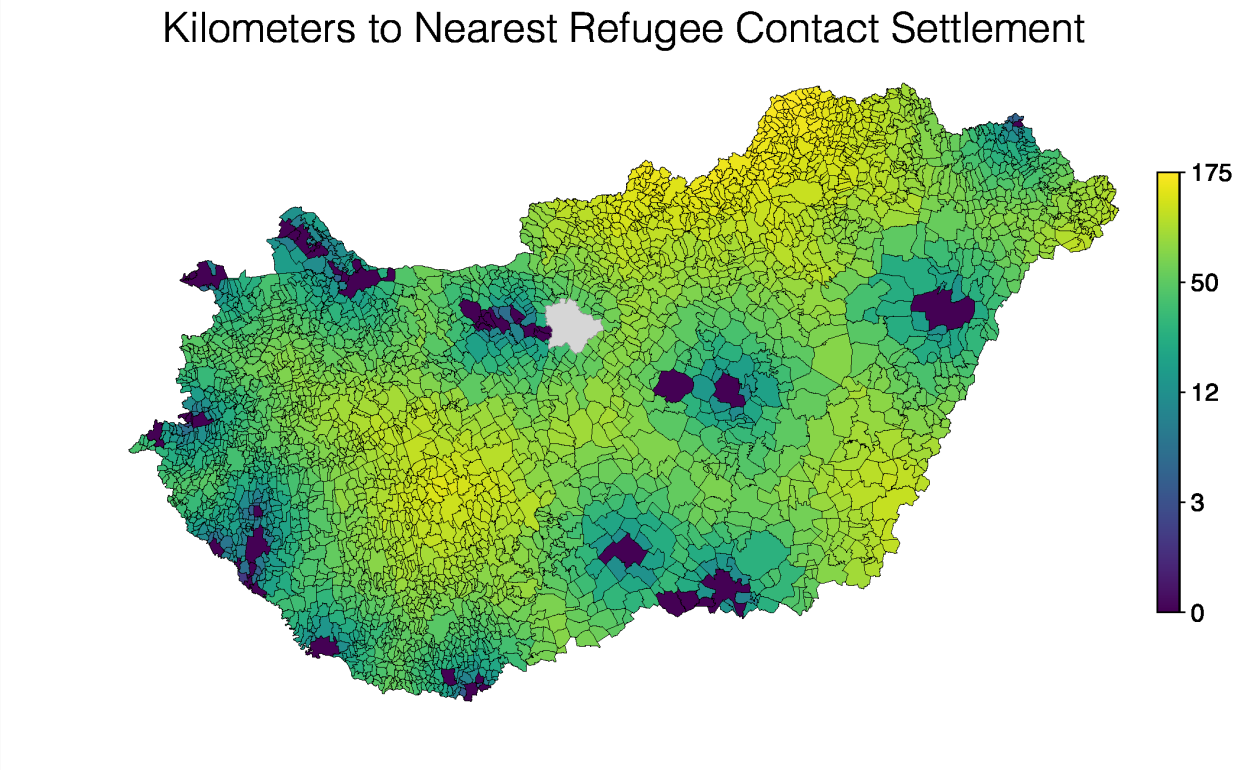
Who: Dr. Theresa Gessler

- **EUI PhD (2019)**
- **Postdoc** at the Digital Democracy Lab / Department of Political Science at University of Zurich
- Co-organizer of UZH Computational Methods Working Group and Summer School for Women in Political Methodology

- **reach me**
  - gessler@ipz.uzh.ch
  - www.theresagessler.eu | @th_ges
- my research: **immigration** | **political parties** | **(digital) democracy** | **gender**

Who: My Research

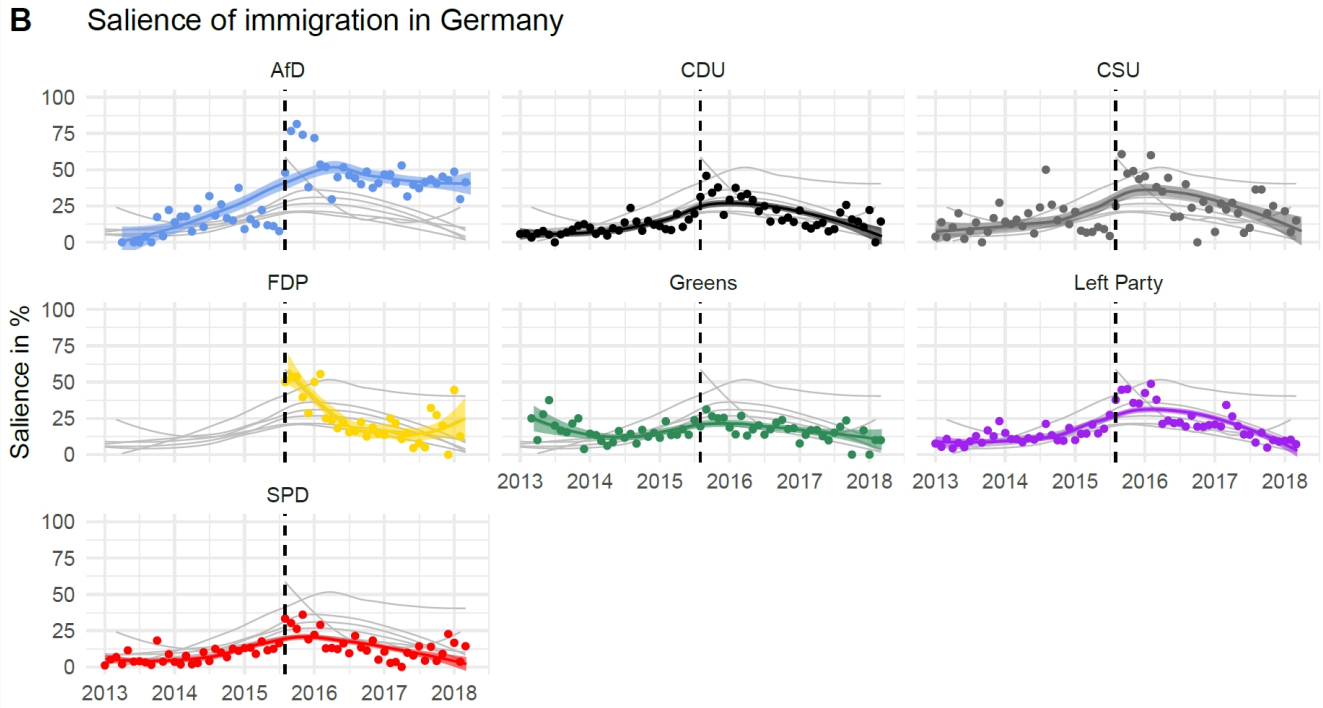
Immigration



Gessler, Theresa, Gergő Tóth, und Johannes Wachs. „No Country for Asylum Seekers? How Short-Term Exposure to Refugees Influences Attitudes and Voting Behavior in Hungary“. Political Behavior, 9. Februar 2021.

