

Fundamentals of Quantitative Text Analysis

Mate Akos

2019 September

MTA TK PTI

For today

1. Introduction
2. Goal of the course
3. What is QTE?
4. Key concepts, assumptions
5. Workflow of QTE
6. Data

About me

- Research Fellow at MTA Centre for Social Sciences
- Previous work on text analysis of IMF Article IV reports, network analysis of fiscal institutions
- PhD (soon to be defended) from CEU (with a brief visiting position at KU Leuven)

Contact

- email: aakos.mate@gmail.com
- all course materials: slides and code:
https://github.com/aakosm/QTASZISZ_2019;
readings in dropbox folder

About you?

- your background, interests
- experience with R (or Python?)
- specific interest in this course

Goals of the course

- Provide bird's-eye view of quantitative text analysis
- Some lecture content but heavy emphasis on applied work
- Befriending R
- Learn to critically evaluate competing methods and when to use which
- Provide the ability to carry out a project using the QTA toolset

course requirements

- Attendance and participation - **15%**
- Small assignments - **5% each**
- Project proposal presentation - **10%**
- Final project - **55%**
 - **3500 word limit**
 - topic up to your imagination
 - structure in the syllabus

Quantitative Text Analysis

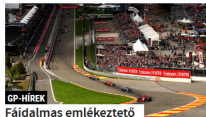


Az ügyészség vádat emelt Hassan F. ellen



Gyanúsan drága és értelmetlen szerződést kötöttek a szocialisták az ügyvédjükkel

Majdnem ötmillió forintot kap azért, hogy az önkormányzati választásról felvilágosítsa az országgyűlési képviselőket. Azokat a képviselőket, akik önkormányzati választáson nem is indulhatnak.



GP-HÍREK

Fájdalmas emlékeztető érkezett: az autósport még mindig nagyon veszélyes

A hétvégi halálos baleset után világossá vált, általában épségben ki lehet szállni a roncsokból. De nem mindig.

→ Még súlyosabb is lehetett volna a spái F2-es baleset

ÍGY LEHET FELISMERNI, HA VALAKI ÉTELFGÜGŐ

Csűrös Karola: Örökké fog tartani az a fájdalom, amit érek

Soha nem tudja már feldolgozni férje, Horváth Ádám elvesztését.

All
IMDbPro
Help
Facebook
Twitter
Instagram
YouTube

Movies, TV & Showtimes
Celebs, Events & Photos
News & Community
Watchlist
Sign in

FULL CAST AND CREW
TRIVIA
USER REVIEWS
IMDbPro
MORE
SHARE

Solo: A Star Wars Story (2018)

PG-13 | 2h 15min | Action, Adventure, Fantasy | 25 May 2018 (USA)

1:29 | Trailer

37 VIDEOS | 504 IMAGES

During an adventure into the criminal underworld, Han Solo meets his future co-pilot Chewbacca and encounters Lando Calrissian years before joining the Rebellion.

Director: Ron Howard

Writers: Jonathan Kasdan, Lawrence Kasdan | [1 more credit »](#)

Stars: Alden Ehrenreich, Woody Harrelson, Emilia Clarke | [See full cast & crew »](#)

[+ Add to Watchlist](#)

Metascore

From metacritic.com

Reviews

2,312 user | 522 critic

Popularity

178 (#106)

What's on David Oyelowo's Watchlist?

The two-time Golden Globe nominee [David Oyelowo](#) has an awards-worthy list of must-see shows. See if you're caught up on his picks.

[Watch now »](#)

Related News

[Phoebe Waller-Bridge To Receive BAFTA La Britannia Award For British Artist Of The Year](#)
30 August 2019 | Deadline

[Woody Harrelson Being Eyed For A Role In The Batman](#)
29 August 2019 | We Got This Covered

Latest Tweets



What's happening?



