quanteda

Quantitative Analysis of Textual Data

INTRODUCTION TO DATA SCIENCE WORKSHOP 2023

KILLIAN, LUCA & ARANXA

AGENDA

- 📚 Text Analysis
- 🔍 Use Cases
- 💡 Intro to Quanteda
- 📝 Quanteda Basics
- 💪 Workflow
- ✅ Why Quanteda?
- 💻 Further Resources

TEXT ANALYSIS 📚

TEXT DATA

Why Text Data?

Text is everywhere!

Larger volumes of text increasingly easily available due to social media and digitalisation

NATURAL LANGUAGE PROCESSING

Natural Language = Human Language

Enables machines to process, understand, interpret or generate natural language

Examples: Chatbots, Speech Recognition, Translation

QUANTITATIVE TEXT ANALYSIS

Subfield of NLP

Use of statistical/computational methods to derive quantitative information from text

Examples: frequency analysis, keyword extraction, sentiment analysis, text visualisation

USE CASES 🔍

SOCIAL SCIENCE USE CASES:

PARTY MANIFESTOS

POLITICAL SPEECHES

SOCIAL MEDIA POSTS

OPEN-ENDED SURVEYS

TODAY’S USE CASE:

BARACK OBAMA'S BEST* SPEECHES

Sharififar, M., & Rahimi, E. (2015). Critical discourse analysis of political speeches: A case study of Obama's and Rouhani's speeches at UN. Theory and Practice in Language studies, 5(2), 343. https://www.academypublication.com/issues2/tpls/vol05/02/14.pdf

QUANTEDA 💡

R package:

For managing and analysing textual data

Available via CRAN as modular packages

quanteda: Core NLP + textual data management functions

quanteda.textmodels: text models + supporting functions

quanteda.textstats: Statistics for textual data

quanteda.textplots: Plots for textual data

Available via GitHub:

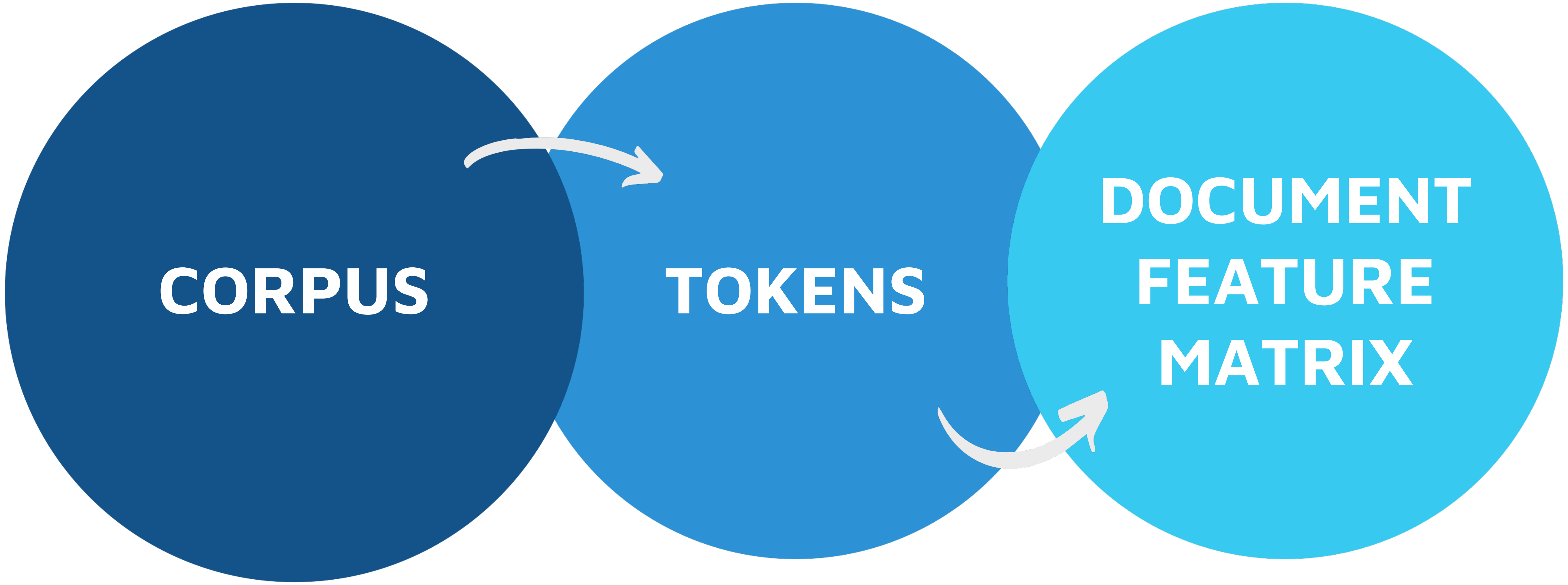
quanteda.sentiment:Sentiment analysis using dictionaries

quanteda.tidy: Extensions for manipulating document variables using tidyverse functions

QUANTEDA BASICS

The quanteda workflow is structured around

THREE MAIN OBJECTS:



CORPUS

TOKENS

DOCUMENT FEATURE MATRIX

CORPUS 📝

What?

Primary data structure for storing and organizing text data and docvars

Why?

