Like a lot of people, It was intrigued by [“I Am Part of the Resistance Inside the Trump Administration”](https://www.nytimes.com/2018/09/05/opinion/trump-white-house-anonymous-resistance.html), an anonymous New York Times op-ed written by a “senior official in the Trump administration”. And like many data scientists, I was curious about what role text mining could play.

Ok NLP people, now’s your chance to shine. Just spitballing here but TF-IDF on “the op-ed” compared to the published writing of every senior Trump admin official? I want likelihood estimates with standard errors. GO!

— Drew Conway (@drewconway) [September 5, 2018](https://twitter.com/drewconway/status/1037475836042117120?ref_src=twsrc%5Etfw)

This is a useful opportunity to demonstrate how to use the [tidytext package](https://www.tidytextmining.com/) that Julia Silge and Its developed, and in particular to apply three methods:

* Using TF-IDF to find words specific to each document (examined in more detail in [Chapter 3 of our book](https://www.tidytextmining.com/tfidf.html))
* Using widyr to compute pairwise cosine similarity
* How to make similarity interpretable by breaking it down by word

Since my goal is R education more than it is political analysis, I show all the code in the post.

**Downloading data**

Getting the text of the op-ed is doable with the rvest R package.

# setup

library(tidyverse)

library(tidytext)

library(rvest)

theme\_set(theme\_light())

url <- "https://www.nytimes.com/2018/09/05/opinion/trump-white-house-anonymous-resistance.html"

# tail(-1) removes the first paragraph, which is an editorial header

op\_ed <- read\_html(url) %>%

html\_nodes(".e2kc3sl0") %>%

html\_text() %>%

tail(-1) %>%

data\_frame(text = .)

The harder step is getting a set of documents representing “senior officials”. An imperfect but fast approach is to collect text from their Twitter accounts. (If you find an interesting dataset of, say, government FOIA documents, I recommend you try extending this analysis!)

We can look at a combination of two (overlapping) Twitter lists containing administration staff members:

* [CSPAN’s list of Cabinet accounts](https://twitter.com/cspan/lists/the-cabinet)
* [digiphile’s list of White House staff](https://twitter.com/cspan/lists/digiphile)

library(rtweet)

cabinet\_accounts <- lists\_members(owner\_user = "CSPAN", slug = "the-cabinet")

staff <- lists\_members(slug = "white-house-staff", owner = "digiphile")

# Find unique screen names from either account

accounts <- unique(c(cabinet\_accounts$screen\_name, staff$screen\_name))

# Download ~3200 from each account

tweets <- map\_df(accounts, get\_timeline, n = 3200)

This results in a set of 136,501 from 69 Twitter handles. There’s certainly no guarantee that the op-ed writer is among these Twitter accounts (or, if they are, that they even write their tweets themselves). But it still serves as an interesting case study of text analysis. How do we find the tweets with the closest use of language?

**Tokenizing tweets**

First, we need to tokenize the tweets: to turn them from full messages into individual words. We probably want to avoid retweets, and we need to use a custom regular expression for splitting it and remove links (just as I’d done when [analyzing Trump’s Twitter account](http://varianceexplained.org/r/trump-tweets/)).

# When multiple tweets across accounts are identical (common in government

# accounts), use distinct() to keep only the earliest

reg <- "([^A-Za-z\\d#@']|'(?![A-Za-z\\d#@]))"

tweet\_words <- tweets %>%

filter(!is\_retweet) %>%

arrange(created\_at) %>%

distinct(text, .keep\_all = TRUE) %>%

select(screen\_name, status\_id, text) %>%

mutate(text = str\_replace\_all(text, "https?://t.co/[A-Za-z\\d]+|&", "")) %>%

unnest\_tokens(word, text, token = "regex", pattern = reg) %>%

filter(str\_detect(word, "[a-z]"))

This parses the corpus of tweets into almost 1.5 million words.

