Title Goes Here

Jane Doe1,2,✉, John Q. Doe1, Peder Ås1, Juan Pérez3, and Max Mustermann

May 20, 2024

Abstract

Text of abstract

1 Formatting Open Science Group  
2 Federation of Planets  
3 Acme Corporation

✉ Correspondence: [Jane Doe <janedoe@fosg.org>](mailto:janedoe@fosg.org)

Keywords: keyword 1; keyword 2; keyword 3

Highlights: These are the highlights.

# Introduction

Here is a citation (Marwick, 2017)

https://annakrystalli.me/rrresearch/10\_compendium.html

# Background

Natural language processing, text analytics

https://github.com/datasciencedojo/IntroToTextAnalyticsWithR/tree/master

# Install all required packages.  
# install.packages(c("ggplot2",   
# "e1071",   
# "caret",   
# "quanteda",   
# "irlba",   
# "randomForest"))

# Load up the .CSV data and explore in RStudio.  
spam.raw <- read.csv("../data/raw\_data/spam.csv",   
 stringsAsFactors = FALSE,   
 fileEncoding = "UTF-16")  
  
# Clean up the data frame and view our handiwork.  
spam.raw <- spam.raw[, 1:2]  
names(spam.raw) <- c("Label", "Text")  
View(spam.raw)  
  
# Check data to see if there are missing values.  
length(which(!complete.cases(spam.raw)))

[1] 0

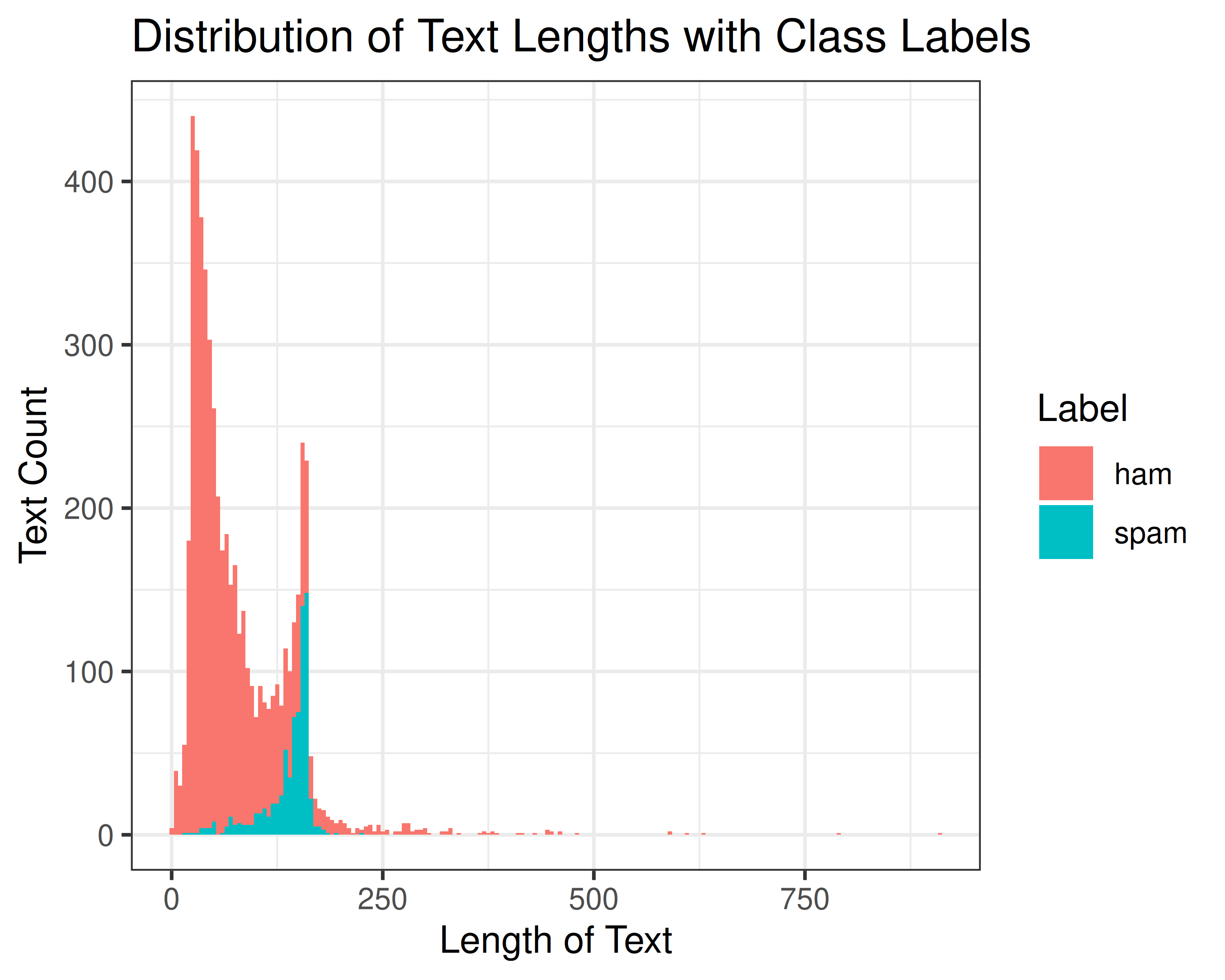
# Convert our class label into a factor.  
spam.raw$Label <- as.factor(spam.raw$Label)  
  