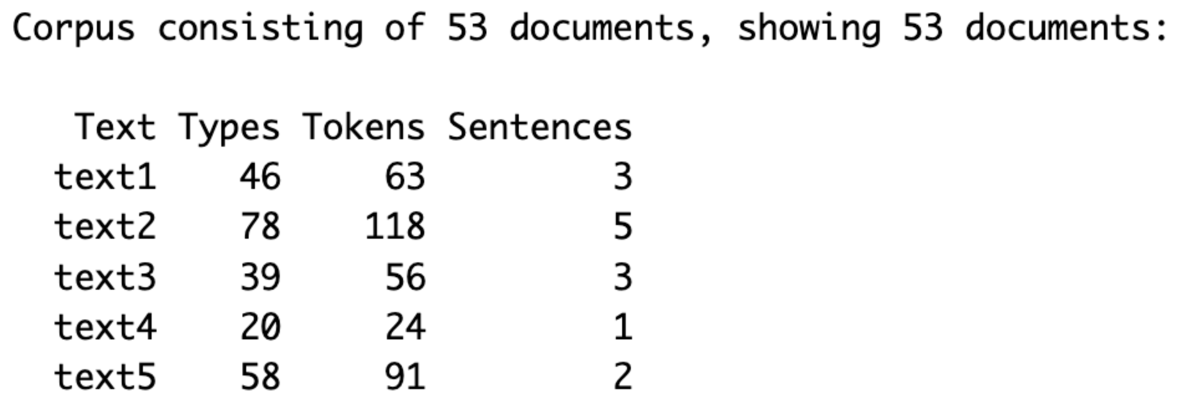
Static “Library” → Copy of original input data

Corpus format required to use quanteda functions

How?

Saves text data with docvars in a data frame

Documents are represented as separate elements --> can be accessed by an index or docvars



```
Corpus consisting of 53 documents, showing 53 documents:

   Text Types Tokens Sentences
  text1    46     63         3
  text2    78    118         5
  text3    39     56         3
  text4    20     24         1
  text5    58     91         2
```

Docvars? 🔍

Document-level or metadata attributes

TOKENS 📝

What?

Tokens = Basic units of text data

Comprise usually of words grouped as semantic units

Preserves the position of words

Why?

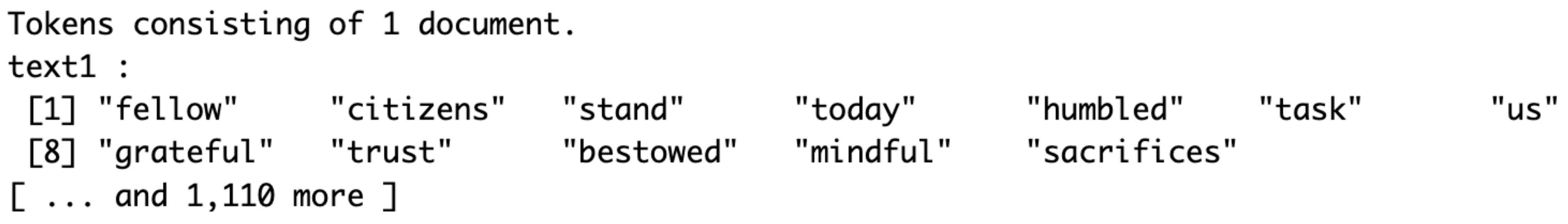
Preprocessing, cleaning and feature extraction operations are performed on tokens

Positional string analysis

How?

Tokenization = Process of splitting text into tokens

Stores tokens in a list of vectors



```
Tokens consisting of 1 document.
text1 :
 [1] "fellow"     "citizens"   "stand"      "today"      "humbled"    "task"       "us"
 [8] "grateful"   "trust"      "bestowed"   "mindful"    "sacrifices"
[ ... and 1,110 more ]
```

DOCUMENT-FEATURE MATRIX (DFM)

What?

Matrix format: Represents frequencies of features in documents

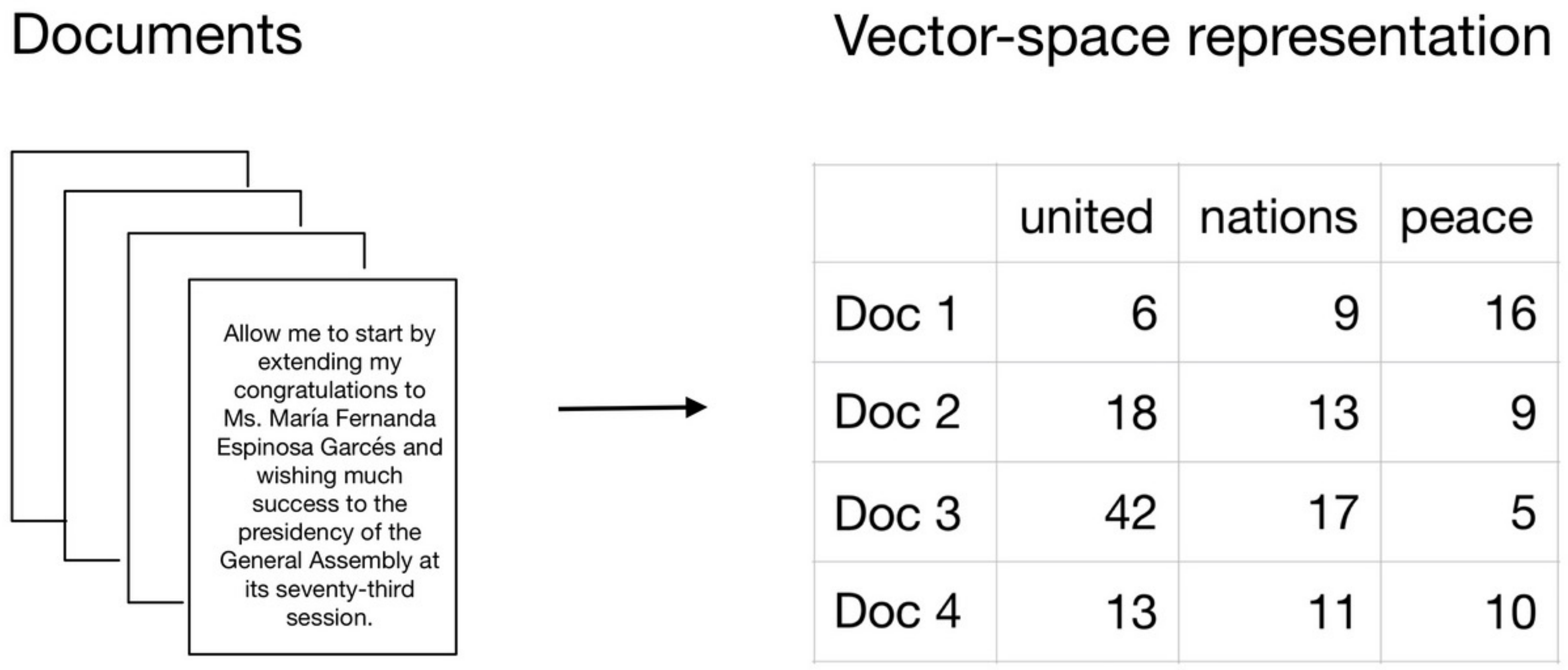
Rows = correspond to documents

Columns = correspond to text features (i.e. tokens)

Why?