Who: My Research

Political Parties



Gessler, Theresa, und Sophia Hunger. „How the Refugee Crisis and Radical Right Parties Shape Party Competition on Immigration“. Political Science Research and Methods, 15. November 2021, 1–21.

Who: My Research

Agenda-setting

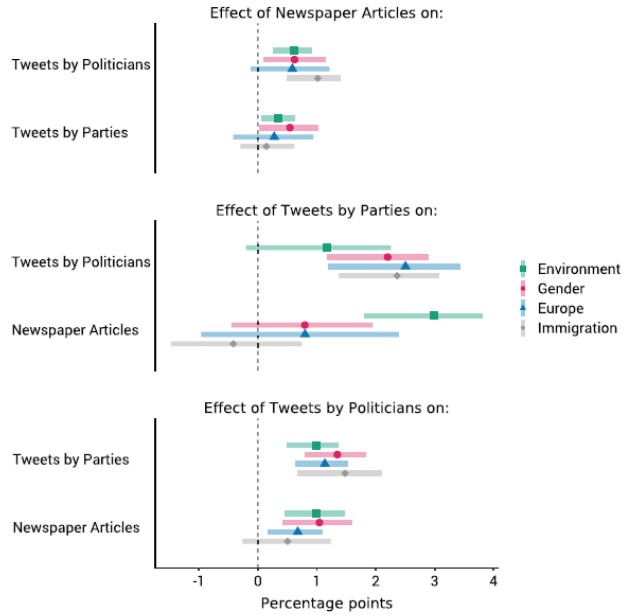


Figure 3. Agenda responsiveness of parties, politicians, and newspapers. Bars denote 95% confidence intervals.

Gilardi F., Gessler T., Kubli M., Müller S. (2022): Social Media and Political Agenda Setting. Political Communication 39 (1): 39-60.

Who: My Research

Digital democracy / Gender

Angela Merkel

Merkel ist eine Weiterleitung auf diesen Artikel. Weitere Bedeutungen sind unter Merkel (Begriffsklärung) aufgeführt.

Angela[1] **Dorothea Merkel** (* 17. Juli 1954 in Hamburg als *Angela Dorothea Kasner*) ist eine deutsche Politikerin (CDU). Sie ist seit dem 22. November 2005 Bundeskanzlerin der Bundesrepublik Deutschland. Vom 10. April 2000 bis zum 7. Dezember 2018 war sie CDU-Bundesvorsitzende. Im Oktober 2018 erklärte sie, sich spätestens mit Ablauf der Legislaturperiode 2021 aus der Politik zurückzuziehen.

Merkel wuchs in der DDR auf und war dort als Physikerin am Zentralinstitut für Physikalische Chemie tätig. Bei der Bundestagswahl am 2. Dezember 1990 errang sie erstmals ein Bundestagsmandat. Bei den folgenden sieben Bundestagswahlen wurde sie in ihrem Wahlkreis in Vorpommern direkt gewählt.[2] Von 1991 bis 1994 war Merkel Bundesministerin für Frauen und Jugend im Kabinett Kohl IV und von 1994 bis 1998 Bundesministerin für Umwelt, Naturschutz und Reaktorsicherheit im Kabinett Kohl V. 1998 bis zu ihrer Wahl zur Bundesvorsitzenden der Partei amtierte sie als Generalsekretärin der CDU.

Nach dem knappen Sieg der Unionsparteien bei der vorgezogenen Bundestagswahl 2005 löste Merkel Gerhard Schröder als Bundeskanzler ab und führte zunächst eine große Koalition mit der SPD bis 2009 (Kabinett Merkel I). Nach der Bundestagswahl 2009 ging sie mit der FDP eine schwarz-gelbe Koalition ein (Kabinett Merkel II), der 2013 eine erneute große Koalition folgte, die auch nach der Bundestagswahl 2017 fortgesetzt wird (Kabinett Merkel III und IV).

Angela Merkel (2019)

Inhaltsverzeichnis [Verbergen]

Gessler T. But is she married? Gender Bias and Users' Gendered Interest in Politicians on Wikipedia. Manuscript.

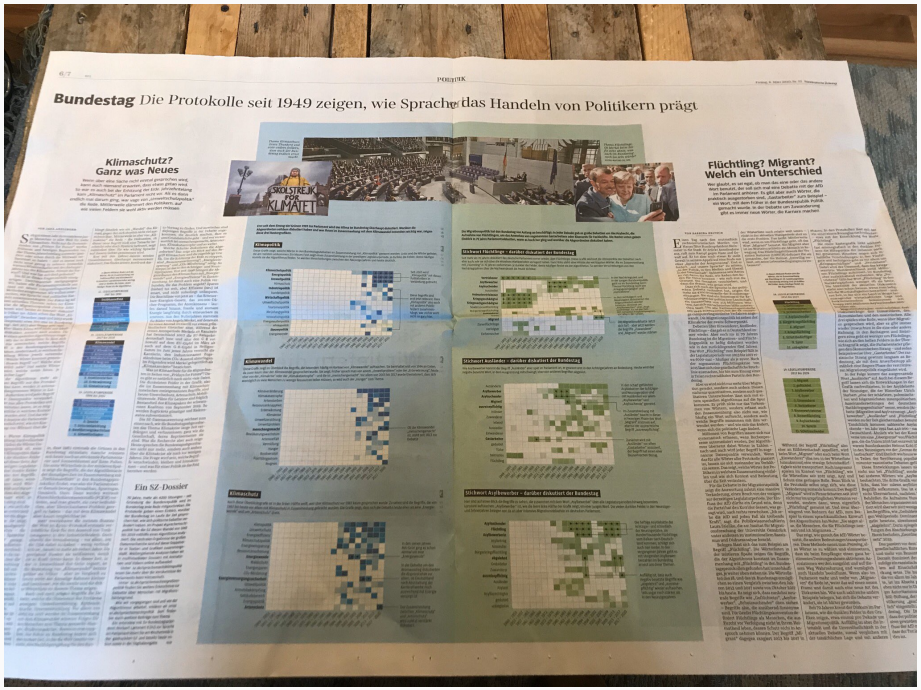
Who: Your turn

- name
- research interests
- why are you taking this course?

