# quanteda

Quantitative Analysis of Textual Data

An Introduction brought to you by:

Laura Menicacci & Dinah Rabe

November 04th 2021

For the course: Intro to Data Science @ Hertie School of Governance

# Overview

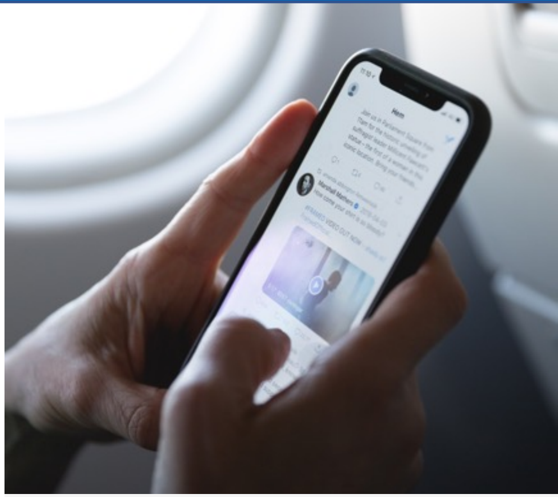
- Motivation
- Getting some buzzwords right
- Examples of quantitative text analysis
- Basics of quanteda
- The simplified workflow
- Main functions
- Some reasons to love quanteda
- Further resources and our references
- The dataset we will work with

# Motivation

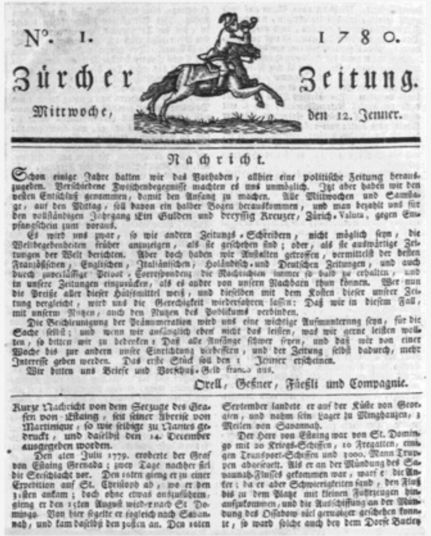
- Most of the data of the world exists in text form
- The volume of available texutal data has increased dramatically
- A lot of data is generated as we speak, tweet or send messages
- But also for example archives are being digitalized (for all german speakers: the swiss news paper NZZ just finalized their digitalized archives with all their newspapers since 1780!)
- This data is highly unstructured in nature

**Definition**

Natural language = human language



NZZ Archiv 1780



N°. 1. 1780.

Zürcher Zeitung.

Mittwoche, den 12. Jenner.

Nachricht.

Schon einige Jahre hatten wir das Vorhaben, allhier eine politiſche Zeitung herauszugeben. Verſchiedene Zwiſchenbegegniſſe machten es uns unmöglich. Jzt aber haben wir den veſten Entſchluß genommen, damit den Anfang zu machen. Alle Mittwochen und Samſtage, auf den Mittag, ſoll davon ein halber Bogen herauskommen, und man bezahlt uns für den vollſtändigen Jahrgang Ein Gulden und dreyſſig Kreuzer, Zürich-Valuta, gegen Empfangſchein zum voraus.

Es wird uns zwar, ſo wie andern Zeitungs-Schreibern, nicht möglich ſeyn, die Weltbegebenheiten früher anzuzeigen, als ſie geſchehen ſind; oder, als ſie auswärtige Zeitungen der Welt berichten. Aber doch haben wir Anſtalten getroffen, vermittelſt der beſten Franzöſiſchen, Engliſchen, Italiäniſchen, Holländiſch- und Deutſchen Zeitungen, und auch durch zuverläſſige Privat-Correſpondenz die Nachrichten immer ſo bald zu erhalten, und in unſere Zeitungen einzurücken, als es andre von unſern Nachbarn thun können. Wer nun die Preiße aller dieſer Hülfsmittel weiß, und dieſelben mit dem Koſten dieſer unſerer Zeitung vergleicht, wird uns die Gerechtigkeit wiederfahren laſſen: Daß wir in dieſem Fall, mit unſerm Nutzen, auch den Nutzen des Publikums verbinden.

Die Beſchleunigung der Pränummeration wird uns eine wichtige Aufmunterung ſeyn, für die Sache ſelbſt; und wenn wir anfänglich eben nicht das leiſten, was wir gerne leiſten wollten, ſo bitten wir zu bedenken: Daß alle Anfänge ſchwer ſeyen, und daß wir von einer Woche bis zur andern unſre Einrichtung verbeſſern, und der Zeitung ſelbſt dadurch, mehr Intereſſe geben werden. Das erſte Stück ſoll den 1 Jenner erſcheinen.

