Project Report

Ye Zhang

10/01/2017

Introduction

The Public Library of Science (PLoS) is a nonprofit open access science, technology and medicine publisher, innovator and advocacy organization with a library of open access journals and other scientific literature under an open content license. This project is to perform an analysis of the statistical analyses in all published PLoS papers, so as to answer quenstions as below:

- What are the most common techniques?
- How do they vary by field?
- Are there any trends over the last 10-15 years?

Step 1

Data: the dataset for this project should include all the published PLoS papers from its 7 journals, PLoS one, PLoS Biology, PLoS Medicine, PLoS Comutational Biology, PLoS Genetic, PLoS Neglected Tropical Diseases and PLos Pathogens. For each publication, there'are a list of information we need to download from the websites into our R program as the dataset:

- Article title
- Authors
- Article DOI
- PLoS journal
- Date of publication
- Materials and Methods part

Step 2

Types of analyses: Usually the statistical analysis techique utilized in a publication is described in the *Materials and Methods* section of the article,thus we should focus on extracting all the types of data analyses techniques mentioned in the *Materials and Methods* section of all the publications. One possibe way is to look for certain key words, such as "Hypothesis testing", "t test", "linear regression", "log linear regression", et al. With this method, it is important to establish a decent pool of key words before extraction, and some refereneces summarizing the statistical analyses methods online could be helpful, such as https://www.statisticallysignificantconsulting.com/Statistical-Tests.htm.

Step 3

Dataset Analysis: After extracting all the key words from articles, we can then start to answer the three questions listed at the beginning. With the dataset established through **Step 1** and **2**, it's possible to figure out the most commonly utilized analyses techniques, and coorelation between these techniques and the fields (the PLoS journal) and publication years. Take the key word "t test" as an example, we can figure out how many times the "t test" is mentioned over the years as well as in articles among 7 different fields.

```
## Loading required package: NLP
```

```
## Warning: package 'NLP' was built under R version 3.4.1
```

```
## Loading required package: rplos
```

```
## Loading required package: fulltext
```

```
## Warning: package 'XML' was built under R version 3.4.1
```

Data Preparation

In order to establish a pool for the key words, first a list of full articles with the word “statistical” in abstract is searched using R package “rplos”, which contains functions that can be used for PLoS article searching and information download. By indicating “statistic” in abstract part, we can achieve result `outide_id` containing all the DOIs of all the full articles that we are interested in.

```
install.packages("tidytext",repos="http://cran.rstudio.com/")
```

```
##
## The downloaded binary packages are in
##  /var/folders/5s/d6trnk_14_lb96_4zy9nc57c0000gn/T//RtmpXXXCW9/downloaded_packages
```

```
library(tidytext)
```

```
## Warning: package 'tidytext' was built under R version 3.4.2
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.1
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:fulltext':
##
##     collect
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 3.4.1
```

```
library(stringr)

out_id_all <- searchplos(q="materials_and_methods: statistics",
                    fl="id", fq='doc_type: full', sort='publication_date desc')

out_id <- searchplos(q="materials_and_methods: statistics",
                     fl="id", fq='doc_type: full', sort='publication_date desc', limit = 100)

# Full text xml given a DOI
out_fulltext <- plos_fulltext(doi=out_id$data$id[1])
data <- xmlParse(out_fulltext[[1]])
out_abstract_1 <- xpathSApply(data, "//abstract", xmlValue)
tidy_abstract_1 <- out_abstract_1 %>% str_replace_all("[[:punct:]]", " ") %>% tidy()

out_fulltext <- plos_fulltext(doi=out_id$data$id[2])
data <- xmlParse(out_fulltext[[1]])
out_abstract_2 <- xpathSApply(data, "//abstract", xmlValue)
tidy_abstract_2 <- out_abstract_2 %>% str_replace_all("[[:punct:]]", " ") %>% tidy()
```

```r
tidy_abstract <- rbind(tidy_abstract_1,tidy_abstract_2)

#for (i in 1:100) {
#  out_fulltext <- plos_fulltext(doi=out_id$data$id[i])
#  xpathSApply(xmlParse(out_fulltext[[1]]), "//abstract", xmlValue)
#  out_abstract <- xpathSApply(data, "//abstract", xmlValue)
#  out_abstract_full
#}

text <- ft_get(out_id$data$id[1])
text$plos$data
```

```
## $backend
## NULL
##
## $path
## [1] "session"
##
## $data
## 1 full-text articles retrieved
## Min. Length: 247753 - Max. Length: 247753
## DOIs: 10.1371/journal.pcbi.1005776 ...
##
## NOTE: extract xml strings like output['<doi>']
```