Why

Why Text Analysis?

- tracks changing discourses over time
  - here: in parliamentary transcripts



Source: Süddeutsche Zeitung, see: Das gehetzte Parlament, Wie der Bundestag den Klimawandel verdrängte and So haben wir den Bundestag ausgerechnet

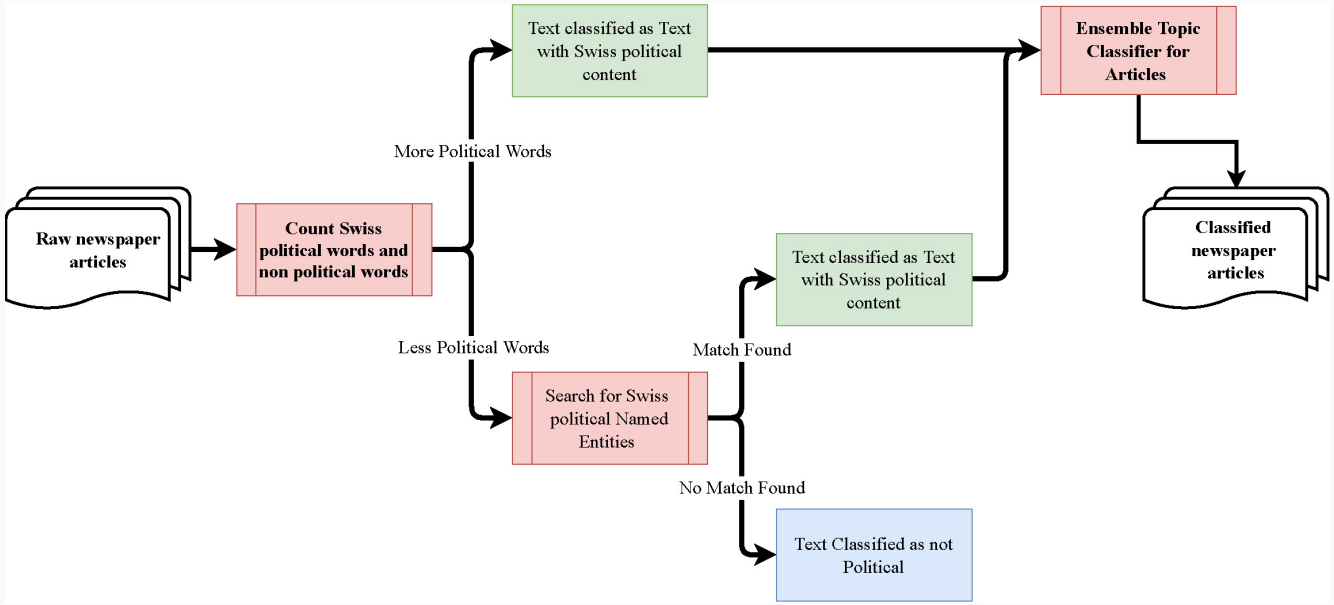
Why Text Analysis?

- provides context to concepts
  - here: to ideological labels, based on open-ended survey questions

Source: Bauer, Paul C., Pablo Barberá, Kathrin Ackermann, and Aaron Venetz. "Is the Left-Right Scale a Valid Measure of Ideology?" Political Behavior 39, no. 3 (2017): 553–83.

Why Text Analysis?

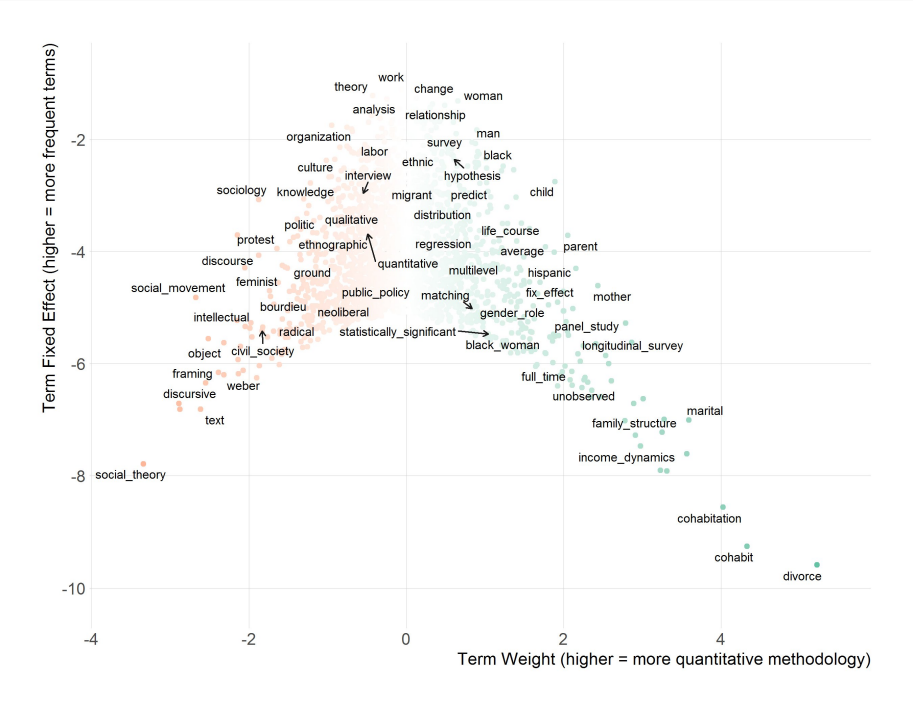
- finds the needle in the haystack
  - here: Media reports on different political topics in Swiss election campain



Source: Gilardi, Fabrizio, Theresa Gessler, Mael Kubli and Stefan Müller. "Social Media and Political Agenda Setting." Work in Progress, 2020.

Why Text Analysis?