Wir bitten uns Briefe und Vorſchuß-Geld franco aus.

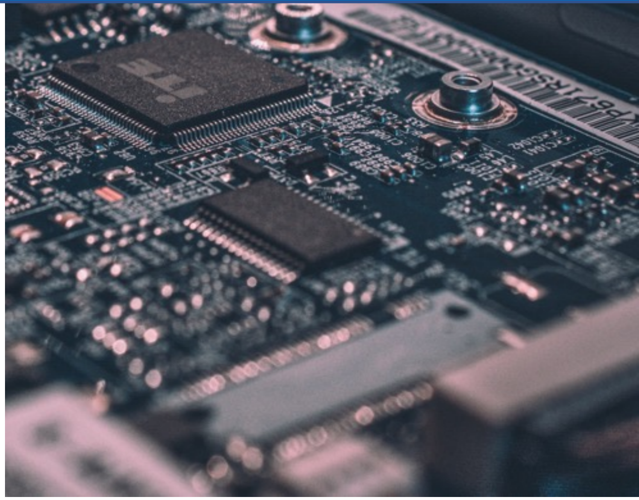
Orell, Geßner, Füeßli und Compagnie.

Kurze Nachricht von dem Seezuge des Grafen von Eſtaing, ſeit ſeiner Abreiſe von Martinique, ſo wie ſelbige zu Nantes gedruckt, und daſelbſt den 14. December ausgegeben worden.

Den 4ten Julii 1779. eroberte der Graf von Eſtaing Grenada; zwey Tage nachher fiel die Seeſchlacht vor. Den 10ten gieng er zu einer Expedition auf St. Chriſtoph ab, wo er den 31ſten ankam; doch ohne etwas auszuführen, gieng er den 15ten Auguſt wieder nach St. Domingo. Von hier ſegelte er ſogleich nach Savannah, und kam daſelbſt den 30ſten an. Den 16ten September landete er auf der Küſte von Georgien, und nahm ſein Lager zu Minghauzen, 3 Meilen von Savannah.

Der Herr von Eſtaing war von St. Domingo mit 20 Kriegs-Schiffen, 10 Fregatten, einigen Transport-Schiffen und 3000. Mann Truppen abgeſegelt. Als er an der Mündung des Savannah-Fluſſes gekommen war, warf e die Anker; da er aber Schwierigkeiten fand, den Fluß bis zu dem Platze mit kleinen Fahrzeugen hinaufzukommen, und die Ausſchiffung an der Mündung des Oſſabow viel geruhiger geſchehen konnte, ſo ward ſolche auch bey dem Dorfe Bucley

# Getting the buzzwords right



**Natural Language Processing,** also referred to as „Computational Linguistics":

- program computers/machines to "read" text (or another input such as speech) by simulating the human ability to understand a *natural* language
- Any kind of computer manipulation of natural language

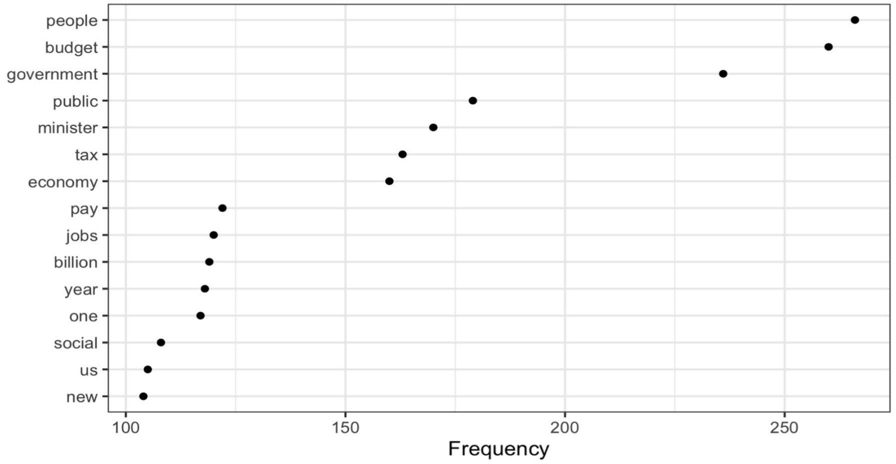
Examples: Chatbots, Speech Recognition, Google Translate