Data format for further analysis

Non-positional analyisis



Documents

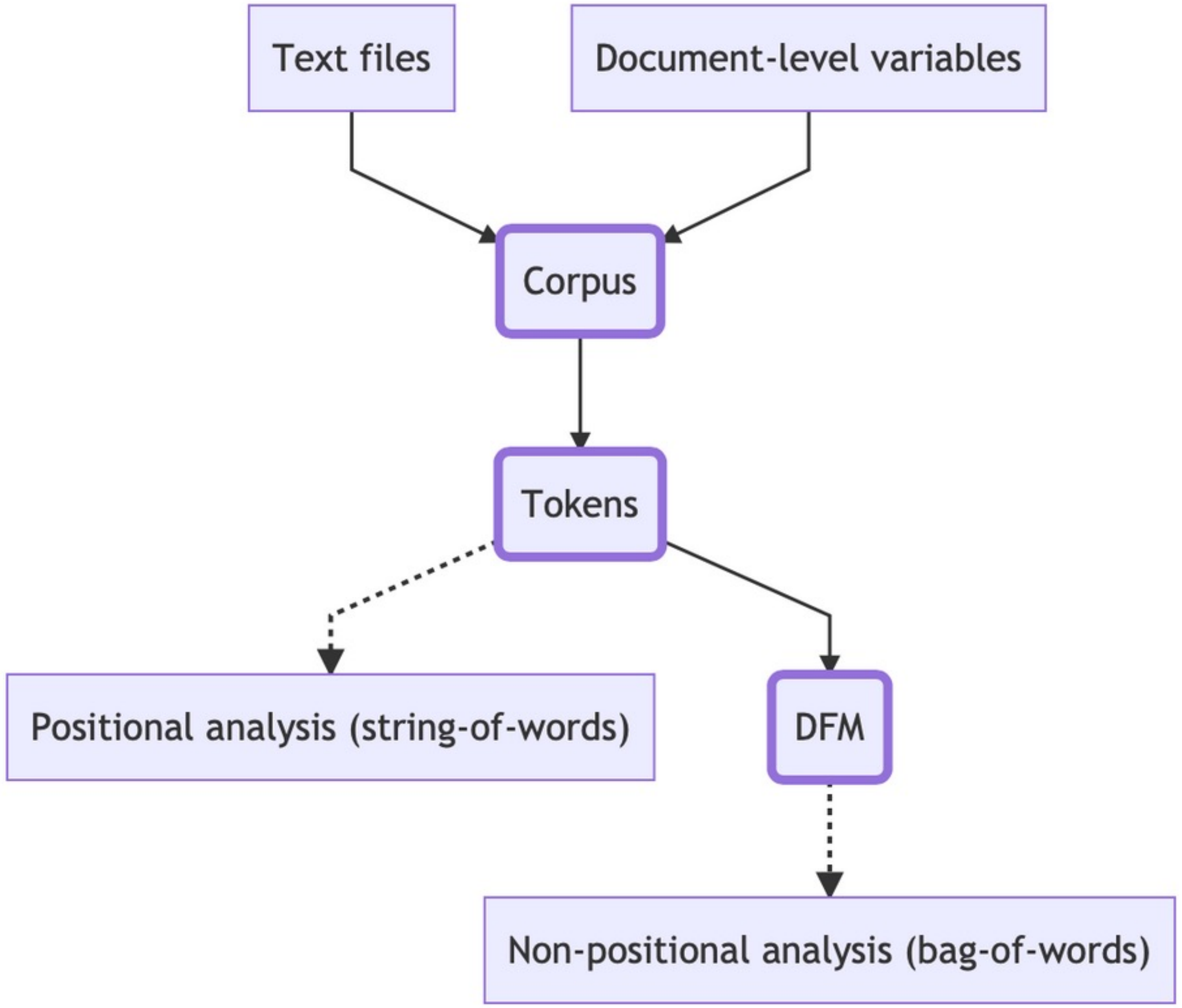
Allow me to start by extending my congratulations to Ms. María Fernanda Espinosa Garcés and wishing much success to the presidency of the General Assembly at its seventy-third session.