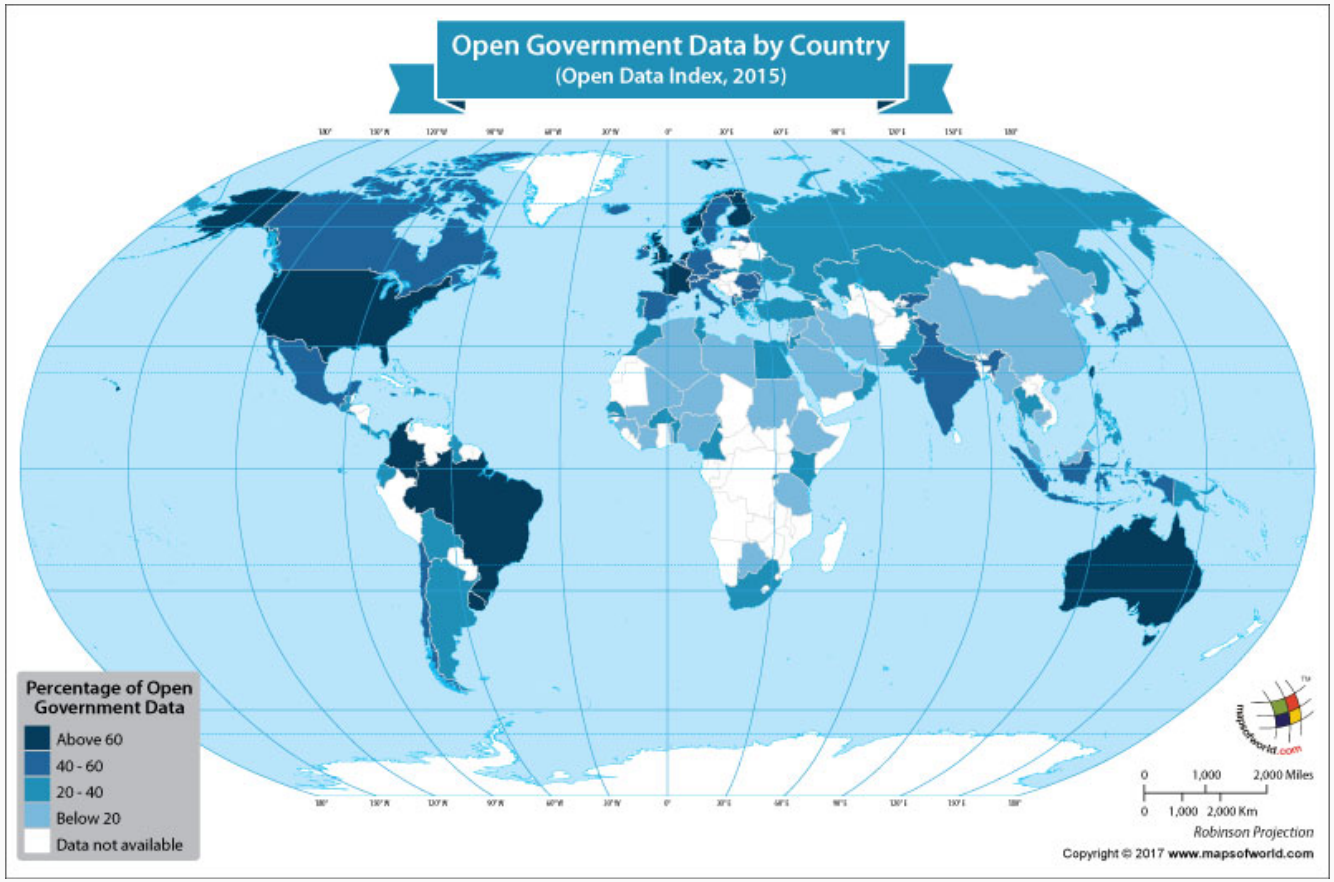
- finds order / dimensionality in vast masses of text
  - here: automated text analysis for 8,737 abstracts of papers published between 1995 and 2017



Source: Schwemmer, Carsten, and Oliver Wieczorek. "The Methodological Divide of Sociology: Evidence from Two Decades of Journal Publications." Sociology 54, no. 1 (2020): 3–21.

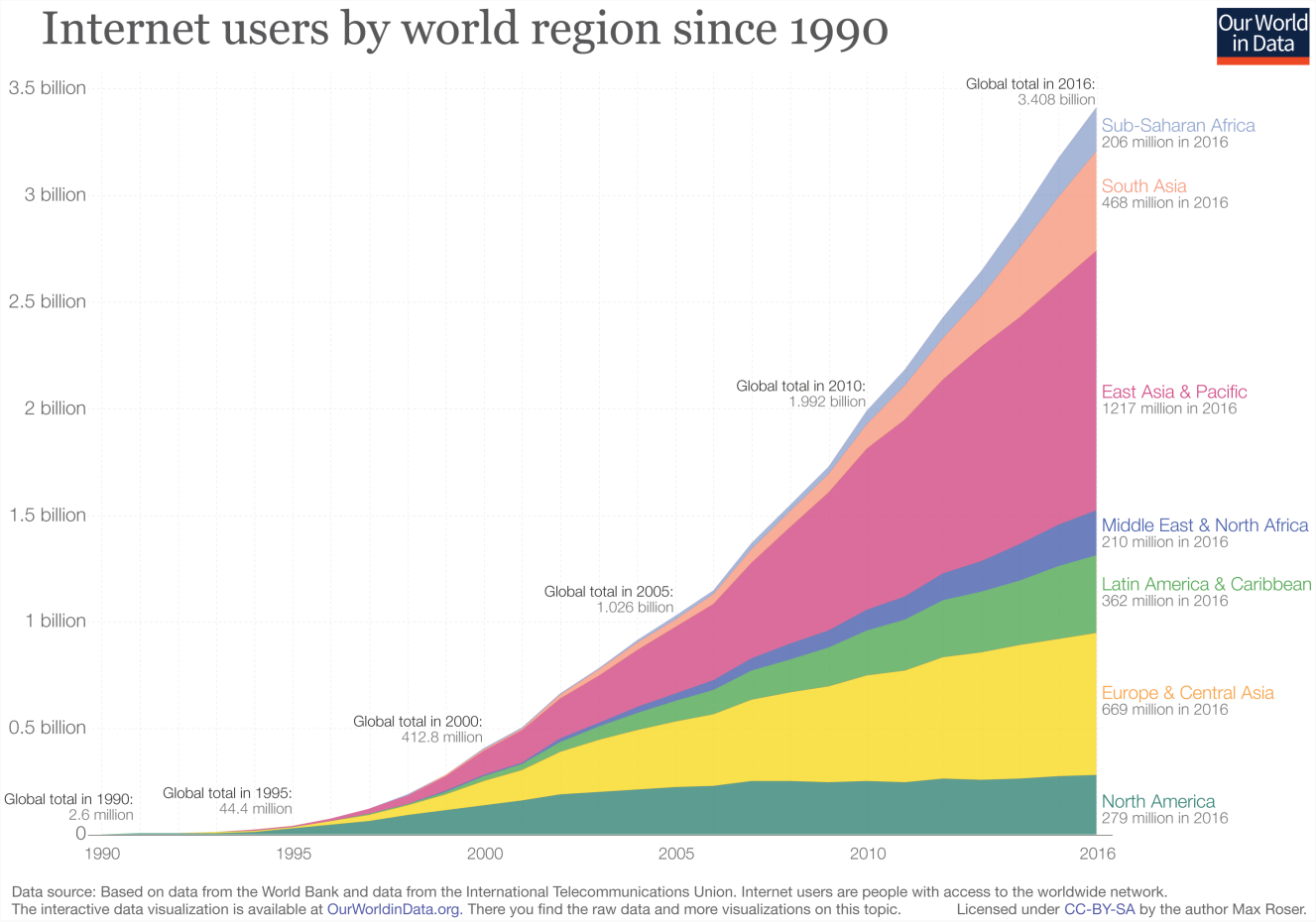
Why Online Data?