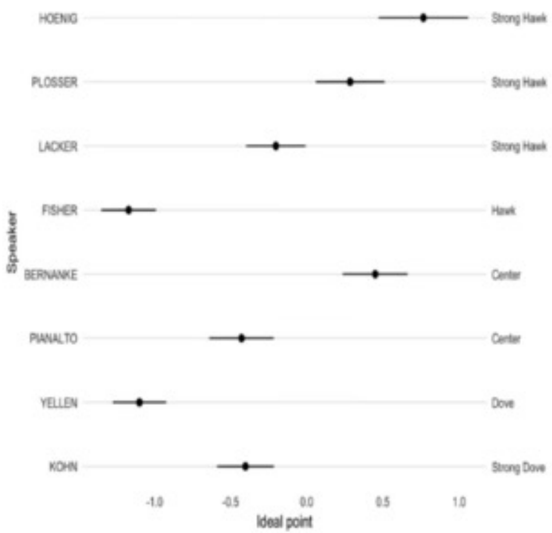
**Quantitative Text Analysis/Text Analytics**

- is the process of deriving meaningful information from natural language text
- it is expressly quantitative, meaning representing textual content numerically but also analysing it as such using computation and statistical methods

# Examples of quantitative text analysis

Descriptive statistics of words



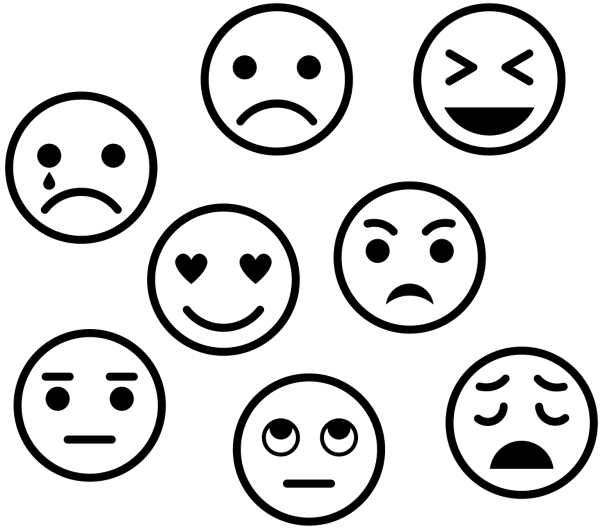


By Hertie Staff:

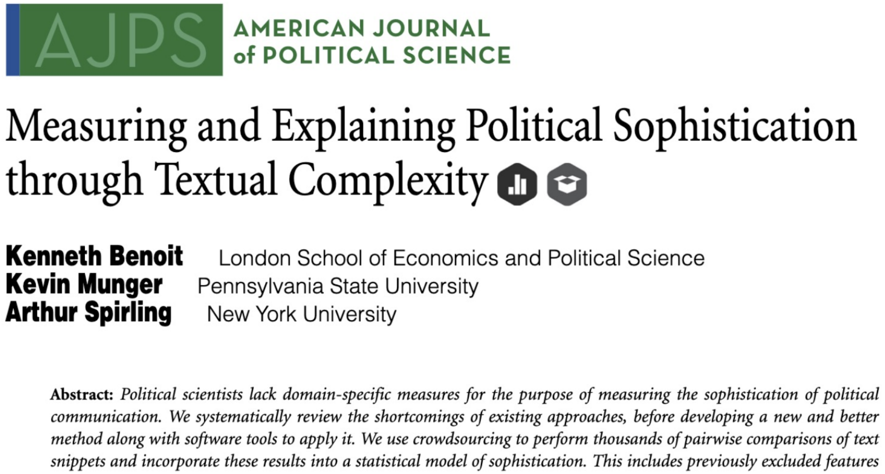
Central Banker's „positions" on inflation: Comparing text based positions to „expert" placements

by Baerg an Lowe (2018)

→ Will Lowe is part of the Hertie Faculty!



Sentiment analysis



AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

Measuring and Explaining Political Sophistication through Textual Complexity

**Kenneth Benoit** London School of Economics and Political Science
**Kevin Munger** Pennsylvania State University
**Arthur Spirling** New York University

**Abstract:** *Political scientists lack domain-specific measures for the purpose of measuring the sophistication of political communication. We systematically review the shortcomings of existing approaches, before developing a new and better method along with software tools to apply it. We use crowdsourcing to perform thousands of pairwise comparisons of text snippets and incorporate these results into a statistical model of sophistication. This includes previously excluded features*

By the creator of the quanteda package:

