

# Time Series Stylometry on the #TravelBan News Writers

*Kanishka Misra*

*April 20, 2018*

## Setup

loading libraries

```
library(tidyverse)
library(jsonlite)
library(tidyjson)
library(widyr)
library(tidytext)
library(ggthemes)
library(lubridate)

options(scipen = 99)
```

## Loading data

```
news_articles <- fromJSON("data/1996_dicts_9_sources.json", simplifyDataFrame = T) %>%
  as.tibble()
```

news\_articles

```
## # A tibble: 1,996 x 10
##   title url text top_words summary source publish_date authors id
## * <chr> <chr> <chr> <list> <chr> <chr> <chr> <list> <chr>
## 1 Appea~ http~ "The ~ <chr [13~ "The 1~ cnn 2017-05-25 <chr [~ 1
## 2 Elon ~ http~ "Tesl~ <chr [16~ "Tesla~ cnn 2017-01-30 <chr [~ 2
## 3 Donal~ http~ "J.P.~ <chr [14~ "J.P. ~ cnn 2017-01-30 <chr [~ 8
## 4 Donal~ http~ "The ~ <chr [14~ "Meanw~ cnn 2017-02-06 <chr [~ 9
## 5 More ~ http~ "Majo~ <chr [18~ "On Tu~ cnn 2017-02-07 <chr [~ 10
## 6 Presi~ http~ "Pres~ <chr [14~ "Presi~ cnn 2017-02-21 <chr [~ 12
## 7 Eliza~ http~ "Know~ <chr [12~ "\"Turn~ cnn 2017-03-16 <chr [~ 13
## 8 Tech ~ http~ "The ~ <chr [14~ "The i~ cnn 2017-01-28 <chr [~ 14
## 9 Here'~ http~ "Afte~ <chr [14~ "After~ cnn 2017-01-28 <chr [~ 15
## 10 GE's ~ http~ "Gene~ <chr [16~ "Gener~ cnn 2017-01-29 <chr [~ 16
## # ... with 1,986 more rows, and 1 more variable: type <chr>
```

## Tidying and Nesting

1. Count authors, only keep single author news articles.
2. Remove duplicates, and keep only text based duplicates if source is different.
3. Remove columns that are not relevant to the scope of this project.
4. Nest based on author, store article count.
5. Select top 10 (or more if similar number of articles exist) authors based on article count.

```
nested_articles <- news_articles %>%
  distinct(text, source, .keep_all = T) %>%
  mutate(length = map_int(authors, length)) %>%
  filter(length < 2) %>%
  unnest(authors) %>%
  select(text, author = authors, publish_date, source) %>%
  group_by(author) %>%
  nest() %>%
  mutate(article_count = map_int(data, nrow)) %>%
  top_n(10, article_count)
```

```
nested_articles
```

```
## # A tibble: 12 x 3
##   author          data          article_count
##   <chr>          <list>          <int>
## 1 laura jarrett  <tibble [31 x 3]>      31
## 2 ariane de vogue <tibble [15 x 3]>      15
## 3 mark sherman   <tibble [15 x 3]>      15
## 4 adam liptak    <tibble [26 x 3]>      26
## 5 editorial board <tibble [23 x 3]>      23
## 6 matt zapotosky <tibble [17 x 3]>      17
## 7 michael a. memoli <tibble [15 x 3]>      15
## 8 maura dolan    <tibble [15 x 3]>      15
## 9 david savage    <tibble [15 x 3]>      15
## 10 jaweed kaleem  <tibble [62 x 3]>      62
## 11 david g. savage <tibble [16 x 3]>      16
## 12 brian bennett  <tibble [32 x 3]>      32
```