Create a decent pool for common statistical methods and select paper with key words in “material and methods” part.

```r
pool <- list("linear regression", "bootstrap", "maximum likelihood", "ANOVA", "clustering",
          "neural network", "support vector machine", "ridge regression", "MCMC")
pl <- c("linear regression", "bootstrap", "maximum likelihood", "ANOVA", "clustering",
          "neural network", "support vector machine", "ridge regression", "MCMC")

# Using keywords in the pool and return "material and methods"
out_LR <- searchplos(q="materials_and_methods: linear regression",
                     fl=c("id","title","journal","publication_date"),
                    fq='doc_type: full', sort='publication_date desc')
out_Boot <- searchplos(q="materials_and_methods: bootstrap",
                     fl=c("id","title","journal","publication_date"),
                    fq='doc_type: full', sort='publication_date desc')
out_MaxL <- searchplos(q="materials_and_methods: maximum likelihood",
                     fl=c("id","title","journal","publication_date"),
                    fq='doc_type: full', sort='publication_date desc')
out_ANOVA <- searchplos(q="materials_and_methods: ANOVA",
                     fl=c("id","title","journal","publication_date"),
                    fq='doc_type: full', sort='publication_date desc')
out_Cluster <- searchplos(q="materials_and_methods: clustering",
                     fl=c("id","title","journal","publication_date"),
                    fq='doc_type: full', sort='publication_date desc')
out_NN <- searchplos(q="materials_and_methods: neural network",
                     fl=c("id","title","journal","publication_date"),
                    fq='doc_type: full', sort='publication_date desc')
out_SVM <- searchplos(q="materials_and_methods: support vector machine",
                     fl=c("id","title","journal","publication_date"),
                    fq='doc_type: full', sort='publication_date desc')
```

```
out_RR <- searchplos(q="materials_and_methods: ridge regression",
                     fl=c("id","title","journal","publication_date"),
                    fq='doc_type: full', sort='publication_date desc')
out_MCMC <- searchplos(q="materials_and_methods: MCMC",
                     fl=c("id","title","journal","publication_date"),
                    fq='doc_type: full', sort='publication_date desc')
```

Data cleaning

After downloading the "abstract"" and "materials and methods" from articles we are intereted in, we clean the data using R package "tidyr" and "tidytext", removing the stopwords and punctuations.

```
# tidy_abstract <- out_abstract %>% str_replace_all("[[:punct:]]", " ") %>% tidy()

file_words <- tidy_abstract %>%
  unnest_tokens(word, x) %>%
  anti_join(stop_words) %>%
  group_by(word) %>%
  tally() %>%
  arrange(desc(n))
```