Kenneth Benoit

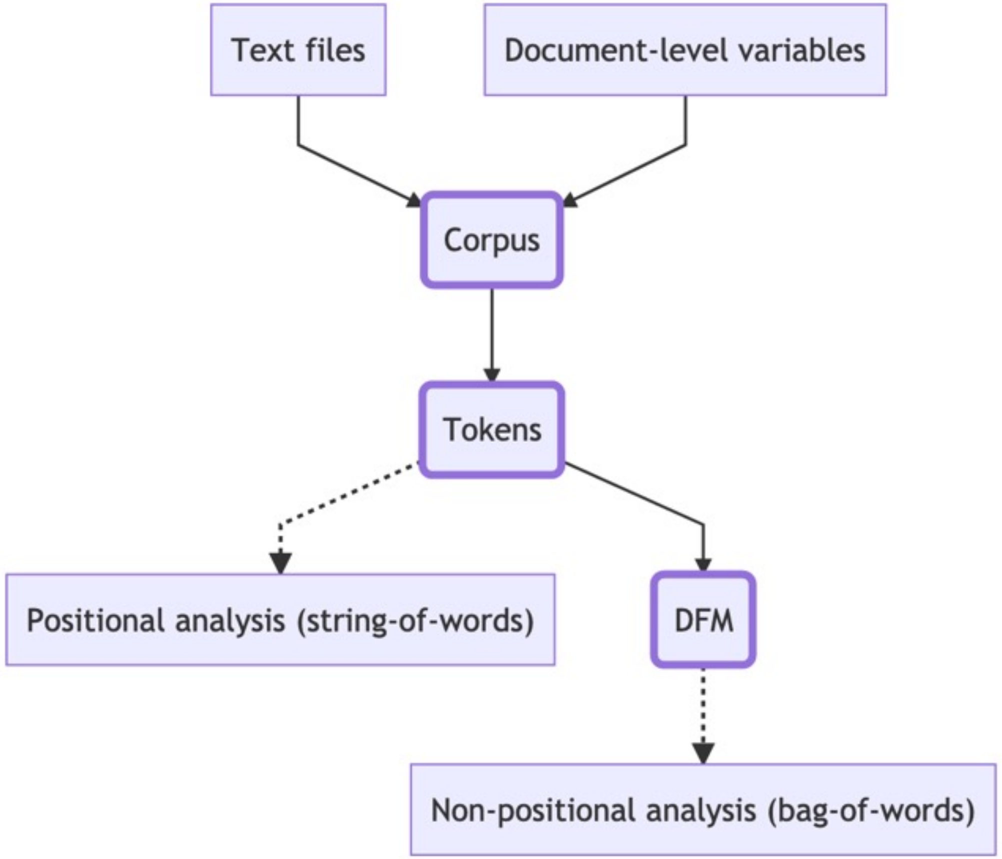
# Basics of quanteda

Quanteda works with 3 main objects tha
you need to know:

**Corpus**

**Tokens**

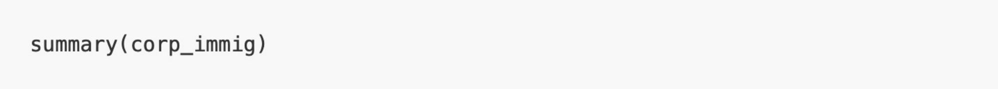
**Document Feature Matrix (DFM)**


Text files
Document-level variables
Corpus
Tokens
Positional analysis (string-of-words)
DFM
Non-positional analysis (bag-of-words)

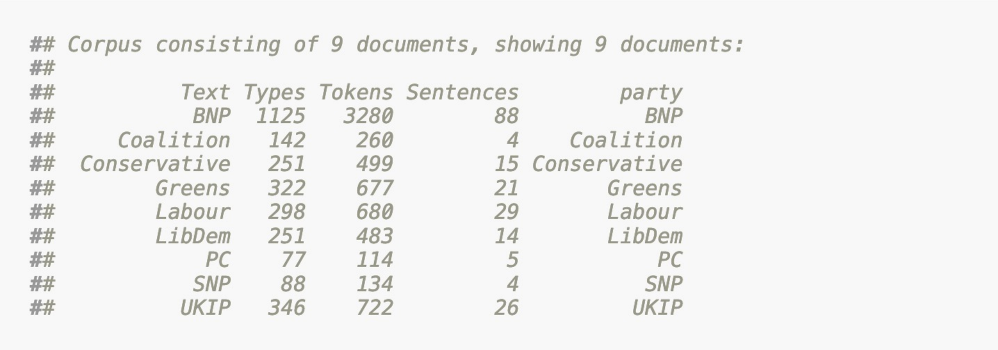
Visualization from quanteda tutorial: https://quanteda.io/