Vector-space representation

	united	nations	peace
Doc 1	6	9	16
Doc 2	18	13	9
Doc 3	42	17	5
Doc 4	13	11	10

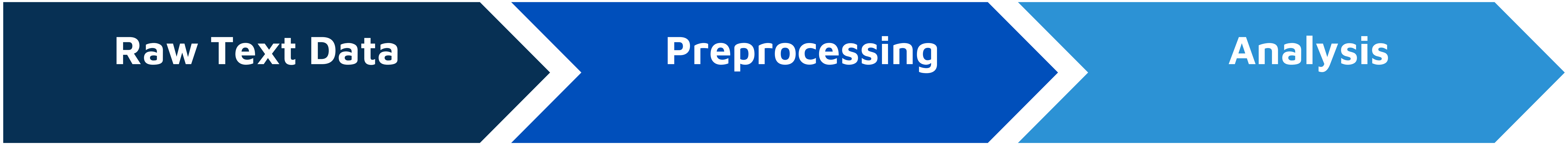
MZES Social Science Data Lab. 2021

WORKFLOW 💪


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

Source: Tutorials.quanteda.io

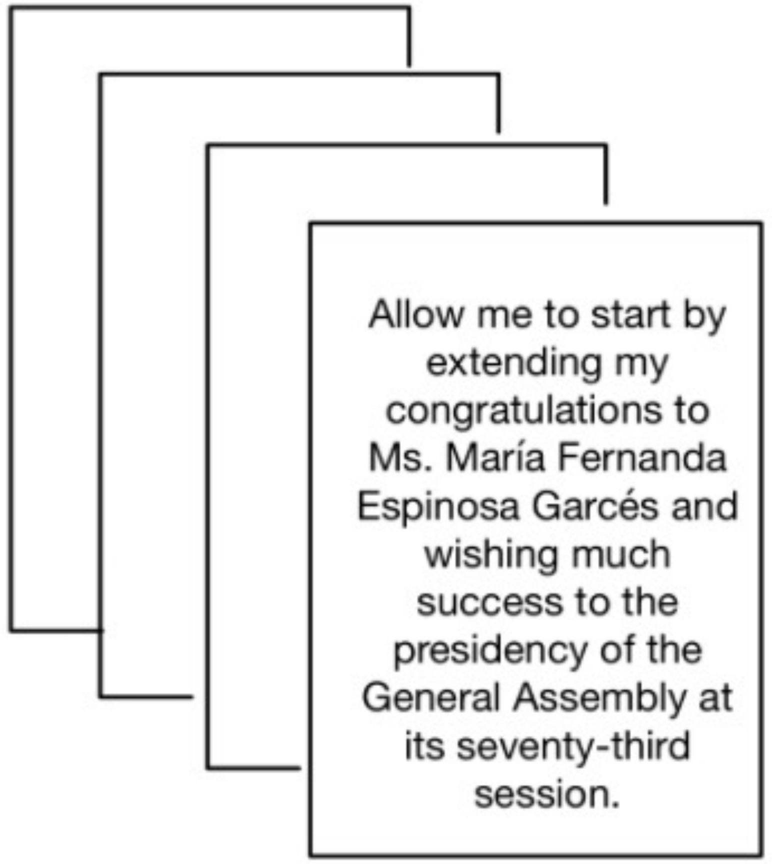
WORKFLOW 💪



Raw Text Data

Preprocessing

Analysis



Allow me to start by extending my congratulations to Ms. María Fernanda Espinosa Garcés and wishing much success to the presidency of the General Assembly at its seventy-third session.

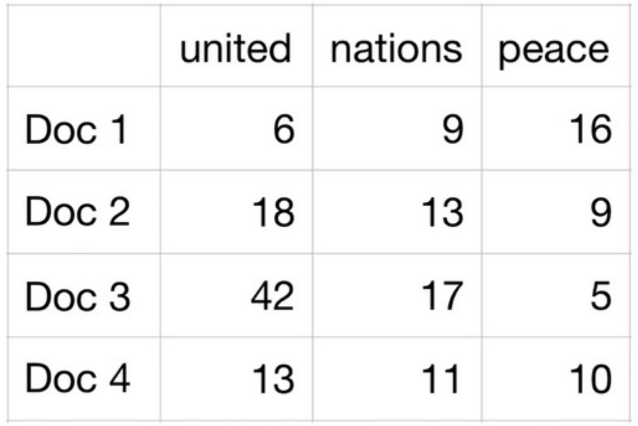
quanteda

Tokenization

Removing stop words, stemming, ...

Feature Selection

DFM



	united	nations	peace
Doc 1	6	9	16
Doc 2	18	13	9
Doc 3	42	17	5
Doc 4	13	11	10

Statistics: **quanteda.textstats**

- Word frequencies
- Key phrases
- Lexical diversity
- Similarity

Models: **quanteda.textmodels**

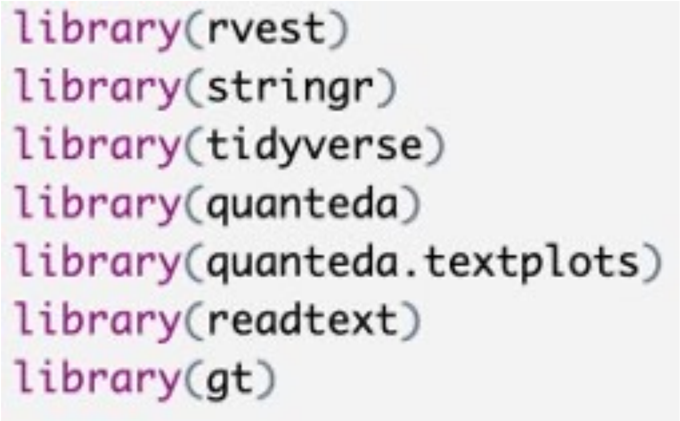
- Supervised ML
- Unsupervised ML
- Word embeddings
- Topic Models

Plots: **quanteda.textplots**

- Networks
- Word clouds
- Keyness

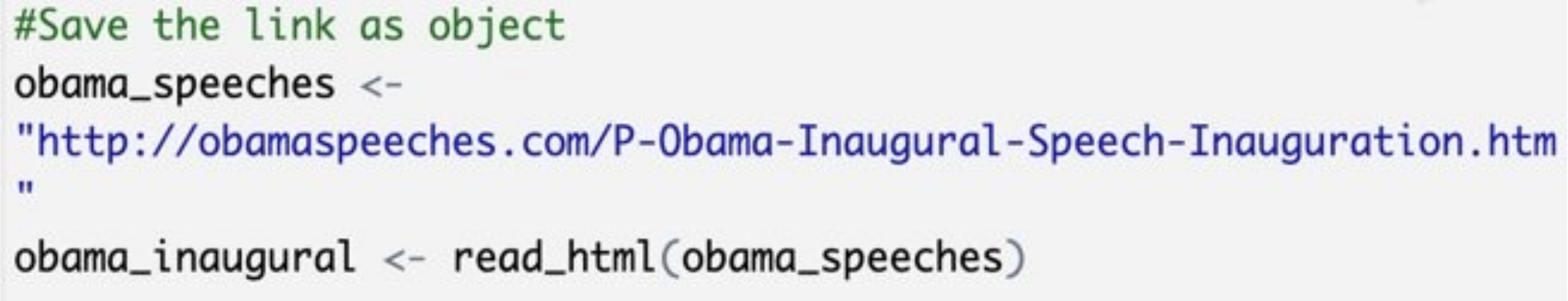
WORKFLOW 💪

GETTING STARTED



```
library(rvest)
library(stringr)
library(tidyverse)
library(quanteda)
library(quanteda.textplots)
library(readtext)
library(gt)
```