Among this population of accounts, and ignoring “stop words” like “the” and “of”, what are the most common words? We can use ggplot2 to visualize this.

tweet\_words %>%

filter(!word %in% stop\_words$word) %>%

count(word, sort = TRUE) %>%

head(16) %>%

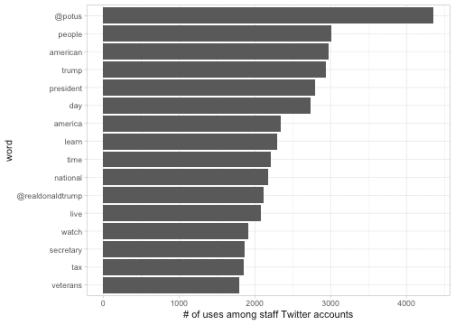
mutate(word = reorder(word, n)) %>%

ggplot(aes(word, n)) +

geom\_col() +

coord\_flip() +

labs(y = "# of uses among staff Twitter accounts")



No real surprises here. Many accounts mention @POTUS often, as well as words like “people”, “American”, and “Trump” that you’d expect from administration accounts.

**Finding a text signature: TF-IDF vectors**

What words make up someone’s “signature”? What make up mine, or Trump’s, or Mike Pence’s, or the op-ed’s?

We could start with the most common words someone uses. But there are some words, like “the” and “of” that just about everyone uses, as well as words like “President” that everyone in our dataset will use. So we also want to downweight words that appear across many documents. A common tool for balancing these two considerations and turning them into a “signature” vector is [tf-idf](https://en.wikipedia.org/wiki/Tf%E2%80%93idf): **term-frequency inverse-document-frequency**. This takes how frequently someone uses a term, but divides it by (the log of) how many documents mention it. For more details, see [Chapter 3 of Text Mining with R](https://www.tidytextmining.com/tfidf.html).

The bind\_tf\_idf function from tidytext lets us compute tf-idf on a dataset of word counts like this. Before we do, we bring in the op-ed as an additional document (since we’re interesting in considering it as one “special” document in our corpus).

# Combine in the op\_ed wordswith the name "OP-ED"

op\_ed\_words <- op\_ed %>%

unnest\_tokens(word, text) %>%

count(word)

word\_counts <- tweet\_words %>%

count(screen\_name, word, sort = TRUE) %>%

bind\_rows(op\_ed\_words %>% mutate(screen\_name = "OP-ED"))

# Compute TF-IDF using "word" as term and "screen\_name" as document

word\_tf\_idf <- word\_counts %>%

bind\_tf\_idf(word, screen\_name, n) %>%

arrange(desc(tf\_idf))

word\_tf\_idf

## # A tibble: 226,410 x 6

## screen\_name word n tf idf tf\_idf

##

## 1 JoshPaciorek #gogreen 170 0.0204 4.25 0.0868

## 2 USTradeRep ustr 147 0.0213 3.56 0.0757

## 3 DeptVetAffairs #vantagepoint 762 0.0173 3.56 0.0614

## 4 DeptofDefense #knowyourmil 800 0.0185 3.15 0.0584

## 5 DanScavino #trumptrain 655 0.0201 2.86 0.0575

## 6 USUN @ambassadorpower 580 0.0154 3.56 0.0548

## 7 USTreasury lew 690 0.0183 2.86 0.0523

## 8 HUDgov hud 566 0.0196 2.30 0.0451

## 9 OMBPress omb 38 0.0228 1.95 0.0444

## 10 SecElaineChao 'can 1 0.00990 4.25 0.0421

## # ... with 226,400 more rows

We can now see the words with the strongest associations to a user. For example, [Josh Paciorek](https://twitter.com/joshpaciorek?lang=en) (the VP’s Deputy Press Secretary) uses the hashtag #gogreen (supporting Michigan State Football) quite often; it makes up 2% of the words (tf, term frequency). Since no one else uses it (leading to an inverse document frequency, idf, of 4.5), this makes it a critical part of his TF-IDF vector (his “signature”).

