university of groningen

Quantitative Text Analysis – Essex Summer School

Introduction to text as data

dr. Martijn Schoonvelde

University of Groningen

Today's plan

- Why automated text analysis?
- Setting up the course
- Steps in a canonical text analysis project

Who am I?

Assistant professor in European Politics & Society at University of Groningen

- Work on automated text analysis of rhetoric of political leaders
  - Understanding political rhetoric. Political logic? Personality? Responsiveness?
  - Does political rhetoric drive or follow public opinion?
- Also interested in non-verbal communication (images as data, emotions and voice pitch)
- Email: martijn.schoonvelde@rug.nl; Twitter: @hjms

TA

Second-year PhD student & Assistant lecturer at the Department of Government, University of Essex

- Research interests: Civil wars, terrorism, political violence
- Contact: m.e.arslan@essex.ac.uk

Contact

- Ask questions, come talk to us – happy to help / set up a Zoom meeting
  - Mehmet is the first person to ask your questions – and I will follow after that.
- Use the Slack workspace for this module to communicate with each other: `essqta2022.slack.com`
  - Mehmet and I will both regularly check the workspace and comment on any issues you raise.
- All materials (slides / code scripts / etc) available at `https://github.com/hjmschoonvelde/essex_summer_school_qta`

Who are you?

- Background, interests
- What do you expect from this course?
- What is your favorite book ever?

Quantitative text analysis

The New York Times

Opinion

I Am Part of the Resistance Inside the Trump Administration

I work for the president but like-minded colleagues and I have vowed to thwart parts of his agenda and his worst inclinations.

Sept. 5, 2018

13133

Leer en español · 阅读简体中文版 · 閱讀繁體中文版 · 한국어로 읽기 · 日本語で読む

Written by **"Anonymous"**

Quantitative text analysis

The New York Times

Opinion

I Am Part of the Resistance Inside the Trump Administration

I work for the president but like-minded colleagues and I have vowed to thwart parts of his agenda and his worst inclinations.

Sept. 5, 2018

13133

Leer en español · 阅读简体中文版 · 閱讀繁體中文版 · 한국어로 읽기 · 日本語で読む

Written by **"Anonymous"**

Quantitative text analysis

The New York Times

Opinion

I Am Part of the Resistance Inside the Trump Administration

I work for the president but like-minded colleagues and I have vowed to thwart parts of his agenda and his worst inclinations.

Sept. 5, 2018

13133

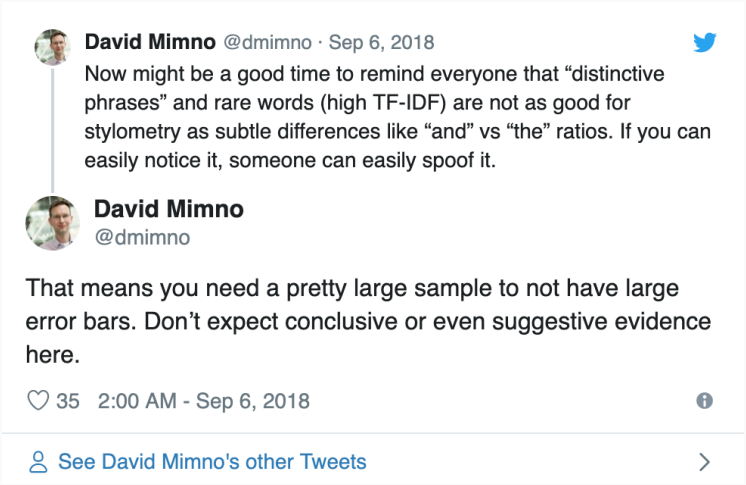
Leer en español · 阅读简体中文版 · 閱讀繁體中文版 · 한국어로 읽기 · 日本語で読む

"We may no longer have Senator McCain. But we will always have his example – a lodestar for restoring honor to public life and our national dialogue."

“Lodestar”

‘a person or thing that serves as an inspiration or guide’

Stylometry



David Mimno @dmimno · Sep 6, 2018

Now might be a good time to remind everyone that “distinctive phrases” and rare words (high TF-IDF) are not as good for stylometry as subtle differences like “and” vs “the” ratios. If you can easily notice it, someone can easily spoof it.