We now have a tibble of the 12 authors with the most number of articles. We can then write a generic function that maps over each author and then calculates the distances of each authors news articles from their first article during the travel ban proposal.

```
temporal_delta <- function(df) {
  df %>%
    mutate(
      text = str_replace_all(tolower(text), paste0(source, "|story highlights"), "")
    ) %>%
    select(-source) %>%
    group_by(publish_date) %>%
    unnest_tokens(word, text) %>%
    count(word) %>%
    ungroup() %>%
    mutate(
      publish_date = lubridate::ymd(publish_date)
    ) %>%
    pairwise_delta(publish_date, word, n, upper = F) %>%
    filter(item1 == min(item1)) %>%
    transmute(
      publish_date = item2,
      delta
    )
}

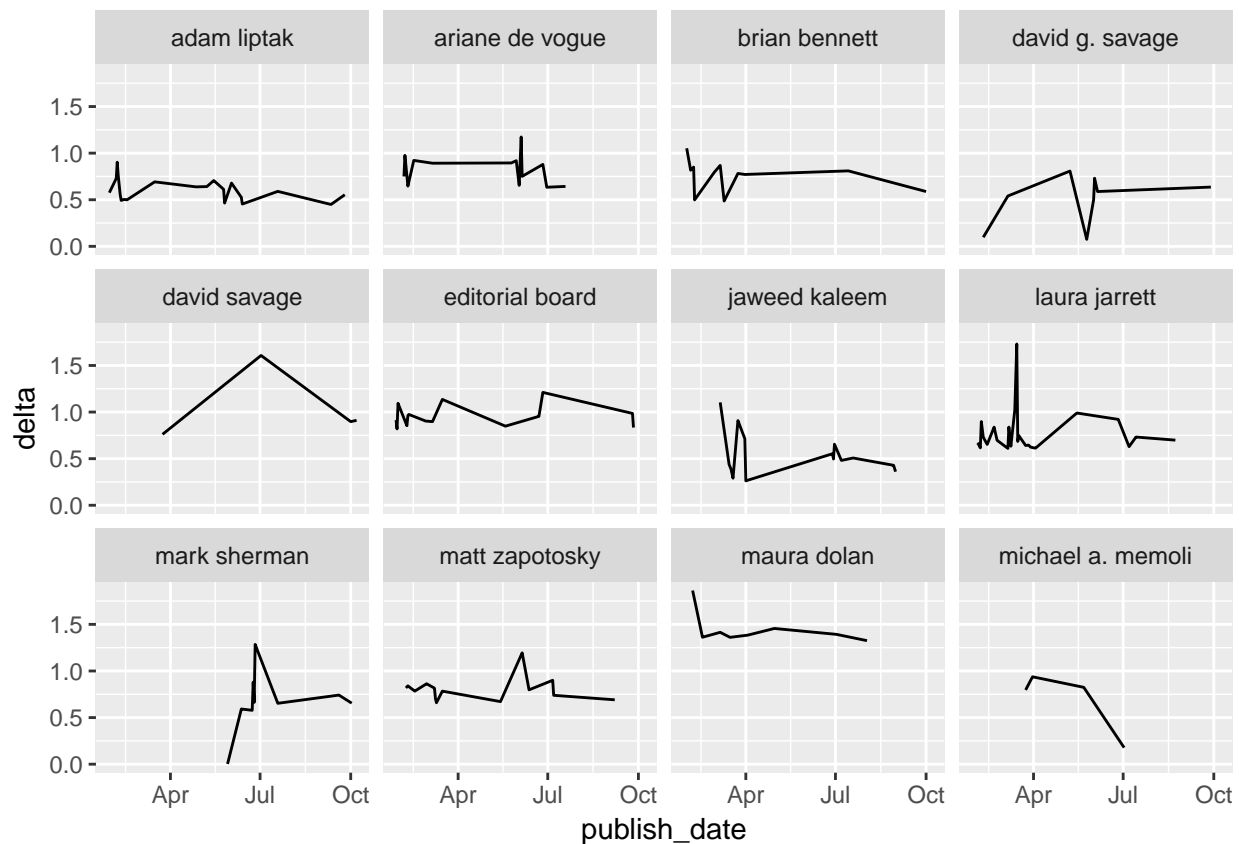
temporal_delta(nested_articles[1, ]$data[[1]])
```

```
## # A tibble: 24 x 2
##   publish_date delta
##   <date>         <dbl>
## 1 2017-02-03     0.668
## 2 2017-02-04     0.658
## 3 2017-02-06     0.616
## 4 2017-02-07     0.898
## 5 2017-02-09     0.729
## 6 2017-02-13     0.653
## 7 2017-02-20     0.838
## 8 2017-02-23     0.697
## 9 2017-03-06     0.610
## 10 2017-03-07    0.839
## # ... with 14 more rows
```

If we map this to all writer's data, we get the following plot:

```
author_deltas <- nested_articles %>%
  mutate(ts_delta = map(data, temporal_delta)) %>%
  unnest(ts_delta)

author_deltas %>%
  ggplot(aes(publish_date, delta)) +
  geom_line() +
  facet_wrap(~author)
```



Help me.