Tweet



ECPR @ECPR · 14s

Tenure-Track-Professorship for #PoliticalCommunication @unikonstanz 
bit.ly/2KaCXku Apply by Friday 20 September



Maarten Lambrechts Retweeted



Daniel Coe @geo_coe · 4h

Wow—#ArcticDEM is such an incredible data resource! This is a single 2m res. tile of the #LenaRiver Delta in #Russia—such a captivating image! Can't wait to visualize the delta in its entirety. #geography
 data: arcgis.com/apps/webappviewer/...
 hi-res image: flickr.com/photos/1657359...



New York Times World @nytimesworld · 44s

Filled with women stripped of hope and children who die before receiving medical care, a detention camp in Syria has become what aid workers, researchers and American military officials warn is a disaster in the making

The goals and reasons of QTE

- Huge amount of textual data (speeches, documents online, social networks, web scraping, etc.)

The goals and reasons of QTE

- Huge amount of textual data (speeches, documents online, social networks, web scraping, etc.)
- Need to quantify analysis: reproducible research, quantitative approaches apply well for text as data

The goals and reasons of QTE

- Huge amount of textual data (speeches, documents online, social networks, web scraping, etc.)
- Need to quantify analysis: reproducible research, quantitative approaches apply well for text as data
- New era of open source statistical computing: R and Python

The goals and reasons of QTE

- Huge amount of textual data (speeches, documents online, social networks, web scraping, etc.)
- Need to quantify analysis: reproducible research, quantitative approaches apply well for text as data
- New era of open source statistical computing: R and Python
- Cheap and accessible way to deal with huge amount of data

The goals and reasons of QTE

- Huge amount of textual data (speeches, documents online, social networks, web scraping, etc.)
- Need to quantify analysis: reproducible research, quantitative approaches apply well for text as data
- New era of open source statistical computing: R and Python
- Cheap and accessible way to deal with huge amount of data
- "*amplifying and augmenting* careful reading and thoughtful analysis" (Grimmer and Stewart 2012)

The other side of the coin

- Be mindful of the assumptions made during the analysis (e.g.: bag of words approach)
- validate, validate, validate
- Be especially cautious with unsupervised learning methods (more on this later)

Four Principles of Quantitative Text Analysis

Grimmer and Steward (2012, 6):

1. "All quantitative models of language are wrong—but some are useful.

Four Principles of Quantitative Text Analysis

Grimmer and Steward (2012, 6):

1. "All quantitative models of language are wrong—but some are useful.
2. Quantitative methods for text amplify resources and augment humans.

Four Principles of Quantitative Text Analysis

Grimmer and Steward (2012, 6):

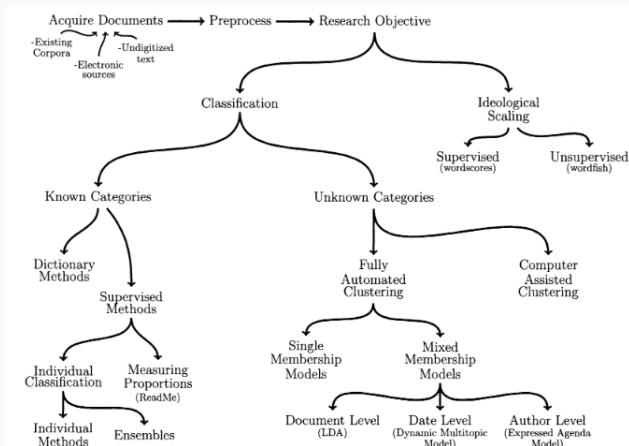
1. "All quantitative models of language are wrong—but some are useful.
2. Quantitative methods for text amplify resources and augment humans.
3. There is no globally best method for automated text analysis.