We could take a look at the “signatures” of a few selected Twitter accounts.

library(drlib)

selected <- c("realDonaldTrump", "mike\_pence", "DeptVetAffairs", "KellyannePolls")

word\_tf\_idf %>%

filter(screen\_name %in% selected) %>%

group\_by(screen\_name) %>%

top\_n(12, tf\_idf) %>%

ungroup() %>%

mutate(word = reorder\_within(word, tf\_idf, screen\_name)) %>%

ggplot(aes(word, tf\_idf, fill = screen\_name)) +

geom\_col(show.legend = FALSE) +

scale\_x\_reordered() +

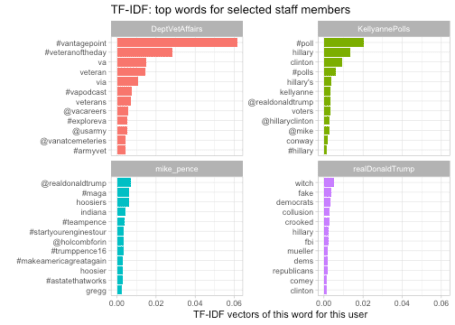
coord\_flip() +

facet\_wrap(~ screen\_name, scales = "free\_y") +

labs(x = "",

y = "TF-IDF vectors of this word for this user",

title = "TF-IDF: top words for selected staff members")



This gives us a set of words that are quite specific to each account. For instance, @DeptVetAffairs uses hashtags like “#vantagepoint” and “#veteranoftheday” that almost no other account in this set would use. Words that are specific to Trump include “witch” (as in “witch hunt”), “fake” (as in “fake news”) and other phrases that he tends to fixate on while other government officials don’t. (See [here](http://varianceexplained.org/r/trump-followup/) for my text analysis of Trump’s tweets as of August 2017).

This shows how TF-IDF offers us a vector (an association of each word with a number) that describes the unique signature of that document. To compare our documents (the op-ed with each Twitter account), we’ll be comparing those vectors.

**The widyr package: cosine similarity**

How can we compare two vectors to get a measure of document similarity? There are many approaches, but perhaps the most common for comparing TF-IDF vectors is [cosine similarity](https://en.wikipedia.org/wiki/Cosine_similarity). This is a combination of a dot product (multiplying the same term in document X and document Y together) and a normalization (dividing by the magnitudes of the vectors).

The widyr package offers a convenient way to compute pairwise similarities on a tidy dataset:

library(widyr)

# Find similarities between screen names

# upper = FALSE specifies that we don't want both A-B and B-A matches

word\_tf\_idf %>%

pairwise\_similarity(screen\_name, word, tf\_idf, upper = FALSE, sort = TRUE)

## # A tibble: 2,415 x 3

## item1 item2 similarity

##

## 1 VPComDir VPPressSec 0.582

## 2 DeptVetAffairs SecShulkin 0.548

## 3 ENERGY SecretaryPerry 0.472

## 4 HUDgov SecretaryCarson 0.440

## 5 usedgov BetsyDeVosED 0.417

## 6 Interior SecretaryZinke 0.386

## 7 SecPriceMD SecAzar 0.381

## 8 SecPompeo StateDept 0.347

## 9 VPPressSec VP 0.343

## 10 USDA SecretarySonny 0.337

## # ... with 2,405 more rows

The top results show that this elementary method is able to match people to their positions. The VP Press Secretary and VP Communications Director unsurprisingly work closely together and tweet on similar topics. Similarly, it matches Shulkin, Perry, Carson, DeVos, and Zinke to their (current or former) cabinet positions, and links the two consecutive Health and Human Services directors (Price and Azar) to each other.

It’s worth seeing this document similarity metric in action, but it’s not what you’re here for. We’re really excited about seeing comparisons between the *op-ed* and Twitter articles. We can