# Corpus

- body/set of texts
- Similar to a dataframe
- Often contains document-level variables (*docvars*)
- Docvars are information associated with the documents



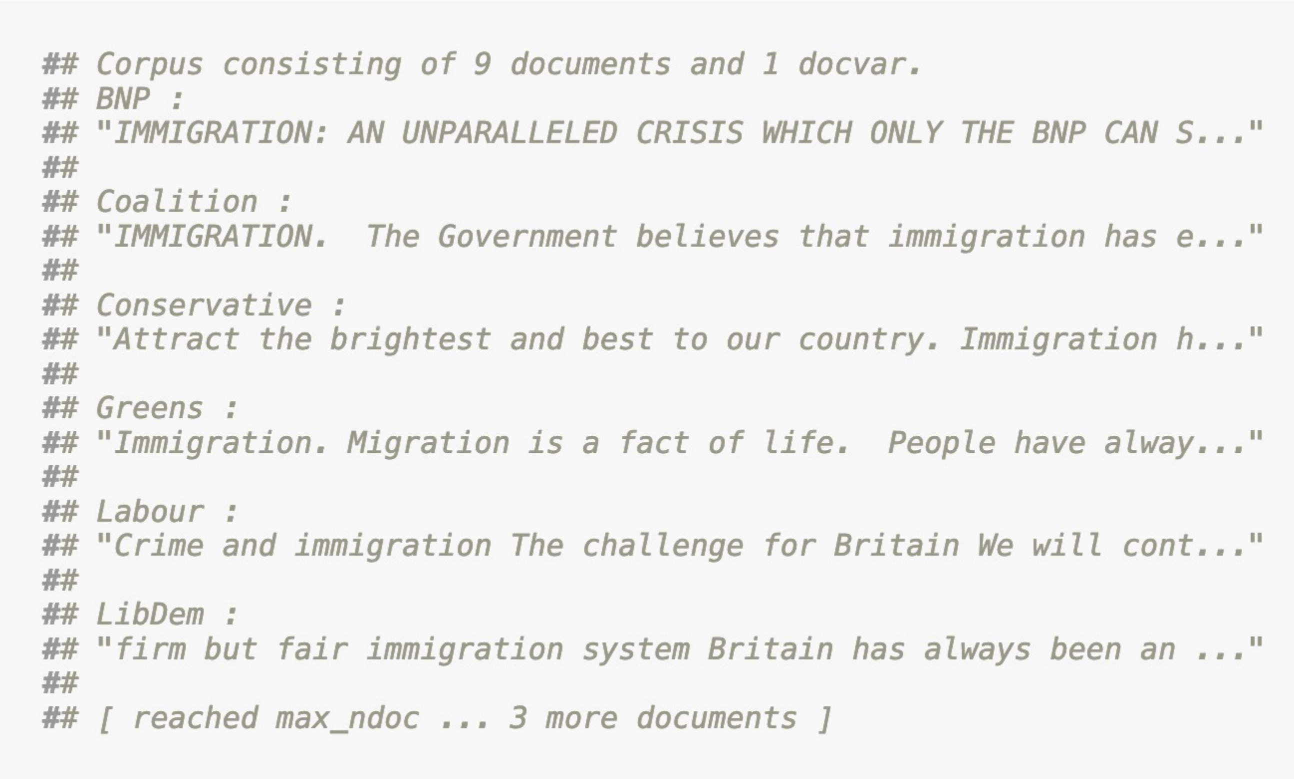
```
summary(corp_immig)
```



```
## Corpus consisting of 9 documents, showing 9 documents:
##
##          Text Types Tokens Sentences        party
##           BNP  1125   3280        88          BNP
##     Coalition   142    260         4    Coalition
##  Conservative   251    499        15 Conservative
##        Greens   322    677        21       Greens
##        Labour   298    680        29       Labour
##        LibDem   251    483        14       LibDem
##            PC    77    114         5           PC
##           SNP    88    134         4          SNP
##          UKIP   346    722        26         UKIP
```

**Example**

Sections of British election manifestos on the topics of immigration and asylum.



```
## Corpus consisting of 9 documents and 1 docvar.
## BNP :
## "IMMIGRATION: AN UNPARALLELED CRISIS WHICH ONLY THE BNP CAN S..."
##
## Coalition :
## "IMMIGRATION.  The Government believes that immigration has e..."
##
## Conservative :
## "Attract the brightest and best to our country. Immigration h..."
##
## Greens :
## "Immigration. Migration is a fact of life.  People have alway..."
##
## Labour :
## "Crime and immigration The challenge for Britain We will cont..."
##
## LibDem :
## "firm but fair immigration system Britain has always been an ..."
##
## [ reached max_ndoc ... 3 more documents ]
```

