Project Report

Ye Zhang

9/17/2017

Project Outline

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:

- Text of the article (including figure legends)
- Article title
- Authors
- PLoS journal
- Date of publication

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
```

Getting data

Search a list of articles with the key words "statistical"

```
out_id <- searchplos(q="materials_and_methods: statistics",
                     fl="id", fq='doc_type: full', sort='publication_date desc')
out_fulltext <- plos_fulltext(doi=out_id$data$id[1])
data <- xmlParse(out_fulltext$`10.1371/journal.ppat.1006607`)
out_abstract <- xpathSApply(data, "//abstract", xmlValue)

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

```
## $backend
## NULL
##
## $path
## [1] "session"
##
## $data
## 1 full-text articles retrieved
## Min. Length: 92268 - Max. Length: 92268
## DOIs: 10.1371/journal.ppat.1006607 ...
##
## NOTE: extract xml strings like output['<doi>']
```

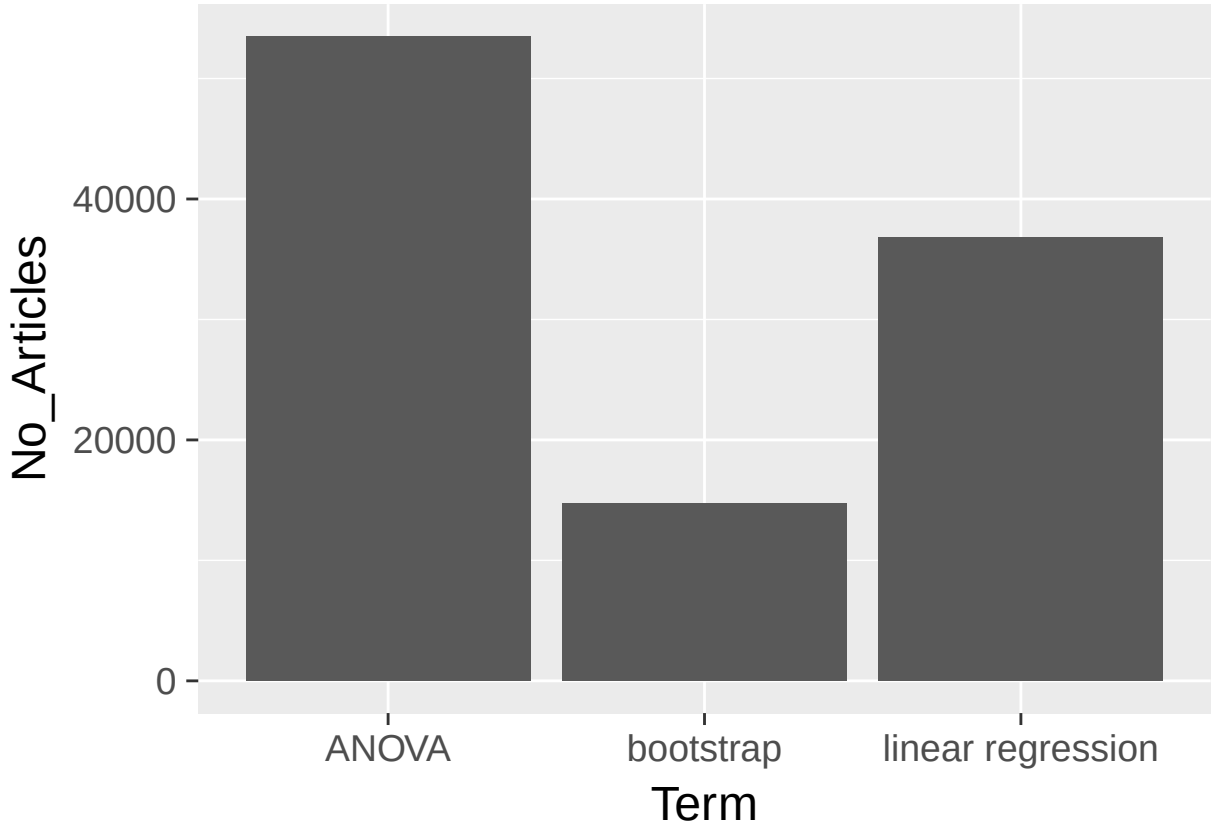
Create a dictionary for common statistical methods and select paper with key words in "material and methods" part