Four Principles of Quantitative Text Analysis

Grimmer and Steward (2012, 6):

1. "All quantitative models of language are wrong—but some are useful.
2. Quantitative methods for text amplify resources and augment humans.
3. There is no globally best method for automated text analysis.
4. **Validate, Validate, Validate."**

Workflow and methods of QTA



Key concepts

- Unit of analysis: **documents** or texts (depends on your research)
- Population of documents: the **corpus** (collection of corpus: corpora)

Bag of words assumption:

- word orders rarely matter
- break the text into smaller units (**tokens**), usually individual words
- Single words: **unigram**
- If word ordering matters: **n-grams** (for 2 words: bi-grams; 3 words: tri-grams)

n-gram example

tokens from 1 document.

text1 :

```
[1] "The"      "quick" "brown" "fox"      "jumps" "over"  
[7] "the"      "lazy"  "dog"
```

tokens from 1 document.

text1 :

```
[1] "The_quick"    "quick_brown" "brown_fox"  
[4] "fox_jumps"    "jumps_over"  "over_the"  
[7] "the_lazy"     "lazy_dog"
```

Key terms

- **types**: unique word
- **tokens**: any word (total count = total words)
- **stemming**: removing suffixes from the words
- **stop words**: set of words to be removed from the documents, because they do not contain relevant information.
- **dictionary**: A set of tokens with equivalent meaning

Stemming example

tokens from 1 document.

text1 :

[1]	"political"	","	"losing"	","
[5]	"parliament"	","	"John"	

tokens from 1 document.

text1 :

[1]	"polit"	","	"lose"	","
[5]	"parliament"	","	"John"	

What would be your stopwords for this sentence?

"My fellow citizens: I stand here today humbled by the task before us, grateful for the trust you've bestowed, mindful of the sacrifices borne by our ancestors. "

Stop words

What would be your stopwords for this sentence?

"My fellow citizens: I stand here today humbled by the task before us, grateful for the trust you've bestowed, mindful of the sacrifices borne by our ancestors. "

"**My** fellow citizens: I stand **here** today humbled **by the** task **before us**, grateful **for the** trust you've bestowed, mindful **of the** sacrifices borne **by our** ancestors. "

Stop words

What would be your stopwords for this sentence?

"My fellow citizens: I stand here today humbled by the task before us, grateful for the trust you've bestowed, mindful of the sacrifices borne by our ancestors. "

"**My** fellow citizens: I stand **here** today humbled **by the** task **before us**, grateful **for the** trust you've bestowed, mindful **of the** sacrifices borne **by our** ancestors. "

"fellow citizens: stand today humbled task, grateful trust you've bestowed, mindful sacrifices borne ancestors. "

Stop words

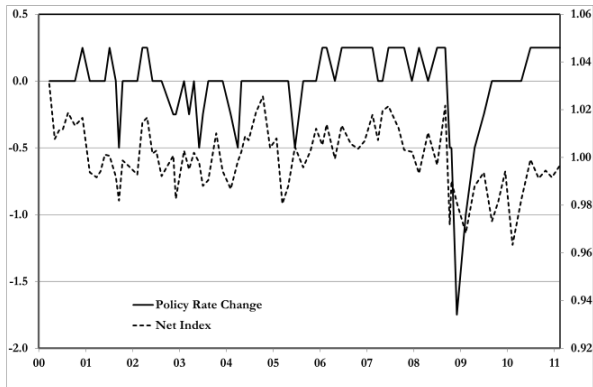
[1]	"i"	"me"	"my"
[4]	"myself"	"we"	"our"
[7]	"ours"	"ourselves"	"you"
[10]	"your"	"yours"	"yourself"
[13]	"yourselves"	"he"	"him"
[16]	"his"	"himself"	"she"
[19]	"her"	"hers"	"herself"
[22]	"it"	"its"	"itself"
[25]	"they"	"them"	"their"
[28]	"theirs"	"themselves"	"what"
[31]	"which"	"who"	"whom"
[34]	"this"	"that"	"these"
[37]	"those"	"am"	"is"
[40]	"are"	"was"	"were"
[43]	"be"	"been"	"being"