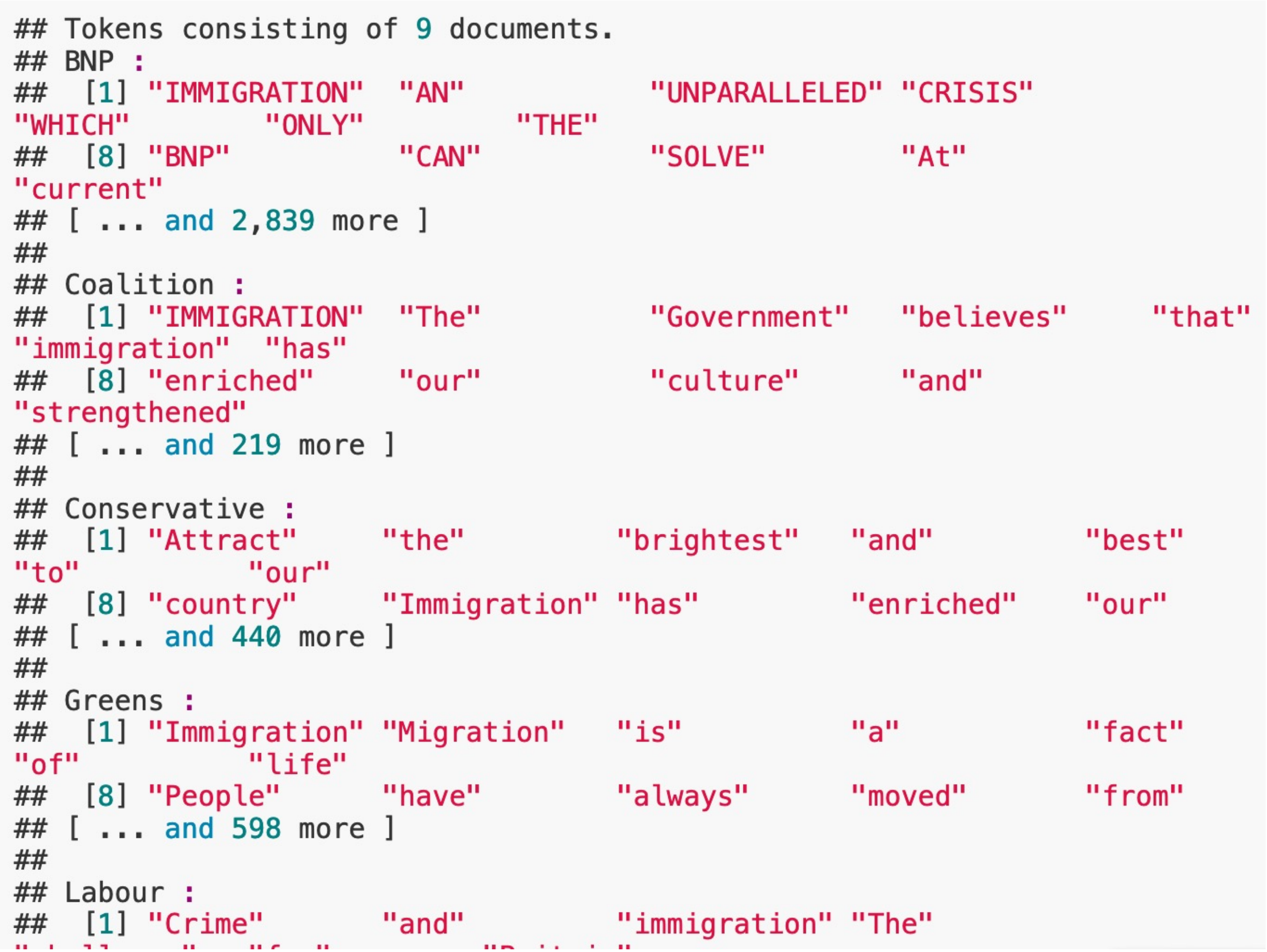
# Token

- a sequence of characters that are grouped together as a useful semantic unit, often a word  (could also be setences)

