EMIL HVITFELDT

(HTTPS://WWW.HVITFELD PAredicting

I love to develop and make things in R. Working on visualization styles, modeling techniques and general workflow problems.

- > Blog (https://www.hvitfeldt.me /blog/)
- > GenArt (https://www.hvitfeldt.me /GenArt/)
- > About (https://www.hvitfeldt.me /about/)
- > ERAN (https://www.hvitfeldt.me /src/contrib/)







/Emil_Hvitfeldt) /in/emilhvitfeldt/) (https://github.com

©2019 Emil Hvitfeldt /E侧dhnitfelda)dt | Template by Bootstrapious.com (https://bootstrapious.com/freetemplates) & ported to Hugo by Kishan B (https://github.com /kishaningithub)

authorship in The Federalist Papers with tidytext

- > Overview
- > The Federalist Papers
- > Libraries
- > Data
- > Predictive modeling
- > Working showcase

Overview

In this post we will

- > talk about The Federalist Papers
- > access and tidy the text using the tidytext package
- (https://(withthooghithhsty/thekkt@kedihccom) apply our model to the data to predict the author of the disputed papers

The Federalist Papers

In the early days of The United States of America around the time when the Constitution was being signed did a series of articles published in various newspapers. These papers where writing under the false name Publius. It was later revealed to have been the collected works of Alexander Hamilton, James

EMIMadiyditffffcLDffn Jay. (HTTPS://WWW.HVITFELDT.ME/)

The Interesting thing in this was that the authorship of I love to develop and make these papers were not consistent. In This is where we things in R. Working on come in, in this blog post will we try to see if we are visualization styles, modeling techniques and general

workflow problems.

If you would like to read more about this story including

- > Blogast attempts to solve this problem please read How
 (hម្រាប់ទេស្តេស្ត្រាស្ត្រស្នេង Planting Year-Old Mystery About
 /bស្ត្រីខ្លាំង American (https://priceonomics.com/how-
- > GenArt > GenArt (https://www.hvitfeldt.me /GenArt/)

Ablutihraries

(https://www.hvitfeldt.me

/about/ill start by loading the libraries which includes

> ERANnet that will be used to construct the predictive (https://www.hvitfeldt.me

/src/contrib/)

library(tidyverse)

ibrary(tidytext)
Frary(gutenbergr)

(https://twiaidiocemilps://wewt@gkedihccom)

/Emil Prince (in/emilhvitfeldt/)

©2019 Emil Hvitfeldt

/E面间Mnitfeldt)dt | Template by
We are lucky today because all of The Federalist
Bootstrapious.com

(https://pers.happens.to/be on gutenberg

templates) & ported to Hugo by Kishan B (https://github.com /kishaningithub)

```
EMIL HVITFELDT
papers <- gutenberg download(1404)
(HTTPS://WWW.HVITFELDT.ME/)
head(papers, n = 10)
```

| ## # A tibble: 10 x 2 things in ## Working on text visualiza##n styles, modeling<int> <chr> 1404 THE FEDERALIST PAPE techniqu##andlgeneral workflow##obl2ms. 1404 1404 By Alexander Hamilt ## 3 > Blog ## 4 (http##/wwww.hvitfeldt.mle404 /blog## 1404 1404 ""

> GenAft / 1404 ""

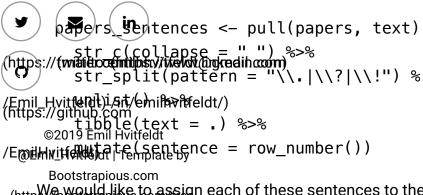
(https://www.hvitfeldt.me / 1404 ""

/GenArt/) 9 1404 ""

10 1404 General Introductio

> About

(https://www.divitivelehoodeling we are going to do later, I /awwwild) like to divide each paper up into sentences. This > ERispa rather complicated affair, but I will take a rather ad (https://www.ch/vitivettyvitle/be good enough for the purpose /set/thisppss). We will do this by collapsing all the lines together and splitting them by .,! and ?'s.