David Mimno
@dmimno

That means you need a pretty large sample to not have large error bars. Don’t expect conclusive or even suggestive evidence here.

♡ 35 2:00 AM - Sep 6, 2018

See David Mimno's other Tweets

“Anonymous”

In October 2020, “Anonymous” revealed himself to be Miles Taylor, a former senior Trump administration official in the DHS.

What is quantitative text analysis

An approach to learning from text that relies on quantification of its textual contents.

- Different from, for example, discourse analysis, which is generally more interested in interpretation, in reading between the lines

We can distinguish between manual approaches and computational approaches to qta (or a combination of both)

- Our focus in this class is on learning about such computational approaches (we'll distinguish between dictionary methods, supervised methods, unsupervised methods)

What is quantitative text analysis

Computational quantitative text analysis is not one-size-fits all, but highly application-dependent but, generally, an application follows three steps

1. Identify texts and units of texts for analysis (identify the **corpus**)
2. Extract quantitatively measured features from these texts and convert them to a quantitative feature matrix
3. Analyse this matrix with statistical methods to draw inferences about these texts

Why quantitative text analysis?

- Political actors (politicians, parties, citizens, etc) produce huge amounts of text, much of which is stored online
  - Speeches, interviews, blog posts, manifestos, social media posts, etc.
- Exciting possibilities to analyze politics beyond elections / beyond election surveys
  - Fine-grained data to learn about public opinion, political behavior, social networks, etc.
- This requires a new set of tools and methods, which computational quantitative text analysis provide

This course

- Introduction of (computational) quantitative text analysis methods in political science using R
- We'll cover the bigger picture of doing research using text
  - However, each of these steps could fill weeks to cover in detail
- Use this course to figure out what you find interesting and want to pursue further
- Ask questions – and help each other out

This course

- No better time to learn these methods than now
- We'll mostly focus on various bag-of-words models but spend some time on word embeddings, as well as images as data
- Lots of cool developments across disciplines!
  - In computer science and linguistics (natural language processing)
  - But also in communication science and psychology, economics and history

This course

- Day 1: What is QTA?
- Day 2: Core assumptions in QTA
- Day 3: Going from text to data
- Day 4: Comparing documents in a corpus
- Day 5: What are dictionaries and how can we validate them?
- Day 6: Human coding and document classification using supervised machine learning
- Day 7: Scaling methods
- Day 8: Topic models
- Day 9: New developments in data
- Day 10: Word embeddings

Requirements: grit

Requirements: fun

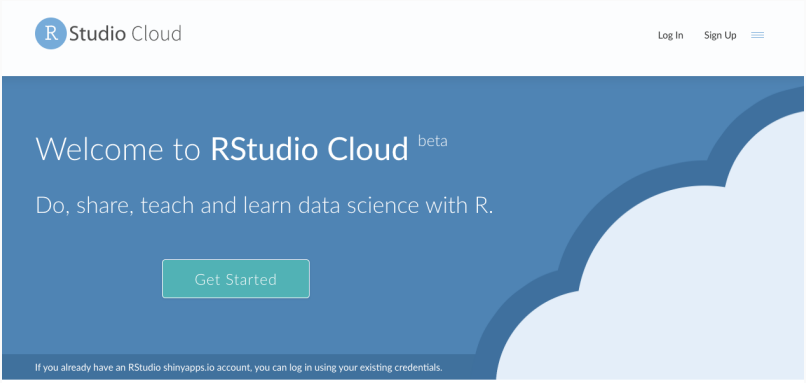
Course objectives

- Learn how computational text analysis methods are used in social science
- Practice preprocessing and analyzing text using R
- Develop a study using text as data
- Critically evaluate existing text as data research

Why R?