- increasing amount of public data online ('open government')



Why Online Data?

- increasing amount of people use the internet

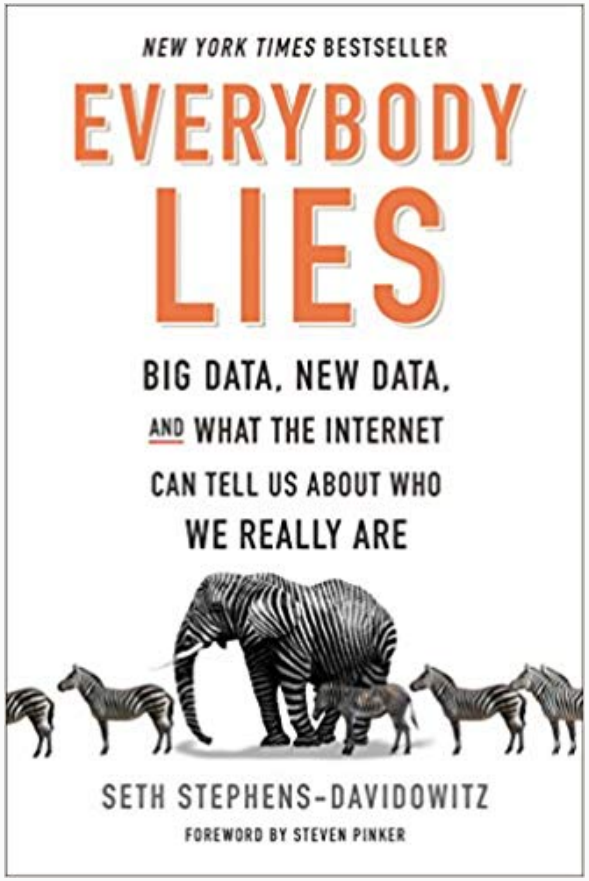


Why Online Data?

- increasing amount of politics happens online

Why Online Data?

- we share everything online



Computational Text Analysis

Computational Text Analysis

- traditional empirical work in political and social science
  - **quantitative methods** with limited understanding of (unstructured) text
  - **qualitative methods** with close analysis of small text collections

↑

↓

- masses of available text
  - e.g. by governments, media, organizations, laws, court decisions, speeches, ...
  - digitalization of existing text collections

→ **untapped potential of interesting (new) data!**

Computational Text Analysis

The spectrum

manual / hermeneutic analysis of content ↔ automated analysis of content

→ we focus on **automated analysis**, with varying degrees of human input

Definitions

- Systematic, objective, quantitative analysis of message characteristics (Neuendorf 2002, *The Content Analysis Guidebook*, 1)
- A variant of content analysis that is expressly quantititative, not just in terms of representing textual content numerically but also in analyzing it, typically using computation and statistical methods. (text analysis course by Ken Benoit)
- many related methods: content analysis, text analysis, text mining, natural language processing, text as data, ...