- Tokenization is the process of splitting text into tokens

- In our  example we will be working with words

Little definition heads up:
A *type* is a unique token



```
## Tokens consisting of 9 documents.
## BNP :
##   [1] "IMMIGRATION"  "AN"            "UNPARALLELED" "CRISIS"
"WHICH"         "ONLY"          "THE"
##   [8] "BNP"          "CAN"           "SOLVE"        "At"
"current"
## [ ... and 2,839 more ]
##
## Coalition :
##   [1] "IMMIGRATION"  "The"           "Government"   "believes"     "that"
"immigration"  "has"
##   [8] "enriched"     "our"           "culture"      "and"
"strengthened"
## [ ... and 219 more ]
##
## Conservative :
##   [1] "Attract"     "the"          "brightest"   "and"          "best"
"to"           "our"
##   [8] "country"     "Immigration" "has"         "enriched"     "our"
## [ ... and 440 more ]
##
## Greens :
##   [1] "Immigration" "Migration"    "is"          "a"            "fact"
"of"           "life"
##   [8] "People"      "have"         "always"      "moved"        "from"
## [ ... and 598 more ]
##
## Labour :
##   [1] "Crime"       "and"          "immigration" "The"
```

# Document Feature Matrix

- Is constructed out of a tokens object

- Like a dataframe with documents in rows and “features” (of the token) as columns

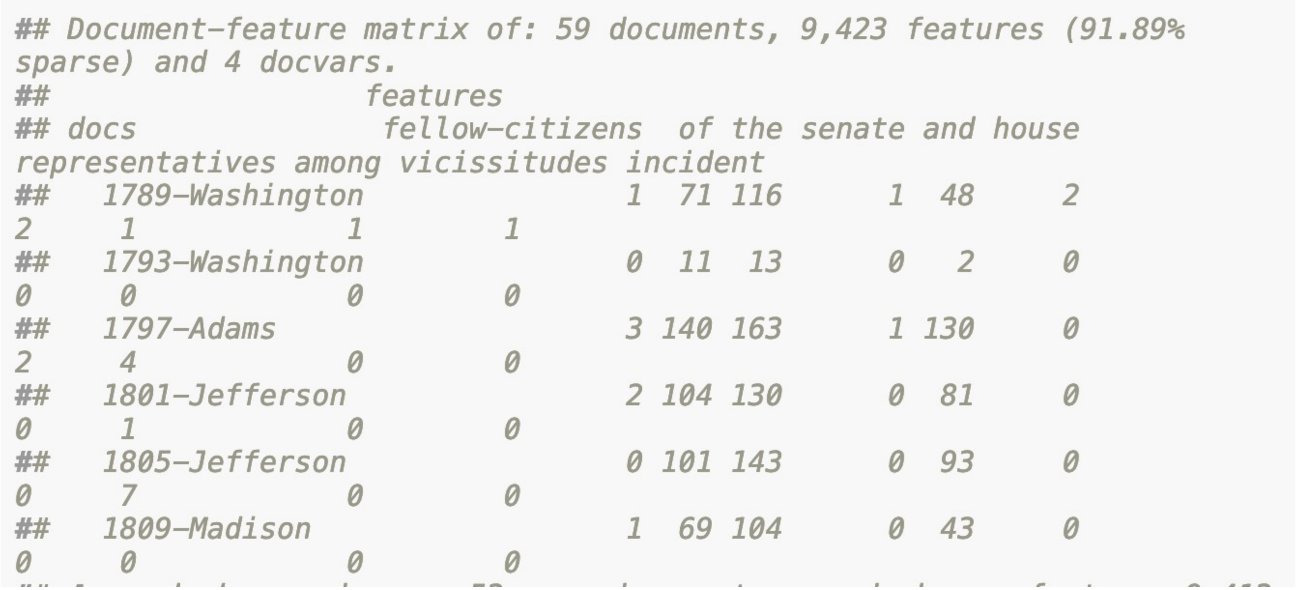
- sparsity/sparseness = the proportion of cells that have zero counts

**Example**

Inaugural speeches of American Presidents

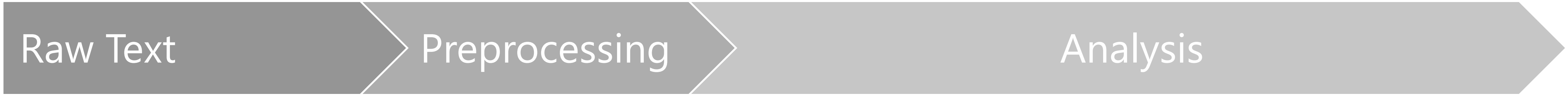
(this dataset is used everyhwere in tutorials on text

analysis)



```
## Document-feature matrix of: 59 documents, 9,423 features (91.89%
sparse) and 4 docvars.
##                    features
## docs                fellow-citizens  of the senate and house
representatives among vicissitudes incident
##   1789-Washington               1  71 116      1  48      2
2     1            1        1
##   1793-Washington               0  11  13      0   2      0
0     0            0        0
##   1797-Adams                    3 140 163      1 130      0
2     4            0        0
##   1801-Jefferson                2 104 130      0  81      0
0     1            0        0
##   1805-Jefferson                0 101 143      0  93      0
0     7            0        0
##   1809-Madison                  1  69 104      0  43      0
0     0            0        0
```