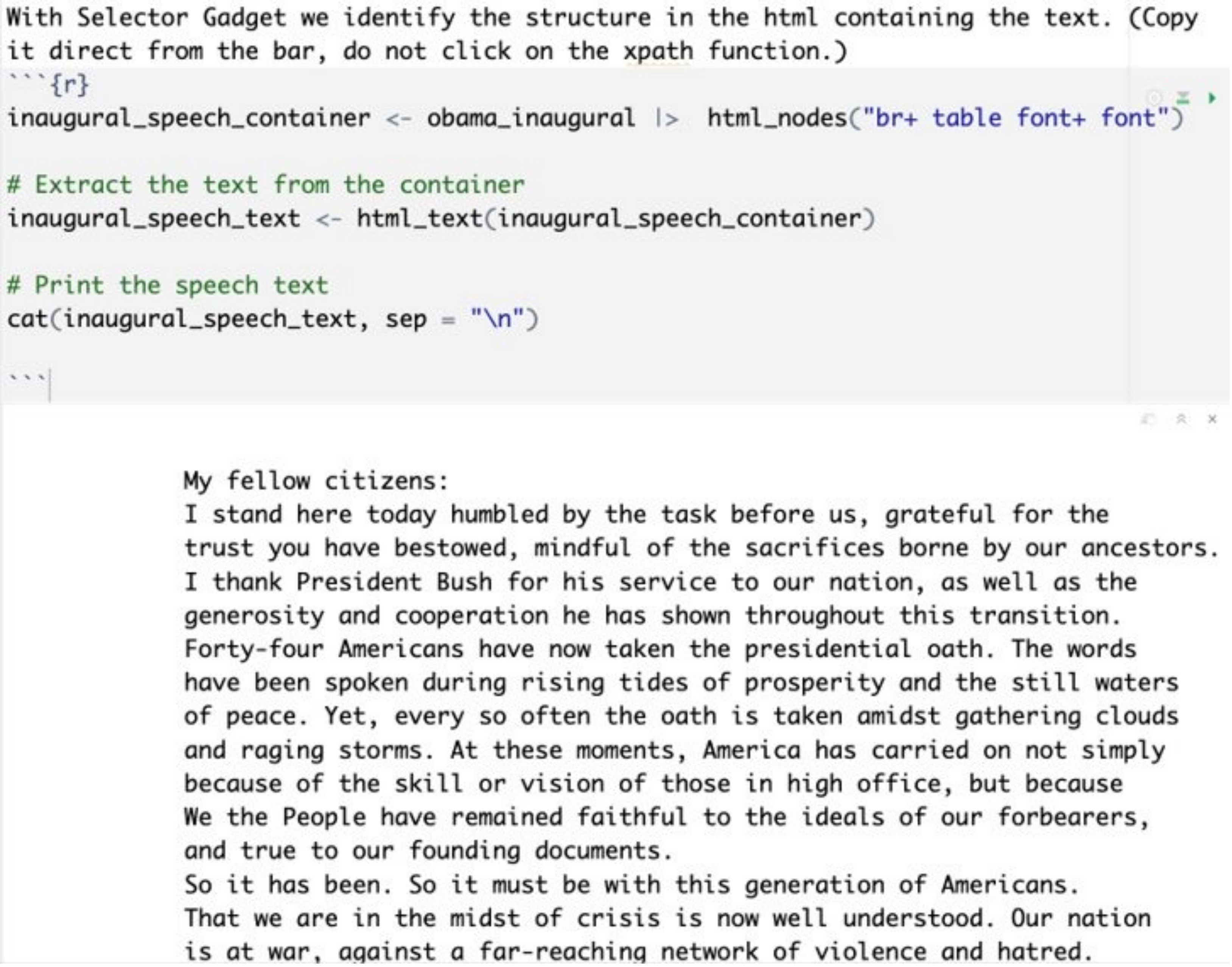
READ DATA INTO R



```
#Save the link as object
obama_speeches <-
"http://obamaspeeches.com/P-Obama-Inaugural-Speech-Inauguration.htm
"
obama_inaugural <- read_html(obama_speeches)
```