# Look only at the similarity of the op-ed to other documents

op\_ed\_similarity <- word\_tf\_idf %>%

pairwise\_similarity(screen\_name, word, tf\_idf, sort = TRUE) %>%

filter(item1 == "OP-ED")

library(drlib)

op\_ed\_similarity %>%

head(12) %>%

mutate(item2 = reorder(item2, similarity)) %>%

ggplot(aes(item2, similarity)) +

geom\_col() +

scale\_x\_reordered() +

coord\_flip() +

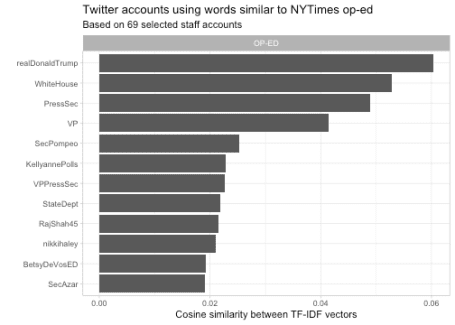
facet\_wrap(~ item1, scales = "free\_y") +

labs(x = "",

y = "Cosine similarity between TF-IDF vectors",

subtitle = "Based on 69 selected staff accounts",

title = "Twitter accounts using words similar to NYTimes op-ed")



This unveils the most similar writer as… Trump himself.

Hmmm. While that would certainly be a scoop, it doesn’t sound very likely to me. And the other top picks (the official White House account, the Press Secretary, and the Vice President) also seem like suspicious guesses.

**Interpreting machine learning: what words contributed to scores?**

The method of tf-idf is a fairly basic one for text mining, but as a result it has a useful trait: it’s based on a linear combination of one-score-per-word. This means we can say *exactly* how much each word contributed to a TF-IDF similarity between the article and a Twitter account. (Other machine learning methods allow *interactions* between words, which makes them harder to interpret).

We’ll try an approach of decomposing our TF-IDF similarity to see how much each . You could think of this as asking “if the op-ed hadn’t used this word, how much lower would the similarity score be?”

# This takes a little R judo, but it's worth the effort

# First we normalize the TF-IDF vector for each screen name,

# necessary for cosine similarity

tf\_idf <- word\_tf\_idf %>%

group\_by(screen\_name) %>%

mutate(normalized = tf\_idf / sqrt(sum(tf\_idf ^ 2))) %>%

ungroup()

# Then we join the op-ed words with the full corpus, and find

# the product of their TF-IDF with it in other documents

word\_combinations <- tf\_idf %>%

filter(screen\_name == "OP-ED") %>%

select(-screen\_name) %>%

inner\_join(tf\_idf, by = "word", suffix = c("\_oped", "\_twitter")) %>%

filter(screen\_name != "OP-ED") %>%

mutate(contribution = normalized\_oped \* normalized\_twitter) %>%

arrange(desc(contribution)) %>%

select(screen\_name, word, tf\_idf\_oped, tf\_idf\_twitter, contribution)

# Get the scores from the six most similar

word\_combinations %>%

filter(screen\_name %in% head(op\_ed\_similarity$item2)) %>%

mutate(screen\_name = reorder(screen\_name, -contribution, sum),

word = reorder\_within(word, contribution, screen\_name)) %>%

group\_by(screen\_name) %>%

top\_n(12, contribution) %>%

ungroup() %>%

mutate(word = reorder\_within(word, contribution, screen\_name)) %>%

ggplot(aes(word, contribution, fill = screen\_name)) +

geom\_col(show.legend = FALSE) +

scale\_x\_reordered() +

facet\_wrap(~ screen\_name, scales = "free\_y") +

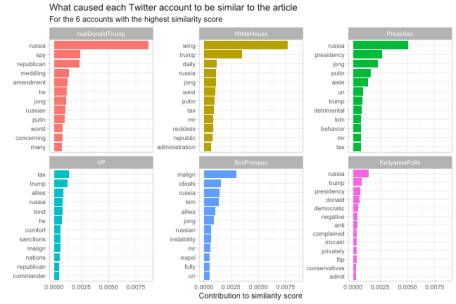
coord\_flip() +

labs(x = "",

y = "Contribution to similarity score",

title = "What caused each Twitter account to be similar to the article",

subtitle = "For the 6 accounts with the highest similarity score")



Now the reasons for the TF-IDF similarities become clearer.

The op-ed uses the words “Russia” five times. The Press Secretary and *especially* Trump mention Russia multiple times on their Twitter accounts, always within the context of defending Trump (as expected). Several accounts also get a high score because they mention the word “Trump” so frequently.