# The first step, as always, is to explore the data.  
# First, let's take a look at distibution of the class labels (i.e., ham vs. spam).  
prop.table(table(spam.raw$Label))

ham spam   
0.8659368 0.1340632

# Next up, let's get a feel for the distribution of text lengths of the SMS   
# messages by adding a new feature for the length of each message.  
spam.raw$TextLength <- nchar(spam.raw$Text)  
summary(spam.raw$TextLength)

Min. 1st Qu. Median Mean 3rd Qu. Max.   
 2.00 36.00 61.00 80.12 121.00 910.00

# Visualize distribution with ggplot2, adding segmentation for ham/spam.  
library(ggplot2)  
  
ggplot(spam.raw, aes(x = TextLength, fill = Label)) +  
 theme\_bw() +  
 geom\_histogram(binwidth = 5) +  
 labs(y = "Text Count", x = "Length of Text",  
 title = "Distribution of Text Lengths with Class Labels")



https://m-clark.github.io/text-analysis-with-R/

# Methods

bla

# Results

bla

# Note the path that we need to use to access our data files when rendering this document  
my\_data <- read.csv(here::here('analysis/data/raw\_data/my\_csv\_file.csv'))

plot(rnorm(10))

|  |
| --- |
| Figure 1: A plot of random numbers |

[Figure 1](#fig-demo-plot) shows how we can have a caption and cross-reference for a plot. Note that figure label and cross-references must both be prefixed with fig-

Here is an example of inline code 3.14 in the middle of a sentence.

# Discussion

# Conclusion

# Acknowledgements

# References

Marwick, B., 2017. Computational reproducibility in archaeological research: Basic principles and a case study of their implementation. Journal of Archaeological Method and Theory 24, 424–450. <https://doi.org/10.1007/s10816-015-9272-9>

### Colophon

This report was generated on 2024-05-20 18:22:17.424868 using the following computational environment and dependencies:

# which R packages and versions?  
if ("devtools" %in% installed.packages()) devtools::session\_info()

─ Session info ───────────────────────────────────────────────────────────────  
 setting value  
 version R version 4.3.1 (2023-06-16)  
 os Ubuntu 23.10  
 system x86\_64, linux-gnu  
 ui X11  
 language (EN)  
 collate en\_US.UTF-8  
 ctype en\_US.UTF-8  
 tz Europe/Madrid  
 date 2024-05-20  
 pandoc 3.1.1 @ /usr/lib/rstudio/resources/app/bin/quarto/bin/tools/ (via rmarkdown)  
  