WORKFLOW 💪

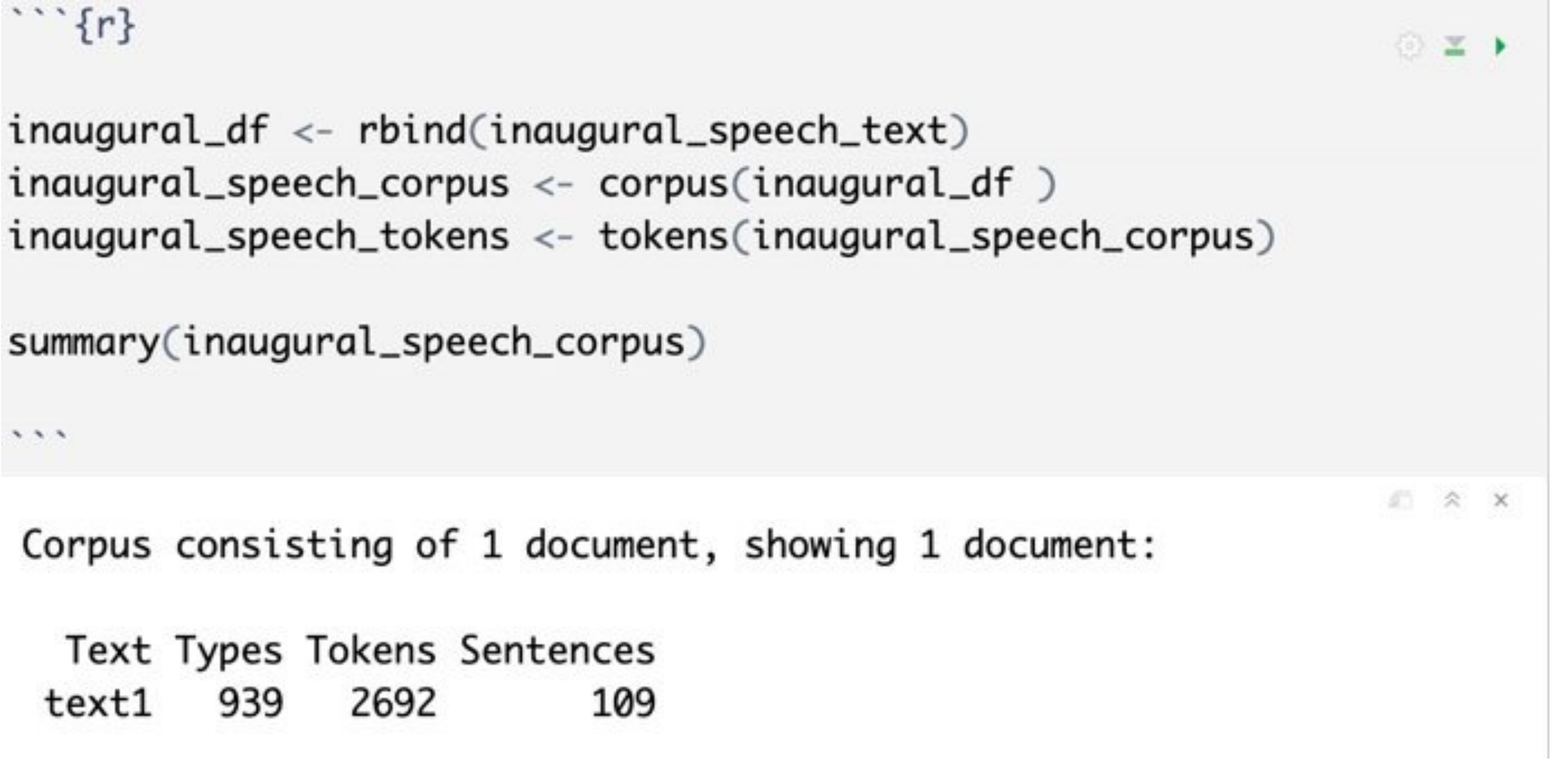
CREATING A CORPUS



With Selector Gadget we identify the structure in the html containing the text. (Copy it direct from the bar, do not click on the xpath function.)

~~~
```{r}
inaugural_speech_container <- obama_inaugural |>  html_nodes("br+ table font+ font")

# Extract the text from the container
inaugural_speech_text <- html_text(inaugural_speech_container)

# Print the speech text
cat(inaugural_speech_text, sep = "\n")

```
~~~

```
          My fellow citizens:
          I stand here today humbled by the task before us, grateful for the
          trust you have bestowed, mindful of the sacrifices borne by our ancestors.
          I thank President Bush for his service to our nation, as well as the
          generosity and cooperation he has shown throughout this transition.
          Forty-four Americans have now taken the presidential oath. The words
          have been spoken during rising tides of prosperity and the still waters
          of peace. Yet, every so often the oath is taken amidst gathering clouds
          and raging storms. At these moments, America has carried on not simply
          because of the skill or vision of those in high office, but because
          We the People have remained faithful to the ideals of our forbearers,
          and true to our founding documents.
          So it has been. So it must be with this generation of Americans.
          That we are in the midst of crisis is now well understood. Our nation
          is at war, against a far-reaching network of violence and hatred.
```

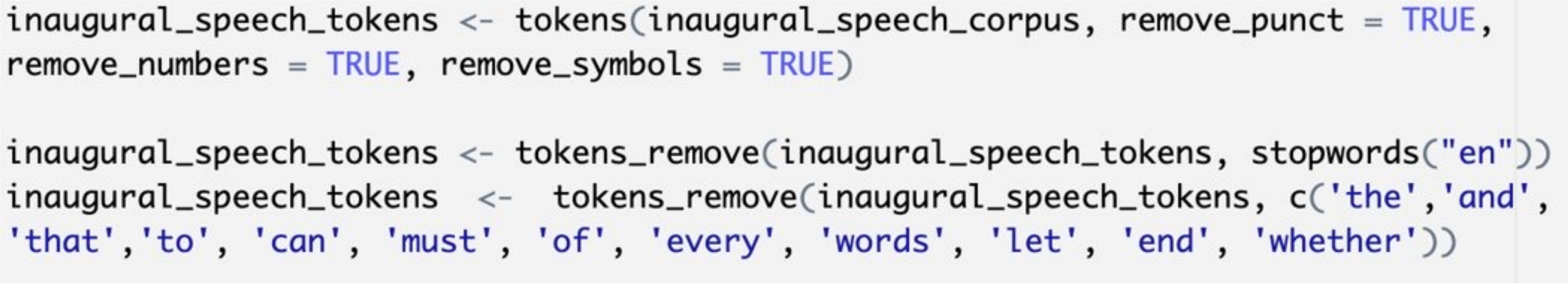


~~~
```{r}

inaugural_df <- rbind(inaugural_speech_text)
inaugural_speech_corpus <- corpus(inaugural_df )
inaugural_speech_tokens <- tokens(inaugural_speech_corpus)

summary(inaugural_speech_corpus)

```
~~~

```
Corpus consisting of 1 document, showing 1 document:

  Text Types Tokens Sentences
 text1   939   2692       109
```