# Reduce the magic to a typical workflow



Raw Text

Preprocessing

Analysis

Social media

News websites

Speeches

Manifestos

Books

... etc

Tokenization

Removing stop words

Stemming

possibly: Create a DFM

**Statistics**:

Word frequencies

(Dis-)similarity

Readability

**Models**:

Word embeddings

Supervised ML

Unsupervised ML

**Plots**:

Key words

networks

word clouds

**Basic package**
quanteda

**Extension package**
quanteda.textstats

**Extension package**
quanteda.textmodels

**Extension package**
quanteda.textplots

# Main function classes

| | |
|---|---|
| Text corpus: | corpus() |
| Tokenization: | tokens() |
| Document-feature matrix: | dfm() |

| | |
|---|---|
| Text statistics: | textstat_() |
| Text models: | textmodel_() |
| Text plots: | textplot_() |

# Corpus functions

- corpus()
- corpus_subset()
- corpus_reshape()
- corpus_segment()
- corpus_sample()

Pre-existing corpora in the quanteda package:

- data_corpus_inaugural
- data_corpus_irishbudget2010

There is an entire package with corpora: quanteda.corpora

# Tokens functions

- tokens()
- tokens_tolower()/tokens_toupper()
- tokens_wordstem()
- tokens_compound()
- tokens_lookup()
- tokens_ngrams()
- tokens_skipgrams()
- tokens_select()/tokens_remove()/tokens_keep()/tokens_replace()
- tokens_sample()
- tokens_subset()

Remember that you can use ?
to lookup the functions

# Some additional terminology of quanteda

**Stems =** words with suffixes removed (using a set of rules)

**Lemmas =** canonical word form

**Stop words =** words that are designed for exclusion from any analysis of text

**Parts of speech =** linguistic markers indicating the general category of a word's inguistic property, e.g. noun, verb, adjective, etc.

**Named entities =** a real-world object, such as persons, locations, organizations, products, etc., that can be denoted with a proper name, often a phrase, e.g. “Hertie School" or "United Kingdom“

**Multi-word expressions =** sequences of words denoting a single concept, e.g. value added tax (in German: Mehrwertsteuer)

# Why quanteda is amazing

- compability with other packages
- You can use a pipelined workflow using magrittr's %>%

THE QUANTEDA INITIATIVE

- UK non-profit organization devoted to the promotion of open-source text analysis software
- software, technical support, teaching and workshops: https://quanteda.org/

# Further resources

Documentation:

- https://quanteda.io
- https://readtext.quanteda.io
- https://spacyr.quanteda.org
- https://github.com/quanteda

Tutorials:

- https://tutorials.quanteda.io

Cheatsheet:

- https://www.rstudio.com/resources/cheatsheets/
- https://github.com/rstudio/cheatsheets/blob/master/quanteda.pdf

# Our references

- All the mentioned further resources

- Workshop presentation of Kenneth Benoit at the University of Münster (27–28 June 2019): https://www.uni-muenster.de/imperia/md/content/ifpol/grasp/2019-06-27_muenster.pdf

- https://manifesto-project.wzb.eu/

# Our example for the tutorial

- The Manifesto Project collects and analyzes parties' electoral programs (manifestos)
- Its data collection is publicly available – data dates back to 1979
- Located at the WZB Berlin Social Science Center and funded by the German Research Foundation

**The Manifesto Corpus** is the digital text collection of the electoral programs

It contains three types of informations:

- machine-readable texts,
- meta-information for each document (such as language and title)
- annotations/codes on the quasi-sentence level(for some documents)

We will make use of these so called CMP codes; they classify sentences with regards to policy topics (isn‘t that amazing?!)

# A few words about ManifestoR

To access the database through Rstudio, you need 2 things:

- The R package MainfestoR
- An API-key

**ManifestoR**

- facilitates downloading and processing the Manifesto Corpus
- it allows bulk downloading several documents at once and transforms the downloaded data into a corpus format

**API-key**

- You need login on the manifesto project website
- There you can create the key on your profile page