(https://pwouldplike.to.assign each of these sentences to the temploreesponding anticleynumber and author. Thus we will Kishasingithub) a group for the papers of interest.

```
I love to develop and make (2.5, 64)
things in R. Working on c(49:58, 62:63)
visualization styles, modeling
Next we will simple look for lines that include
techniques and general
workflow properties T No" as they would indicate the start of a
     paper and then label them accordingly.
> Blog
   (https://www.hwitfoldt.ene papers sentences %>%
  /blog/) mutate(no = cumsum(str detect(text,
> GenArt
   (https://www.stviitedesne(word, text) %>%
   /GenArtyutate(author = case_when(no %in% ha
                                               no %in% ma
> About
                                               no %in% ja
   (https://www.hvitfeldt.me
                                               no %in% un
   /about/)
                    id = paste(no, sentence, sep
> ERAN
   (htetos://a/wevavohuittkeboblumtebefore we move on
   /src/contrib/)
        papers_words %>%
          count(author)
              A Mibble: 4 x 2
                        tkekkt@gkedihccom)
/Emil_H##teldt)hanvertionitfeld(6)88
(https://aithub.com
                              8540
                            45073
/Emilihvitaling of is an experience by
                             24471
       B##tsflapliblik.00Wn
 (https://bootstrapious.com/free-
 tem Meeses thated have this drop post as many articles as the
  Kishther (two side it learners o we will exclude him from
     further analysis
        papers_words <- papers_words %>%
          filter(author != "jay")
```

(HT Predictive modeling

I love to developting predictive model we will use the termthings in R. Working on I frequency matrix as our input and as the response will visualization styles, modeling be an indicator that Hamilton wrote the paper. For this techniques and general modeling we will use the glmnet package which fits a generalized linear model via penalized maximum

> Bldikelihood. It is quite fast and works great with sparse (https://www.hybergetheterm-frequency matrix.

/blog/)

The response is set to the binomial family because of GenArt
the binary nature of the response (did Hamilton write (https://www.hvitfeldt.me the sentence).
/GenArt/)

- > About two get the term-frequency matrix with the cast_ (https://www.hyitfeldt.me /about/)
- > ERAN papers_dtm <- papers_words %>% (https://www.hvitfeldt.me count(id, word, sort = TRUE) %>% /src/contrib()_dtm(id, word, n)

We will need to define a response variable, which we will with an indicator for our training set which will be the articles with https://(withdoor) known authors.

/Emil_Hvitfeldt) /in/emilhvitfeldt/) (https://github.com

@ 例母 每 min Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper / Emil Hvitter a. frame (id = dimnames (paper a. frame (id = dimnames (paper a. frame (id = dimnames (paper a

We will use cross-validation to obtain the best value of the models tuning parameter. This part takes a couple of minutes.

```
papers dtm[meta$train, ]
                      meta$y[meta$tŕáin]
I love to develop and make
things in R. Working on cv.glmnet(predictor, response
visualization styles, modeling
techniques and general model, we will add the predicted
workfyalues being meta data frame.
> Blog meta <- meta %>%
  (https://www.withevitherlettme predict(model, newx =
  /blog/)
                                       s = model$lamb
> GenArt
        now time to visualize the results. First we will look://www.hvitfeldt.me
  at how the training set have been separated.
> About
  (https://ww
           ilter(train) %>%
  /about/
          ggplot(aes(factor(no), pred)) +
> ERAN
          geom_boxplot(aes(fill = author)) +
  (https://www.hvitfeldt.me() +
  /src/contrib/y = "predicted probability",
                x = "Article number") +
          theme(legend.position = "top") +
         ♥cale<sup>in</sup>fill manual(values = c("#30489
                ndpsy/wewww.act.x = element_text(ang
/Emil_Hvitfeldt) /in/emilhvitfeldt/) hamilton 🗯 madison
(https://github.com
  Kiskan B (https://github.com
       /kishaningithub
      0.00
                              Article number
```

The box plot of predicted probabilities, one value for

EMILE ACTIVE THE FORTH THE FORTH BY BABILITY

represents the extent to which the model believe the I love to develop and make sentence was written by Alexander Hamilton. things in R. Working on

visualization styles, modeling techniques and general model can settle the dispute of the 12 workfpapproblems will plot the predicted probabilities of the unknown papers alongside the training set.