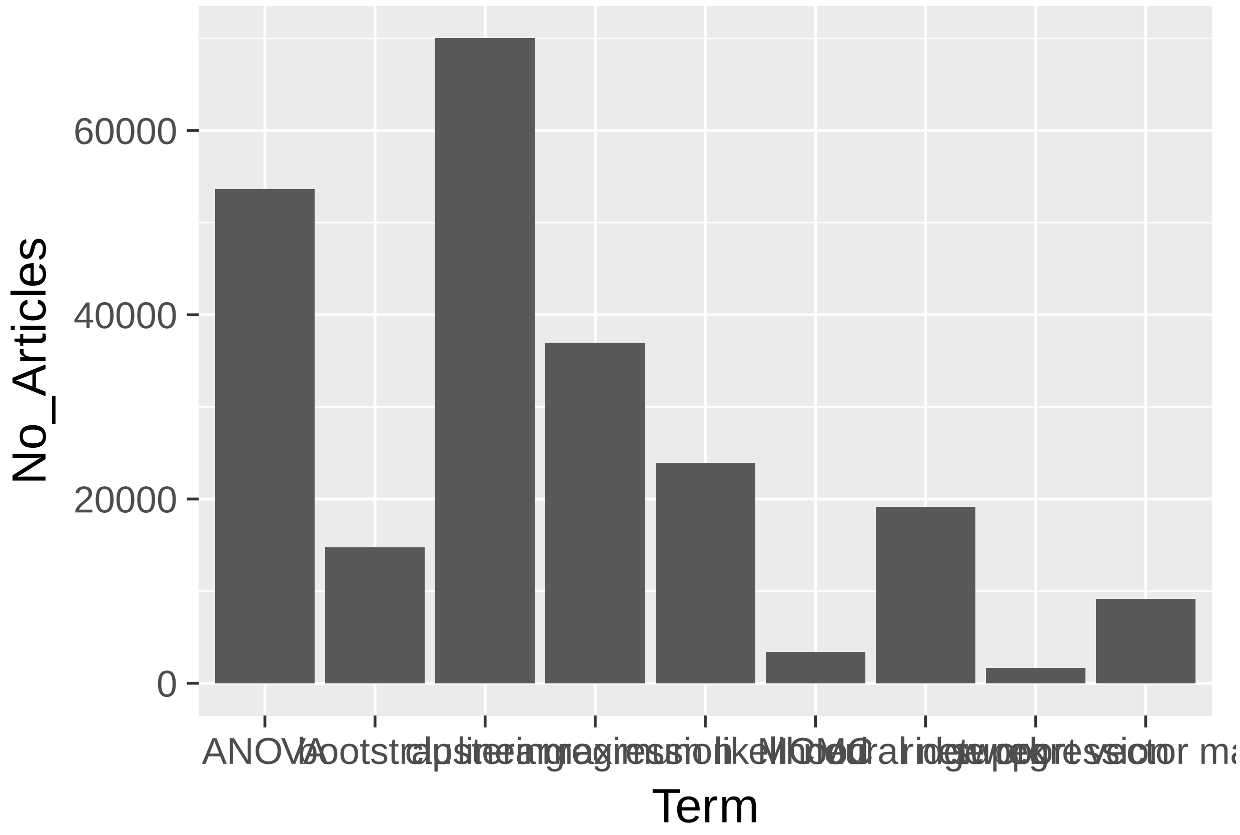
```
## Joining, by = "word"
```

Data Analysis

1. Compare the frequency of statistical methods by counting the number of articles using the key words

```
plosword(pool, vis = TRUE)
```

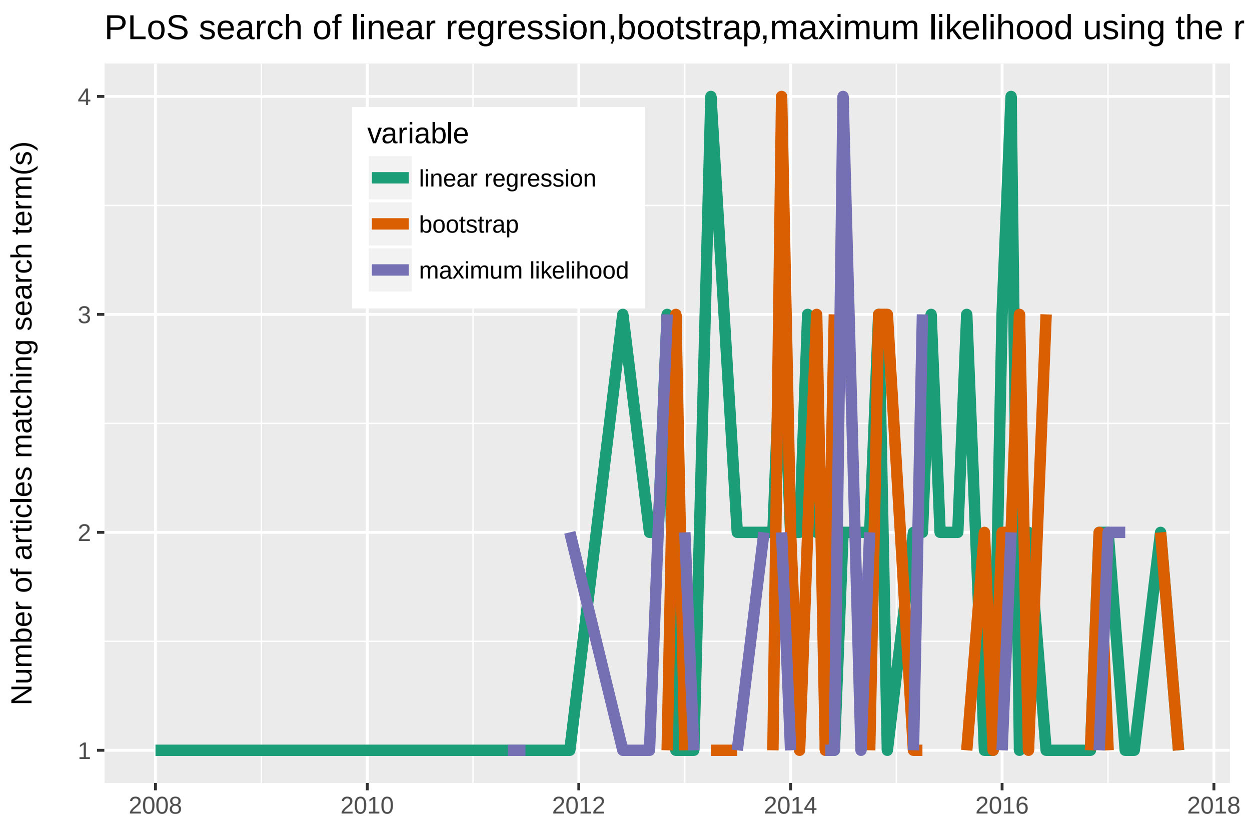
```
## $table
##   No_Articles                  Term
## 1       36942     linear regression
## 2       14767             bootstrap
## 3       23928    maximum likelihood
## 4       53608                 ANOVA
## 5       70014            clustering
## 6       19159        neural network
## 7        9167 support vector machine
## 8        1667      ridge regression
## 9        3393                  MCMC
##
## $plot
```