Unfortunately, with a document this short and topical, that’s all it takes to get a high similarity score (a bag of words method can’t understand the context, such as mentioning Russia in a negative or a defensive context). This is one reason it’s worth taking a closer look at what goes into an algorithm,

Having said that, there’s one signature I think is notable.

**“Malign behavior”**

[Many others have noted “lodestar” as a telltale word in the piece](https://www.huffingtonpost.com/entry/lodestar-mike-pence-anonymous-new-york-times_us_5b905dd5e4b0511db3dec1e1?section=politics&ncid=tweetlnkushpmg00000016). None of the relevant documents included that. I’d like to focus on another word that did: **malign**. Emphasis mine:

He complained for weeks about senior staff members letting him get boxed into further confrontation with Russia, and he expressed frustration that the United States continued to impose sanctions on the country for its **malign** behavior.

“Malign” isn’t as rare a word as “lodestar”, but it’s notable for being used in the exact same context (discussing Russia or other countries’ behavior) in a number of tweets from both Secretary of State Pompeo and the @StateDepartment account. (Pompeo has actually used the term “malign” an impressive [seven times since May](https://twitter.com/search?f=tweets&vertical=default&q=from%3Asecpompeo%20malign&src=typd), though all but one were about Iran rather than Russia).

In [#Finland](https://twitter.com/hashtag/Finland?src=hash&ref_src=twsrc%5Etfw) tonight. Monday, [@POTUS](https://twitter.com/POTUS?ref_src=twsrc%5Etfw) & I will meet with our Russian counterparts in [#Helsinki](https://twitter.com/hashtag/Helsinki?src=hash&ref_src=twsrc%5Etfw) A better relationship with the Russian government would benefit all, but the ball is in Russia’s court. We will continue to hold Russia responsible for its malign activities [@StateDept](https://twitter.com/StateDept?ref_src=twsrc%5Etfw) [pic.twitter.com/K4C1PSrRMb](https://t.co/K4C1PSrRMb)

— Secretary Pompeo (@SecPompeo) [July 15, 2018](https://twitter.com/SecPompeo/status/1018608807612690432?ref_src=twsrc%5Etfw)

“Malign behavior” has been common language for Pompeo this whole year, as it has for other State Department officials like [Jon Huntsman](https://www.reuters.com/article/us-usa-russia-summit/trump-will-focus-on-russias-malign-activity-at-summits-us-officials-idUSKBN1JV2AO). What’s more, you don’t need data science to notice the letter spends three paragraphs on foreign policy (and praises “the rest of the administration” on that front).  
I’m not a pundit or a political journalist, but I can’t resist speculating a bit. Pompeo is named by the [Weekly Standard](https://www.weeklystandard.com/michael-warren/these-are-the-four-people-most-likely-to-be-behind-the-anonymous-new-york-times-op-ed-from-the-resistance-inside-the-trump-administration) as one of four likely authors of the op-ed, but even if he’s not the author my guess would be **someone in the State Department.**

**Conclusion: Opening the black box**

It’s worth emphasizing again that this article is just my guess based on a single piece of language (it’s nowhere close to the certainty of my [analysis of Trump’s Twitter account](http://varianceexplained.org/r/trump-tweets/) during the campaign, which was statistically significant enough that I’d be willing to consider it “proof”).

I was fairly skeptical from the start that we could get strong results with document-comparison methods like this, especially on such a small article. That opinion mirrored people with much more expertise than I have:

That means you need a pretty large sample to not have large error bars. Don’t expect conclusive or even suggestive evidence here.

— David Mimno (@dmimno) [September 6, 2018](https://twitter.com/dmimno/status/1037490608015925248?ref_src=twsrc%5Etfw)

But I’m satisfied with this analysis both as a demonstration of tidytext methods and one on the importance of model interpretability. When we ran a TF-IDF comparison, we knew it was wrong because @realDonaldTrump appeared at the top. But what if Trump *hadn’t* been the one to mention Russia the most, or if another false positive had caused an account to rise to the top? Breaking similarity scores down by word is a useful way to interrogate our model and understand its output.