- Encompasses all steps of the research process (from scraping to data viz / analysis)
- Tremendously helpful user community
- Lots of development, new packages (we'll mostly rely on **quanteda**, tidyverse, ggplot2, and stringr)
- Other languages possible as well
  - E.g., Python for some tasks and R for other tasks

R Studio Cloud



A canonical text project

- Obtain text data – develop a **corpus**
- Clean the text data
- Pre-process the text data (select the most relevant features into a quantitative feature matrix)
- Analyse this matrix with statistical methods to draw inferences about these texts

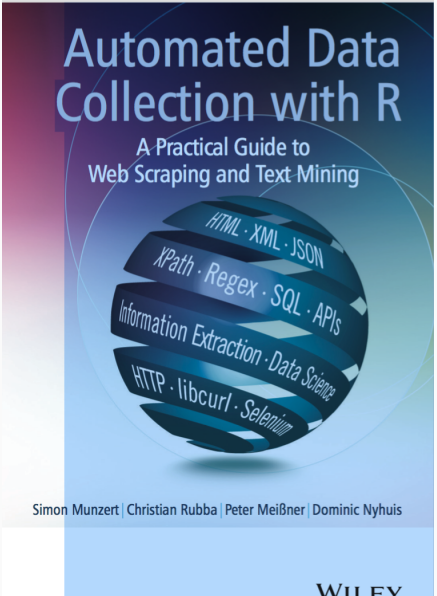
Where to find text data?

- Repositories such as Lexis Nexis (newspaper data)
- Existing text datasets. For example:
  - EUSpeech (Schumacher *et al.*): Harvard Dataverse
  - Parlspeech (Rauh *et al.*): Harvard Dataverse
  - Party manifestos: `https://manifesto-project.wzb.eu/`
  - A general repository of political datasets: `https://github.com/erikgahner/PolData`
- Replication data repositories
- Getting data from the web

Getting data from the web

- APIs (Application Program Interface)
  - Makes parts of a website available to your computer
  - Various R libraries: **GuardianR, RTweet, WikidataR**
  - Oftentimes data stored in JSON and XML format – we'll need to turn that into an R object (e.g., a dataframe)
- Web scraping / screen scraping
  - `rvest`
  - Oftentimes data stored in HTML and XML format
- Last resort: copy-paste text
  - Not recommended – but if you do, make sure to use a plain text editor

Getting data from the web



Cleaning data

- Expect a lot of trial and error
- Important consideration: character encoding
  - Mapping of bits to understandable characters
  - Many coding schemes exists, with different methods of encoding “extended” characters
- Some background: `http://kunststube.net/encoding/`

Preprocessing data

- Stemming, lemmatization, number removal, stopword removal, etc.
- We'll discuss these steps (and how to do them in R) in detail tomorrow
- Goal: select most relevant **features**, and allow for meaningful comparisons between documents in a corpus

Analysis

- Three broad types of analysis (Boumans & Trilling 2016), from most deductive to most inductive:
  - **counting and dictionary methods**: the researcher can fully specify relevant features, and will categorise text accordingly
  - **supervised methods**: the researcher knows how to categorise documents, and uses machine learning methods to learn which features drive this categorisation
  - **unsupervised methods**: the researcher uses qta tools to learn about textual categorisation inductively
- We'll encounter many applications in this course

Tools in R

- Quanteda library

  `In R > Install.packages("Quanteda")`

- Tidyverse

  `In R > install.packages("Tidyverse")`

- We'll be using these packages a lot; Check them out; see what they can do; and be flexible with using one or the other.

Tomorrow

- Read the assigned papers
- Make sure you have an up to date version of R and R Studio installed (or familiarize yourself with R Studio Cloud)
  - If you are new to R, go through `http://qpolr.com/index.html` or start reading R for Data Science (Wickham & Grolemund, 2017) `https://r4ds.had.co.nz/`
- Look at the following snippet of text and list all the ways (you can think of) that it needs to be cleaned:

```
<p>Ladies and gentlemen,</p><p>It is an honour to be here today to introduce the
theme of 'recession and recovery'. If you will permit,  I would like to suggest that
this afternoon we focus more on recovery than on recession. I think we know enough
about the recession side  of the story.</p><p>It started with the fall of Lehman
Brothers on 15 September 2008.. I happened to be here, at the Blouin Creative
Leadership Summit, only ten days later. Everyone was talking about the collapse of
Lehman. They were shocked and alarmed.  But even then we could hardly imagine that
its impact would be so dramatic, so historic.</p><p>As we now know, this event
triggered a global financial and economic crisis. Governments were forced to give
cash injections running into billions to prevent an economic and financial meltdown.
When credit dried up and demand fell, businesses struggled  to keep their heads above
water, and many went under. Ordinary people's jobs, homes and pensions were at
risk.</p><p>
```