Text analysis

Data Science in a Box

datasciencebox.org

Tidytext analysis

Tidytext

- Using tidy data principles can make many text mining tasks easier, more effective, and consistent with tools already in wide use
- Learn more at tidytextmining.com

What is tidy text?

```r
text <- c("Oh! Get me away from here, I'm dying",
          "Play me a song to set me free",
          "Nobody writes them like they used to",
          "So it may as well be me",
          "Here on my own now after hours",
          "Here on my own now on a bus",
          "Think of it this way",
          "You could either be successful or be us",
          "With our winning smiles, and us",
          "With our catchy tunes or worse",
          "Now we're photogenic",
          "You know, we don't stand a chance")

text
```

```
##  [1] "Oh! Get me away from here, I'm dying"
##  [2] "Play me a song to set me free"
##  [3] "Nobody writes them like they used to"
##  [4] "So it may as well be me"
##  [5] "Here on my own now after hours"
##  [6] "Here on my own now on a bus"
##  [7] "Think of it this way"
##  [8] "You could either be successful or be us"
##  [9] "With our winning smiles, and us"
## [10] "With our catchy tunes or worse"
## [11] "Now we're photogenic"
## [12] "You know, we don't stand a chance"
```

What is tidy text?

```r
text_df <- tibble(line = 1:12, text = text)

text_df %>% print(n = 12)
```

```
## # A tibble: 12 × 2
##     line text
##    <int> <chr>
##  1     1 Oh! Get me away from here, I'm dying
##  2     2 Play me a song to set me free
##  3     3 Nobody writes them like they used to
##  4     4 So it may as well be me
##  5     5 Here on my own now after hours
##  6     6 Here on my own now on a bus
##  7     7 Think of it this way
##  8     8 You could either be successful or be us
##  9     9 With our winning smiles, and us
## 10    10 With our catchy tunes or worse
## 11    11 Now we're photogenic
## 12    12 You know, we don't stand a chance
```

What is tidy text?

```
text_df %>%
  unnest_tokens(word, text) %>%
  print(n = 10)
```

```
## # A tibble: 80 × 2
##     line word
##    <int> <chr>
##  1     1 oh
##  2     1 get
##  3     1 me
##  4     1 away
##  5     1 from
##  6     1 here
##  7     1 i'm
##  8     1 dying
##  9     2 play
## 10     2 me
## # … with 70 more rows
```

Case study: FM's COVID-19 speeches

github.com/mine-cetinkaya-rundel/fm-speeches-covid19