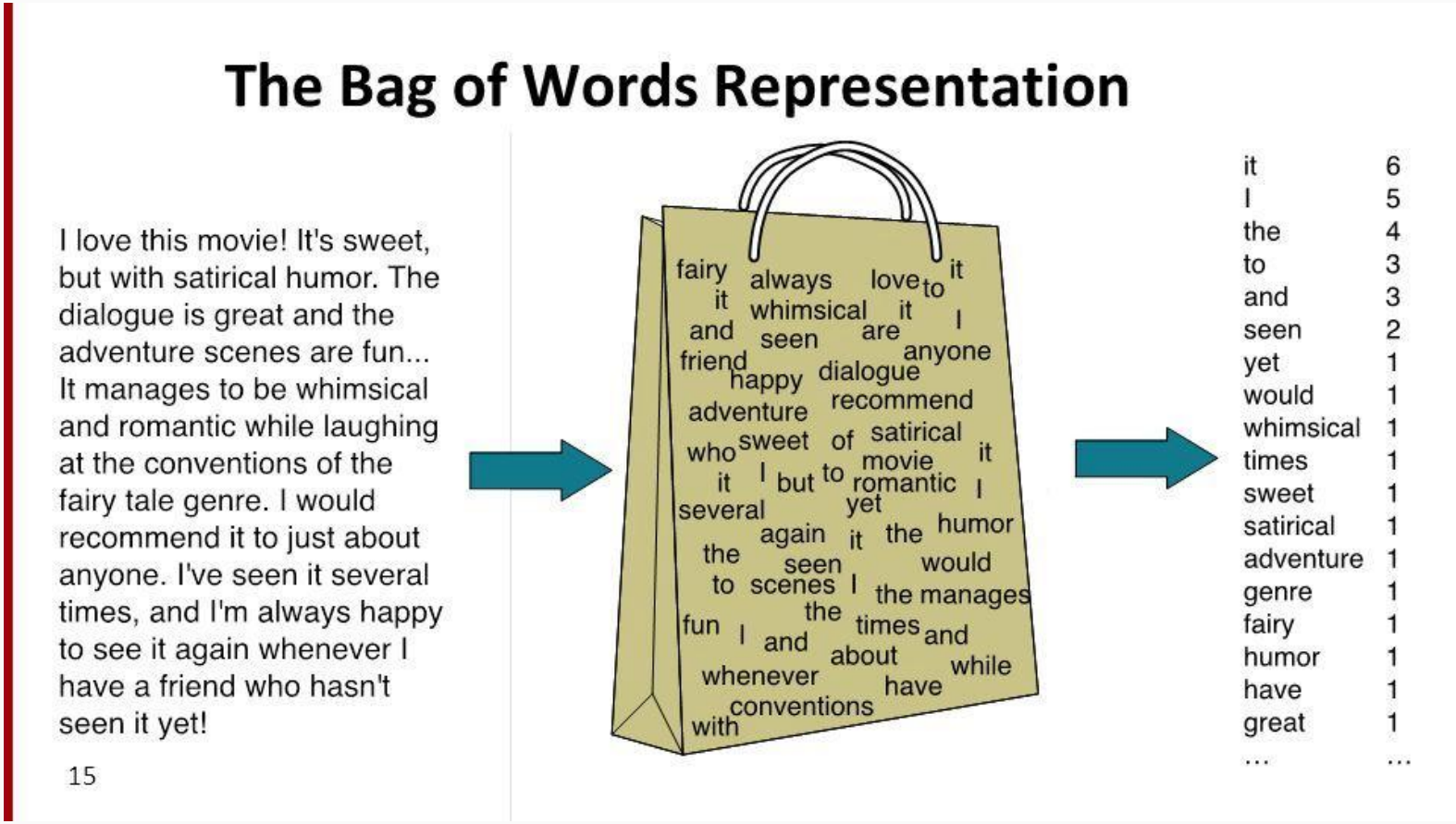
Basic assumptions

When doing quantitative text analysis, we assume...

- ...That texts represent an **observable implication** of some **underlying characteristic** of interest (an attribute of the author, the subject, ...)
- ...That texts can be represented through extracting their **features**, e.g. words
- ...That we can analyze the **frequency of features** (as a *document-feature matrix*) with quantitative methods to measure these underlying characteristics

Computational Text Analysis

The 'bag of words' assumption



Source: programmersought.com

The 'bag of words' assumption

A text

```
course ← "Over the past years, the availability of new data through the
digitalization of legal, political, journalistic corpora as well as the growth of
online sources has allowed researchers to answer new research questions across
political science.

The aim of this course is to introduce students to the quantitative analysis of
textual data. We will cover both applications in recent empirical research and the
implementation of text analysis techniques through hands-on experiences using the R
statistical programming language.

The course will cover the collection of text data with webscraping techniques, text
preprocessing, dictionaries and descriptive analysis of texts, as well as supervised
and unsupervised learning methods to classify the content of text corpora."
```

The 'bag of words' assumption

Its features

```
tokens(course)
```

```
## Tokens consisting of 1 document.
## text1 :
##  [1] "Over"         "the"          "past"         "years"        ","
##  [6] "the"          "availability" "of"           "new"          "data"
## [11] "through"      "the"
## [ ... and 110 more ]
```

The 'bag of words' assumption

Their frequency

```
dfm(course)
```

```
## Document-feature matrix of: 1 document, 71 features (0.00% sparse) and 0 docvars.
##        features
## docs    over the past years , availability of new data through
##   text1    1  11    1     1 6            1  9   2    3       2
## [ reached max_nfeat ... 61 more features ]
```

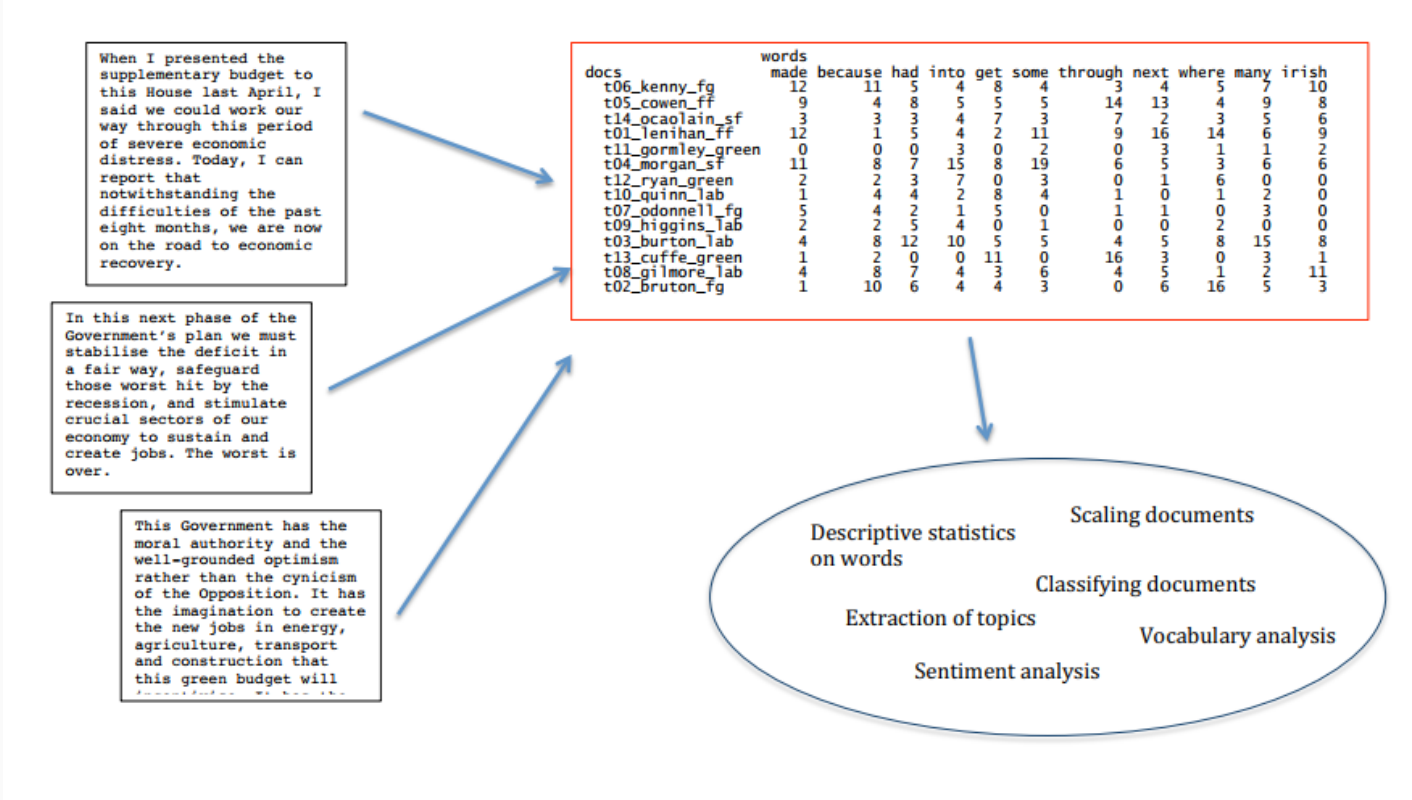
Their sorted frequency

```
dfm(course,remove_punct=T) %>% topfeatures()
```

```
##      the       of       as       to     text     data analysis      and      new
##       11        9        4        4        4        3        3        3        2
##  through
##        2
```

Computational Text Analysis

The 'bag of words' assumption



Source: Slapin, J. B., and S.-O. Proksch (2008). A Scaling Model for Estimating Time-Series Party Positions from Texts." American Journal of Political Science 52 (3): 705-22.