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

Analysis

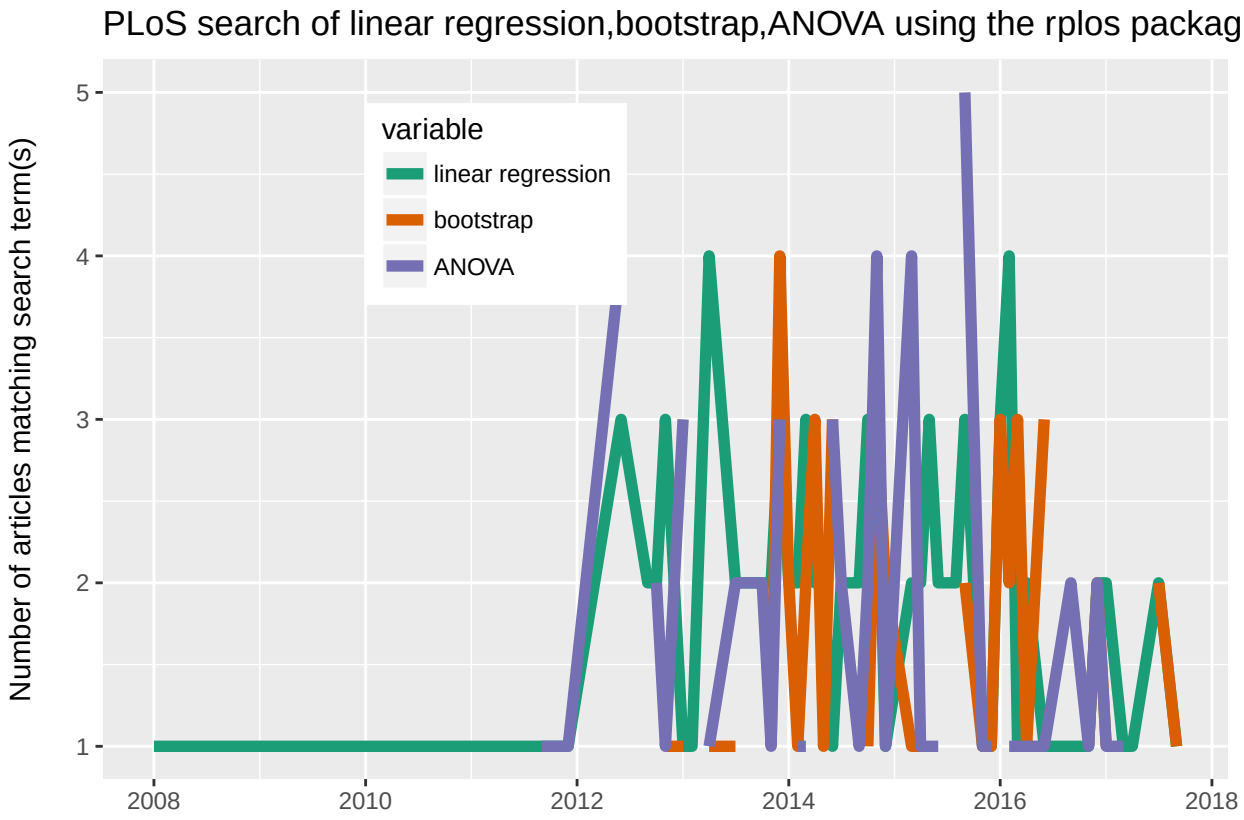
```
plosword(list('linear regression', 'bootstrap','ANOVA'), vis = TRUE)
```

```
## $table
##   No_Articles              Term
## 1       36854 linear regression
## 2       14734         bootstrap
## 3       53506             ANOVA
##
## $plot
```




```
plot_throughtime(terms = c('linear regression','bootstrap','ANOVA'), limit = 100)
```

```
## Warning: Removed 9 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
```