2. Observe the

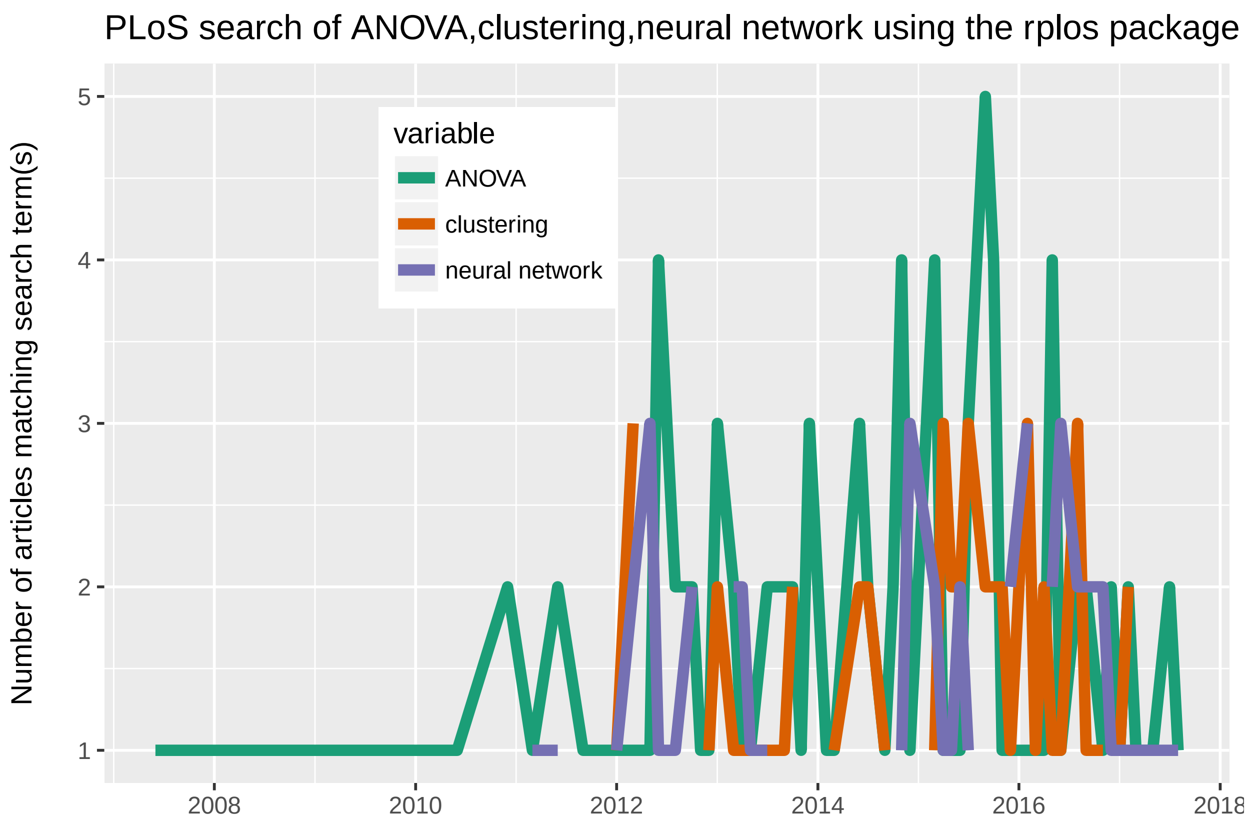
```
plot_throughtime(terms = pl[1:3], limit = 100)
```

```
## Warning: Removed 7 rows containing missing values (geom_path).
```