Computational Text Analysis

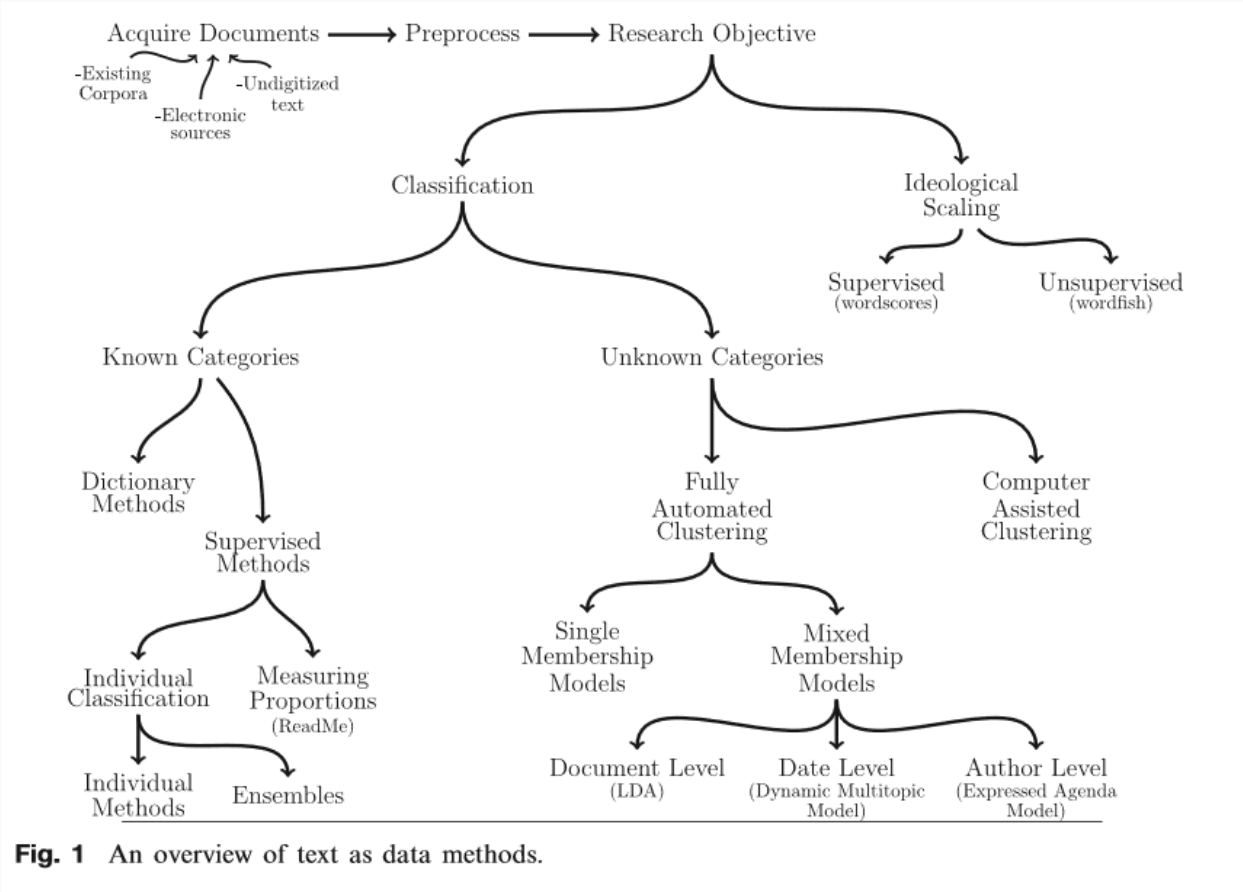


Fig. 1 An overview of text as data methods.

Grimmer, J. and B. M. Stewart (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political Analysis 21, 267-297.

Overview

Overview

Sessions

10/05 Introduction to Text Analysis

10/05 Descriptive Analyses, Dictionaries

11/05 Supervised Learning Methods

11/05 Unsupervised Learning Methods

12/05 Advanced Text Analysis Methods

12/05 Webscraping & the Text Analysis Pipeline

Overview

10/05 Descriptive Analyses, Dictionaries

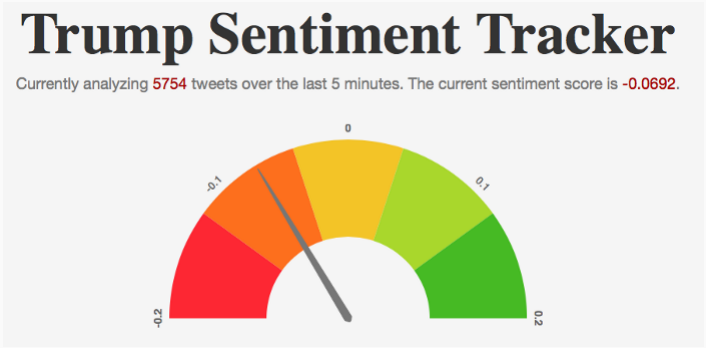
- How to prepare text for analysis
- descriptive overviews of text corpora
- dictionary analyses of word frequencies

```
## Keyword-in-context with 6 matches.
##    [167_Trump, 9]            to you, the | country | would have been left
##  [167_Trump, 150] should have closed our | country | . Wait a minute
##    [169_Trump, 9] should have closed our | country | because you thought it
##   [215_Trump, 36]     the history of our | country | . And by the
##    [226_Trump, 9]      to shut down this | country | and I want to
##   [228_Trump, 29]       to shut down the | country | . We just went
```