WORKFLOW 💪

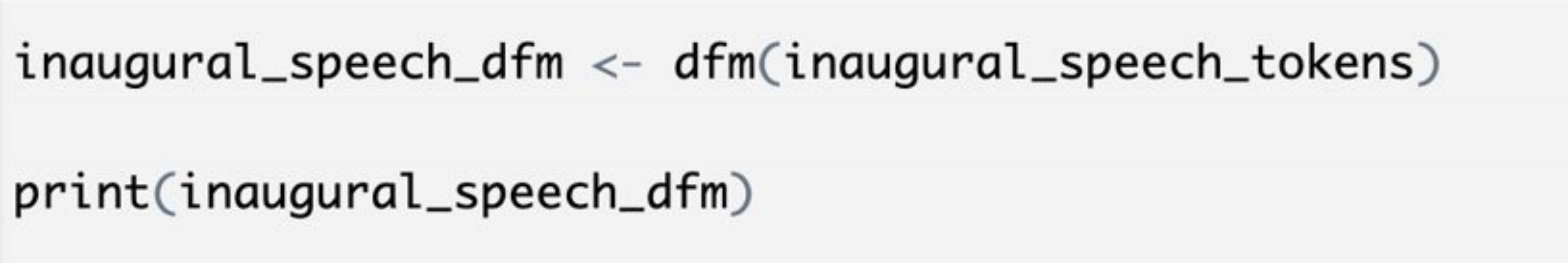
PREPROCESSING: TOKENS + CLEANING



```
inaugural_speech_tokens <- tokens(inaugural_speech_corpus, remove_punct = TRUE,
remove_numbers = TRUE, remove_symbols = TRUE)

inaugural_speech_tokens <- tokens_remove(inaugural_speech_tokens, stopwords("en"))
inaugural_speech_tokens  <-  tokens_remove(inaugural_speech_tokens, c('the','and',
'that','to', 'can', 'must', 'of', 'every', 'words', 'let', 'end', 'whether'))
```

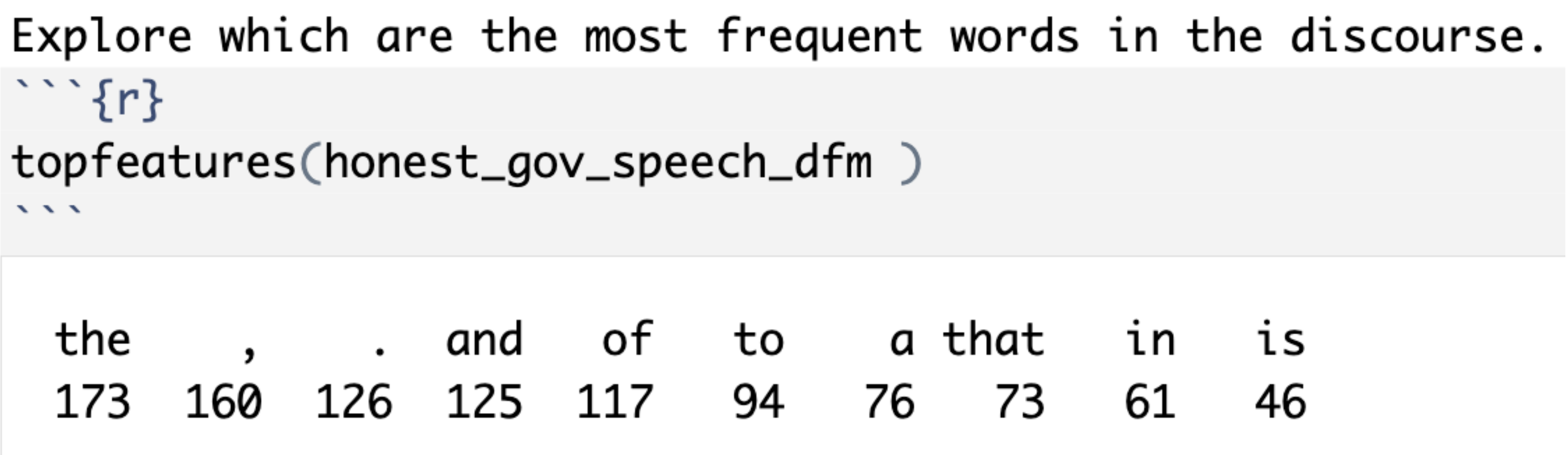
WORKFLOW 💪

DOCUMENT FEATURE MATRIX



```
inaugural_speech_dfm <- dfm(inaugural_speech_tokens)

print(inaugural_speech_dfm)
```