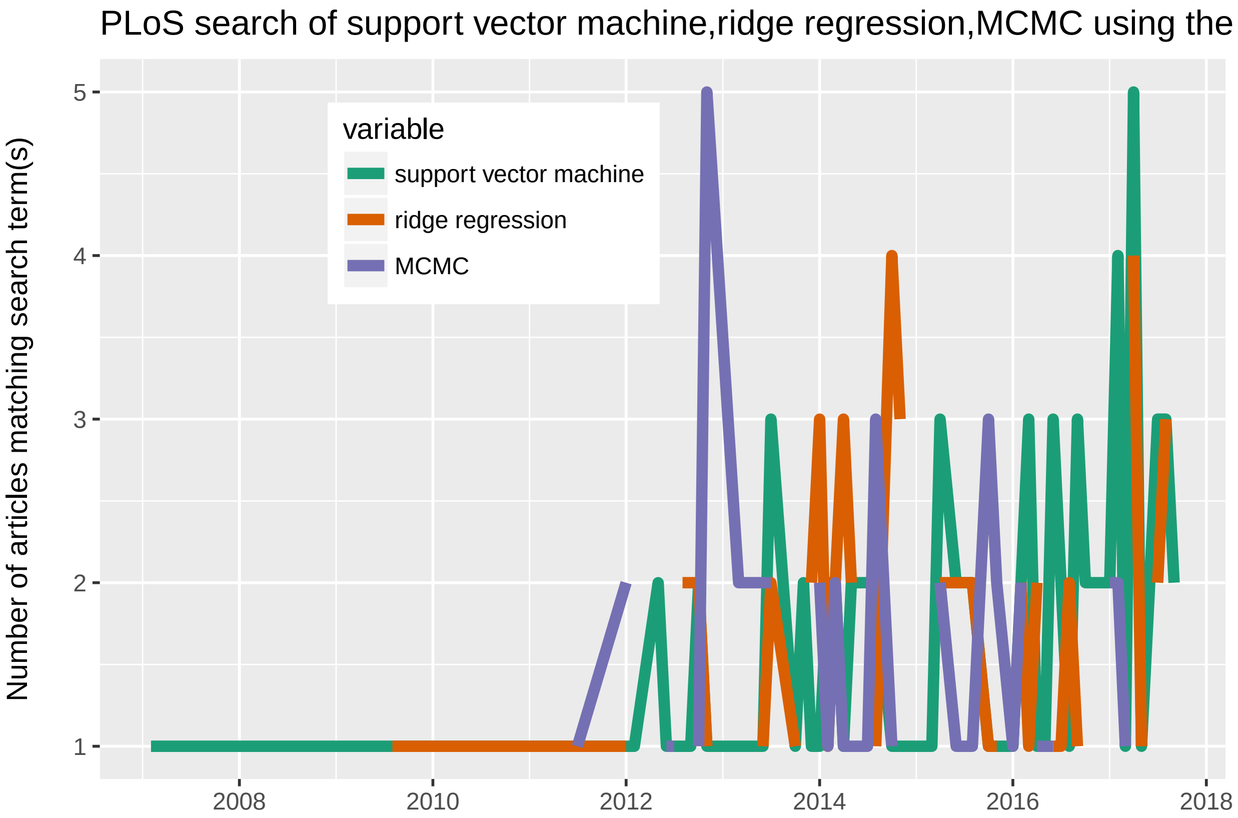
```
plot_throughtime(terms = pl[4:6], limit = 100)
```

```
## Warning: Removed 9 rows containing missing values (geom_path).
```



```r
plot_throughtime(terms = pl[7:9], limit = 100)
```

```
## Warning: Removed 11 rows containing missing values (geom_path).
```



```
output <- highplos(q='linear regression', hl.fl = 'Materials and Methods')
```

```
## http://api.plos.org/search?wt=json&q=linear regression&start=0&hl=true&hl.fl=Materials and Methods
```

Reference

1. https://www.statisticallysignificantconsulting.com/Statistical-Tests.htm.