Typical workflow

1. Selecting texts, compiling the corpus
2. Cleaning text
 - Make the text machine readable
 - Stemming
 - Removing stop words, numbers, separators
 - Define documents/unit of analysis (reports, sentences, paragraphs, etc.)
3. Preprocess text
 - From words to numbers
 - Defining features (ngrams)
 - Creating a document-feature matrix
4. Reporting

Some examples of applications

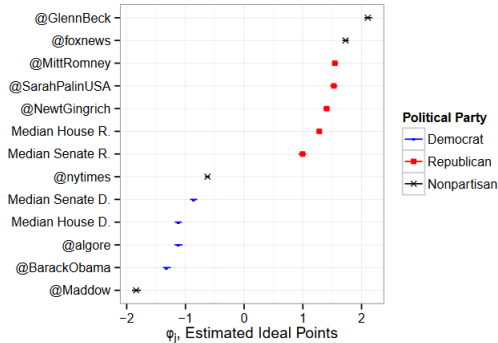
Figure 1. The Net Index and the change of the Riksbank's policy rate



Note: Net Index is on the right-hand axis, policy rate change is on the left-hand axis.

Sentiment analysis of central bank minutes (dictionary method) in Apel, M., & Grimaldi, M. (2012). The information content of central bank minutes. *Riksbank Research Paper Series*, (92).

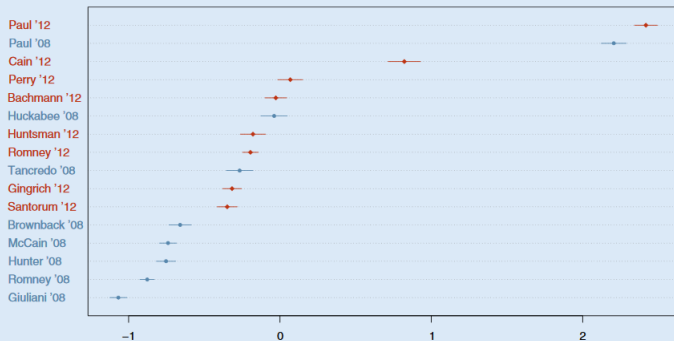
Figure 5: Estimated Ideal Points for Key Political Actors



Ideational scaling of twitter users, in Barberá, Pablo. 2015. "Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data." *Political Analysis* 23(1):76–91.

Figure 1

Candidate Positions



Candidate positions extracted from their pre-Iowa debate speeches with bootstrapped 95% confidence intervals (1,000 replications). 2008 candidates denoted by circles (blue) and 2012 candidates by diamonds (red).

Unsupervised scaling of US politicians, in Medzihorsky, J., Littvay, L., & Jenne, E. K. (2014). Has the tea party era radicalized the republican party? Evidence from text analysis of the 2008 and 2012 republican primary debates. *PS: Political Science & Politics*, 47(4), 806-812.

Where to get data

- already existing datasets (party manifesto data, un speeches, etc)
- news aggregators (lexis nexis is the cheapest)
- webscraping websites (rvest for R, beautifulsoup for Python)
- replication data repositories
- social/news media APIs (Twitter, NYT, etc.)
- Optical Character Recognition (OCR)

Why R?

- Open source and free
- Designed for statistical analysis
- Widespread use in academic and data science community
- Great transferable skillset
- Huge and versatile package ecosystem

For the next week

Download and install R and RStudio (both free)

- Download R for windows from here: <https://cran.r-project.org/bin/windows/base/>
- Download R for OSX from here:
<https://cran.r-project.org/bin/macosx/>
- Download RStudio: <https://www.rstudio.com/products/rstudio/download/>

What we'll do with R

- Learn the basic operations for handling data in R
- Learn to use packages designed for text analysis and handling texts
- Learn how to write papers/reports in R in publication quality