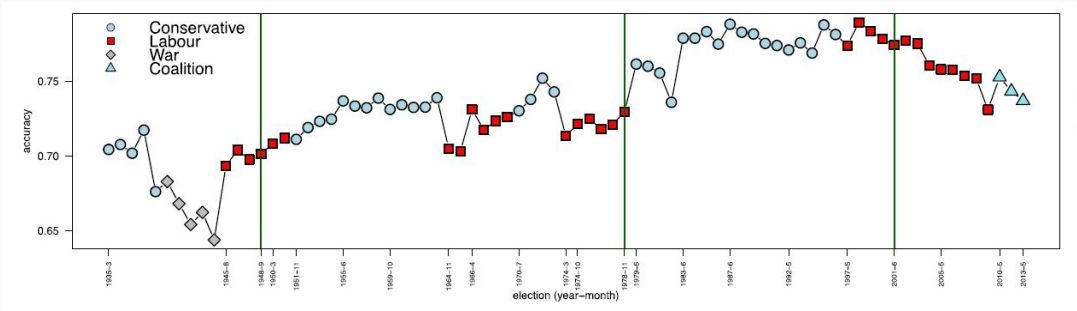
Overview

11/05 Supervised Learning Methods

- classifying text into **known categories**
  - e.g. sentiments
  - e.g. topics
- discrete and continuous classifications
- substantive uses of classification uncertainty



Conor Dewey's Trump Sentiment Tracker

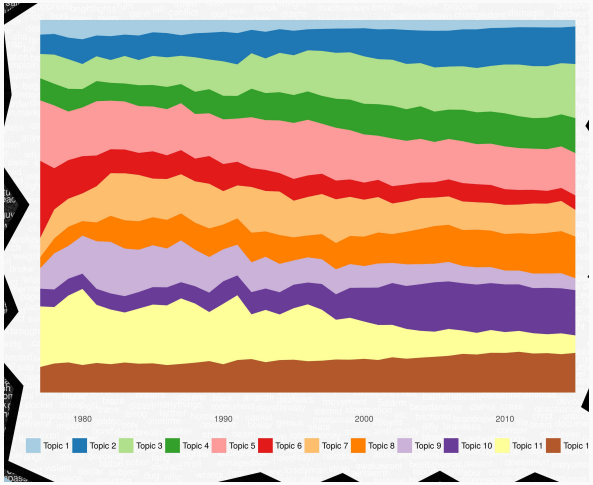


e.g. Parliamentary Polarization estimates by Andrew Peterson and Arthur Spirling

Overview

11/05 Unsupervised Learning Methods

- classifying text into **unknown categories**
- methods and choices
  - topic models
  - clustering methods
- introducing elements of supervision



e.g. *Martin Mölder, Federico Vegetti. "What do they talk about when they play punk? A quantitative analysis of punk rock lyrics from 1977 to 2015."*

Overview

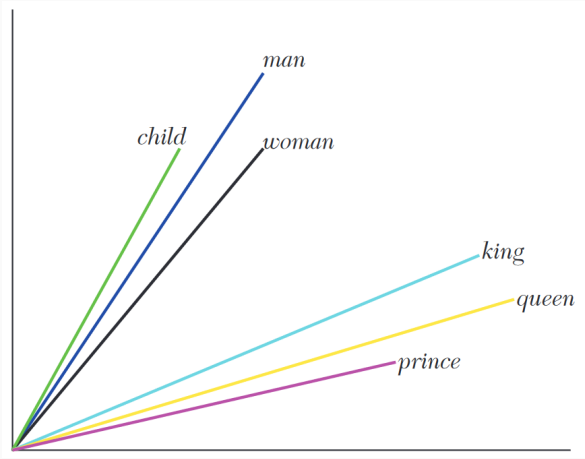
12/05 Advanced Text Analysis Methods

Text Analysis is a dynamic field

- methods we have not covered
- word embeddings
- transformer models
- ...

→ What do I need to know about this?

→ What should I use?



Overview

12/05 Webscraping & the Text Analysis Pipeline

- gathering data from webpages
  - simple HTML pages: texts, tables
- an overview of more advanced techniques
  - Selenium
  - APIs
- small project on US presidential speeches → your chance to do a small analysis

```
<div class="navbar navbar-default  navbar-fixed-top" ro
  <div class="container">
    <div class="navbar-header">
      <button type="button" class="navbar-toggle collap
        <span class="icon-bar"></span>
        <span class="icon-bar"></span>
        <span class="icon-bar"></span>
      </button>
      <a class="navbar-brand" href="index.html">EUI Com
    </div>
    <div id="navbar" class="navbar-collapse collapse">
      <ul class="nav navbar-nav">
        <li>
  <a href="index.html">Overview</a>
</li>
<li>
  <a href="readings.html">Readings</a>
</li>
<li>
  <a href="code.html">Code &amp;amp; Slides</a>
</li>
```

Overview

How this course works

- **learning by doing**
  - **'lecture'** to introduce method
  - **joint exercises** in class
  - **additional exercises for individual work** / in lab session
- **asking questions, whenever you have them!**

Material

- **Course documentation** with syllabus, slides, exercises & reading list: http://theresagessler.eu/eui_cta/
  - *updates throughout the course*
- RStudio Cheat Sheets
- R for Data Science
- Quanteda Tutorials

Overview

Packages that we will use

- scraping: **rvest** (potentially: **httr**, **RSelenium**)
- text analysis: **quanteda**, **quanteda.textstats**, **quanteda.textplots**, **quanteda.textmodels**, **caret**, **stm** (pontentially: stringr, readtext)
- data wrangling and visualization: **tidyverse** (especially **dplyr**, **tidyr**, **lubridate** and **ggplot2**)
- creating documents and reports: **rmarkdown** and **knitr**

```
first_packages ← c("tidyverse","rvest","quanteda","quanteda.textstats",
"quanteda.textplots","quanteda.textmodels","rmarkdown","knitr")
install.packages(first_packages)
```

Recommended Setup

- create a folder for this course
  - decide on a structure: slides, exercises, example code, ...
- create an RStudio Project in that folder
- Homeworks, Exercises etc. in RMarkdown
- Find your way to write good code - some inspiration:
  - tidyverse style guide
  - SoftwareCarpentry
  - R for Reproducible Research
  - Code and Data for the Social Sciences: A Practitioner's Guide

→ **RMarkdown Exercise**

→ **Text Analysis Process Exercise** (*optional*)

Homework

In R

- *installing packages: you know the drill!*
- **familiarize yourself with for loops** (exercise file on Slides and Code page)

HTML (until Thursday)

No need to be able to write HTML pages but scraping is much easier with some concepts

- familiarize yourself with the structure of HTML
  - HTML Tags / Elements
  - HTML Attributes
  - how a hyperlink works
  - ideally: sections HTML Basic to HTML Paragraphs plus HTML Links on W3schools
- install a suitable web browser

After the break

After the break

What we'll cover

- **Pre-processing** Text Data
  - from text to data
- **creating and applying dictionaries** to measure latent concepts
- **packages**: `quanteda` & its siblings

Thank you - see you in a few minutes!