─ Packages ───────────────────────────────────────────────────────────────────  
 package \* version date (UTC) lib source  
 cachem 1.1.0 2024-05-16 [1] CRAN (R 4.3.1)  
 cli 3.6.2 2023-12-11 [1] CRAN (R 4.3.1)  
 colorspace 2.1-0 2023-01-23 [1] CRAN (R 4.3.1)  
 devtools 2.4.5 2022-10-11 [1] CRAN (R 4.3.1)  
 digest 0.6.35 2024-03-11 [1] CRAN (R 4.3.1)  
 dplyr 1.1.4 2023-11-17 [1] CRAN (R 4.3.1)  
 ellipsis 0.3.2 2021-04-29 [1] CRAN (R 4.3.1)  
 evaluate 0.23 2023-11-01 [1] CRAN (R 4.3.1)  
 fansi 1.0.6 2023-12-08 [1] CRAN (R 4.3.1)  
 farver 2.1.2 2024-05-13 [1] CRAN (R 4.3.1)  
 fastmap 1.2.0 2024-05-15 [1] CRAN (R 4.3.1)  
 fs 1.6.4 2024-04-25 [1] CRAN (R 4.3.1)  
 generics 0.1.3 2022-07-05 [1] CRAN (R 4.3.1)  
 ggplot2 \* 3.5.1 2024-04-23 [1] CRAN (R 4.3.1)  
 glue 1.7.0 2024-01-09 [1] CRAN (R 4.3.1)  
 gtable 0.3.5 2024-04-22 [1] CRAN (R 4.3.1)  
 htmltools 0.5.8.1 2024-04-04 [1] CRAN (R 4.3.1)  
 htmlwidgets 1.6.4 2023-12-06 [1] CRAN (R 4.3.1)  
 httpuv 1.6.15 2024-03-26 [1] CRAN (R 4.3.1)  
 jsonlite 1.8.8 2023-12-04 [1] CRAN (R 4.3.1)  
 knitr 1.46 2024-04-06 [1] CRAN (R 4.3.1)  
 labeling 0.4.3 2023-08-29 [1] CRAN (R 4.3.1)  
 later 1.3.2 2023-12-06 [1] CRAN (R 4.3.1)  
 lifecycle 1.0.4 2023-11-07 [1] CRAN (R 4.3.1)  
 magrittr 2.0.3 2022-03-30 [1] CRAN (R 4.3.1)  
 memoise 2.0.1 2021-11-26 [1] CRAN (R 4.3.1)  
 mime 0.12 2021-09-28 [1] CRAN (R 4.3.1)  
 miniUI 0.1.1.1 2018-05-18 [1] CRAN (R 4.3.1)  
 munsell 0.5.1 2024-04-01 [1] CRAN (R 4.3.1)  
 pillar 1.9.0 2023-03-22 [1] CRAN (R 4.3.1)  
 pkgbuild 1.4.4 2024-03-17 [1] CRAN (R 4.3.1)  
 pkgconfig 2.0.3 2019-09-22 [1] CRAN (R 4.3.1)  
 pkgload 1.3.4 2024-01-16 [1] CRAN (R 4.3.1)  
 profvis 0.3.8 2023-05-02 [1] CRAN (R 4.3.1)  
 promises 1.3.0 2024-04-05 [1] CRAN (R 4.3.1)  
 purrr 1.0.2 2023-08-10 [1] CRAN (R 4.3.1)  
 R6 2.5.1 2021-08-19 [1] CRAN (R 4.3.1)  
 Rcpp 1.0.12 2024-01-09 [1] CRAN (R 4.3.1)  
 remotes 2.5.0 2024-03-17 [1] CRAN (R 4.3.1)  
 rlang 1.1.3 2024-01-10 [1] CRAN (R 4.3.1)  
 rmarkdown 2.27 2024-05-17 [1] CRAN (R 4.3.1)  
 rstudioapi 0.16.0 2024-03-24 [1] CRAN (R 4.3.1)  
 scales 1.3.0 2023-11-28 [1] CRAN (R 4.3.1)  
 sessioninfo 1.2.2 2021-12-06 [1] CRAN (R 4.3.1)  
 shiny 1.8.1.1 2024-04-02 [1] CRAN (R 4.3.1)  
 stringi 1.8.4 2024-05-06 [1] CRAN (R 4.3.1)  
 stringr 1.5.1 2023-11-14 [1] CRAN (R 4.3.1)  
 tibble 3.2.1 2023-03-20 [1] CRAN (R 4.3.1)  
 tidyselect 1.2.1 2024-03-11 [1] CRAN (R 4.3.1)  
 urlchecker 1.0.1 2021-11-30 [1] CRAN (R 4.3.1)  
 usethis 2.2.3 2024-02-19 [1] CRAN (R 4.3.1)  
 utf8 1.2.4 2023-10-22 [1] CRAN (R 4.3.1)  
 vctrs 0.6.5 2023-12-01 [1] CRAN (R 4.3.1)  
 withr 3.0.0 2024-01-16 [1] CRAN (R 4.3.1)  
 xfun 0.44 2024-05-15 [1] CRAN (R 4.3.1)  
 xtable 1.8-4 2019-04-21 [1] CRAN (R 4.3.1)  
 yaml 2.3.8 2023-12-11 [1] CRAN (R 4.3.1)  
  
 [1] /home/jordi/R/x86\_64-pc-linux-gnu-library/4.3  
 [2] /usr/local/lib/R/site-library  
 [3] /usr/lib/R/site-library  
 [4] /usr/lib/R/library  
  
──────────────────────────────────────────────────────────────────────────────

The current Git commit details are:

# what commit is this file at?   
if ("git2r" %in% installed.packages() & git2r::in\_repository(path = ".")) git2r::repository(here::here())

Local: master /home/jordi/Documents/nlprrcompendium  
Remote: master @ origin (https://github.com/sandpiles/nlprrcompendium.git)  
Head: [c9b1253] 2023-11-13: small changes