To and from Tidy Formats

Alfredo Hernandez Sanchez

Central European University

May 16, 2019

The Tidy Format

- In the Tidy format, instead of analyzing a sparse document feature matrix, we obtain a one-token-per-document-per-row with the function unnest_tokens This makes the analysis more compatible with TIDYVERSE tools like GGPLOT, DPLYR and TIDYR.
- But we cannot use other applications like topic models since they are constructed around sparse-matrix formats, like QUANTEDA.

The Syntax

- In the Tidy format, instead of analyzing a sparse document feature matrix, we obtain a one-token-per-document-per-row with the function `unnest_tokens()`.
- The package tidytext has two verbs to change to and from tidy formats:
  - `Tidy()` turns a sparse matrix into a tidy data frame
  - `Cast()` the one-term-per-row a matrix format:
    - For a quanteda object (DFM) we must use: cast_dfm()

Example: Tidying the “Speeches”" DFM

```
# Our trimmed speeches DFM
speeches.dfm
```

```
## Document-feature matrix of: 22 documents, 3,598 features (46.4% sparse)
```

```
# We turn the sparse matrix into a one token per row format
speeches.dfm %>%
  tidy()
```

```
## # A tibble: 42,396 x 3
##    document term     count
##    <chr>    <chr>    <dbl>
##  1 text1    in-offic    30
##  2 text2    in-offic     4
##  3 text3    in-offic    15
##  4 text5    in-offic    19
```

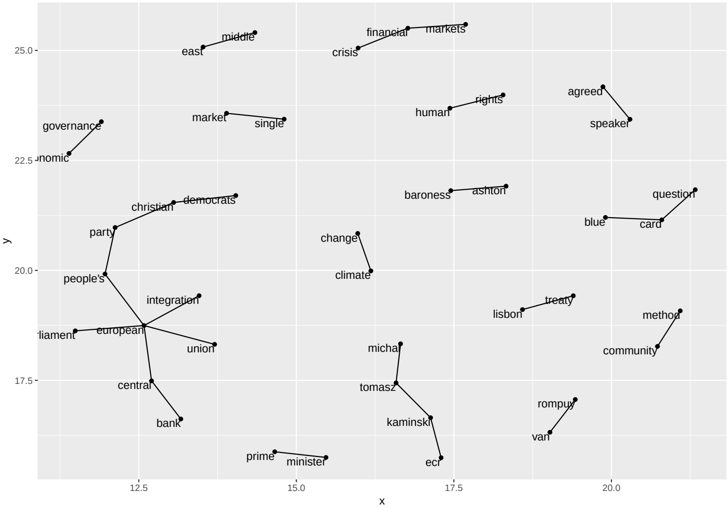
TIDY() to pre-process the CORPUS directly

We can also use `tidy()` directly on the corpus, to turn it into a tibble and then analyze ngrams!

```r
td_speeches <- tidy(speeches)
my_bigrams <- td_speeches %>%
  unnest_tokens(bigram, text, token = "ngrams", n = 2)
my_bigrams
```

```
## # A tibble: 871,231 x 2
##    speaker       bigram
##    <chr>         <chr>
##   1 Brian Crowley in office
##   2 Brian Crowley office of
##   3 Brian Crowley of the
##   4 Brian Crowley the of
##   5 Brian Crowley of the
##   6 Brian Crowley the the
```

We can create several visualization with Tidy formats & GGPLOT


middle
east
financial
markets
crisis
agreed
rights
human
speaker
market
single
governance
nomic
christian
democrats
baroness
ashton
question
blue
card
party
change
climate
people's
integration
treaty
lisbon
method
rliament
european
union
michal
community
central
tomasz
rompuy
kaminski
bank
van
prime
minister
ecr
x
y