TOP FEATURES



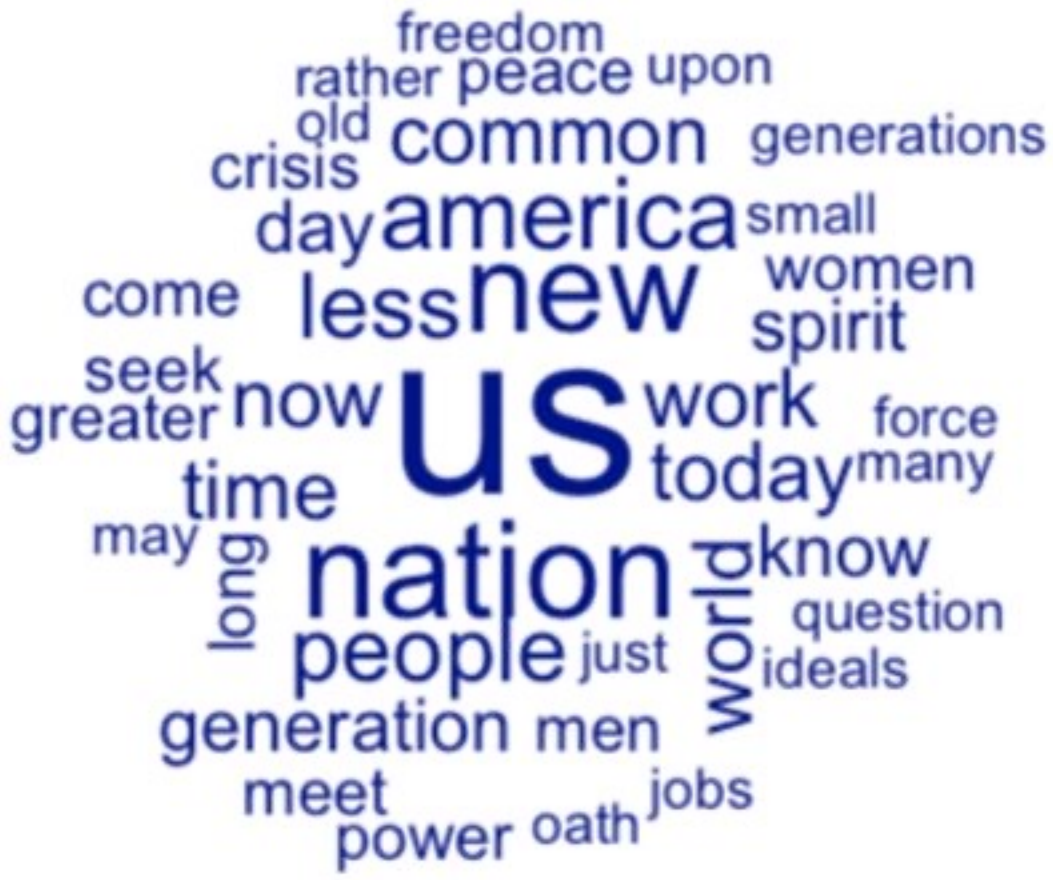
Explore which are the most frequent words in the discourse.

```{r}
topfeatures(honest_gov_speech_dfm )
```

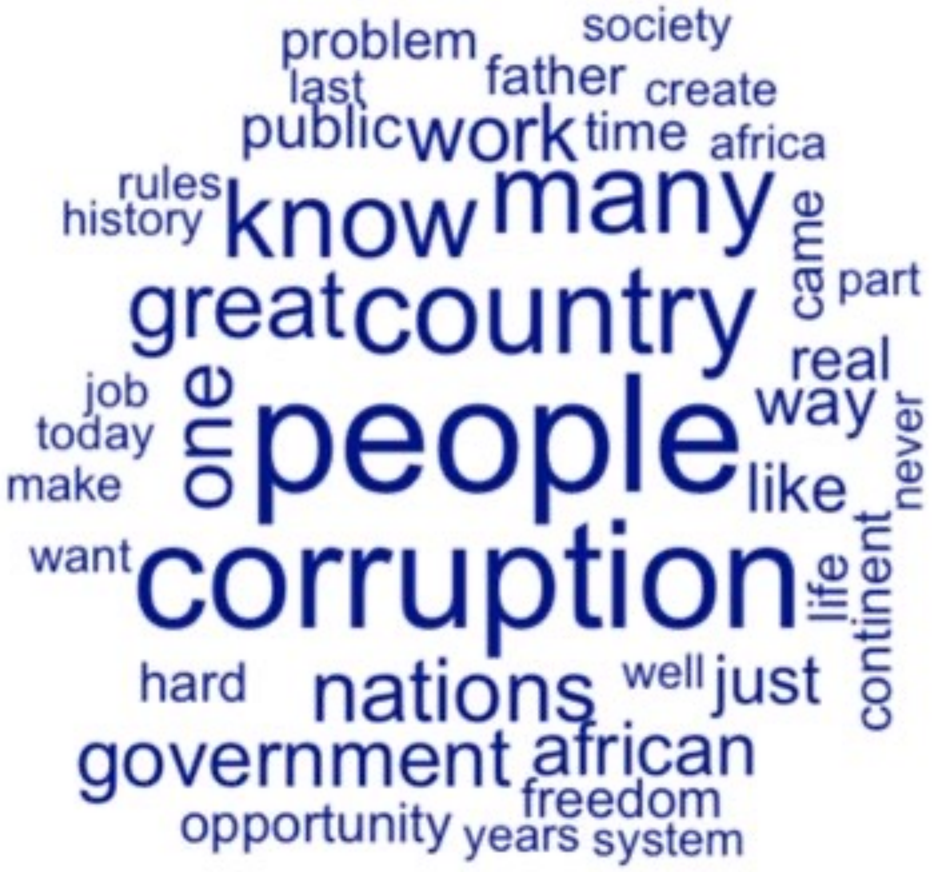
```
  the    ,    .  and   of   to    a that   in   is 
  173  160  126  125  117   94   76   73   61   46 
```

WORKFLOW 💪

ANALYSIS



Inaugural Speech (2009)



An Honest Government – A Hopeful Future (2006)

WHY QUANTEDA? ✅

1. **Compatibility:** E.g.: Tidyverse and |>

2. **Well maintained package:** Quanteda Initiative

3. **Easy to use:** for beginners but offers complex functions too

4. **Efficient:** Fast and efficient package for processing large text data

FURTHER RESOURCES 💻

🔍 Quanteda Website

🔍 Quanteda Tutorial on the Quanteda Website

🔍 Quanteda Cheat Sheet

🔍 Presentation by quanteda founder Kenneth Benoi at the University of Münster

🔍 A Beginner's Guide to Text Analysis with quanteda (University of Virginia)

🔍 quanteda: An R package for the quantitative analysis of textual data, JOSS, 2018

🔍 An Introduction to Text as Data with quanteda (Penn State and Essex courses in "Text as Dara")

🔍 Advancing Text Mining with R and quanteda: Methods Bites

🔍 Quanteda initiative