> Blog (http

(httpm://wwwy.byitfeldt.me /blog/)ggplot(aes(factor(no), pred)) +

> GenArt geom_boxplot(aes(fill = author)) +

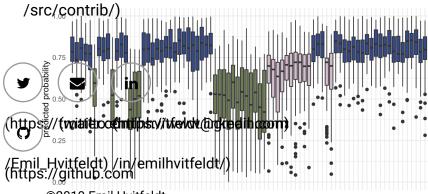
(https://twown_ntiedimanle() +

/GenArt/ bs(y = "predicted probability",

theme(legend.position = "top") +
(https://www.hvitfeldt.me
 scale_fill_manual(values = c("#30489
/about/)
 theme(axis.text.x = element text(ang)

> ERAN

 $(https://www.hvitfeldthame \ {}_{\text{hamilton}} \ {}_{\text{madison}} \ {}_{\text{pi}} \ {}_{\text{unknown}}$



©2019 Emil Hvitfeldt

/Eppil州Witfeldeldt | Template by

Bootstrapious.com

(http://publicalphat.the predicted probabilities don't quite temphates upredicted probabilities upredicted probabil

Kishan B (https://github.com/variety of different reasons. One of /kishaningithub) them could be that Madison wrote them and Hamilton edited them.

Despite the unsuccessful attempt to predict the secret author we still managed to showcase the method

EMINHOWN Free Leng unsuccessful in this case could (HT Treste Length House Length House Length Land Length Land Length Le

```
I love to develop and make
things Working showcase
visualization styles, modeling
techniques and general
Since the method proved unsuccessful in determining workflow problems.
    the secret author did I decide to add an example where
> Blod authorship is know. We will use the same data from
  (hearlier only look at known Hamilton and Madison
  /bpapers, train on some of them and show that the
     algorithm is able to detect the authorship of the other.
  (https://www.hvitfeldt.me
  /Genpapers_dtm <- papers_words %>%
          filter(author != "unknown") %>%
  count(id, word, sort = TRUE) %>% (https://www.hvitfeldt.me cast_dtm(id, word, n) /about/)
> About
> ERActe we let the first 16 papers that they wrote be the
  (https://www.dhtviefeebstt.loneetraining set.
  /src/contrib/)
       meta <- data.frame(id = dimnames(paper</pre>
          left_join(papers_words[!duplicated(p
         Ήψtatሆ(y = as.numeric(author == "ham
(https://(withithocthings.v/novorming.com)
       ## Warning: Colŭmn `id`´joining factor
/Emil_H##feidt)a/fia/@failhvivf@idt/)r
(https://github.com
      ©2019 Emil Hvitfeldt
/Emilinvitantie tempfate by apers_dtm[meta$train, ]
      Bbetspaphose.com meta$y[meta$train]
 (https://bootstrapious.com/free-
 templates)&ported to Hugo hymnet(predictor, response
  Kishan B (https://github.com
       /kishaningithub)
       meta <- meta %>%
          mutate(pred = predict(model, newx =
                                       s = model slamb
```

```
EMIL HYLTFELDT
| geom_boxplot(aes(fill = author)) +
things in R. Working on the things in R. Working on
visualization aps ynodelingredicted probability",
techniques and gen&ra # "Article number") +
workflow prableme (legend.position = "top") +
          scale fill manual(values = c("#30489)
          theme(axis.text.x = element text(ang
> Blog
  (https://www.hwltfeldf.ares(xintercept = 16.5), c
  /blog/)
> GenArt
  (https://ww
> Abœut
> ERAN
  (https://www.hvitfeldt.me
  /src/contrib/)
                              Article numbe
     So we see that while it isn't as crystal clear what what
    the test set predictions are giving us, they still give a
     pretty good indication.
(https://(withithocethhillps://thekat@gkedihccom)
(Emil_Hyjtfeldt) /in/emillovitfeldtb) utteranc.es
      ©2019 Emil Hvitfeldt
/Emile | Template by
      Bootstrapious.com
 (https://bootstrapious.com/free-
 templates) & ported to Hugo by
  Kishan B (https://github.com
       /kishaningithub)
       MJ Styling with Markdown is
                                      Sign in to comment
       supported
```

EMIL HVITFELDT (HTTPS://WWW.HVITFELDT.ME/)

I love to develop and make things in R. Working on visualization styles, modeling techniques and general workflow problems.

- > Blog
 (https://www.hvitfeldt.me
 /blog/)
- > GenArt
 (https://www.hvitfeldt.me
 /GenArt/)
- > About
 (https://www.hvitfeldt.me
 /about/)
- > ERAN
 (https://www.hvitfeldt.me
 /src/contrib/)







(https://(withtelocethnthlpstv/twevkt@gkedihccom)

/Emil_Hvitfeldt) /in/emilhvitfeldt/) (https://github.com

©2019 Emil Hvitfeldt /**EmilHvitfeldb)**dt | Template by Bootstrapious.com

(https://bootstrapious.com/freetemplates) & ported to Hugo by Kishan B (https://github.com /kishaningithub)