Data Science Project

Gege Gui

9/11/2017

Introduction:

PLoS (Public Library of Science) is a nonprofit Open Access publisher, innovator and advocacy organization with a mission to accelerate progress in science and medicine by leading a transformation in research communication. The data analysis project is to find out the most common statistical techniques in all published PLoS papers and the trends over the last 10-15 years.

The website has an API tool for finding related articles. ALM API can print out information about articles we are interesed in, such as author, publication date, abstract, etc. There is an R package "rplos", which contains functions that can be used easily to look at article matrics.

Besides the "rplos" R package for summary statistics, we want to conduct text and data mining on our own by obtaining entire corpus. The raw data are from "Documentation: Text and Data Mining" page. As bulk downloading of article HTML/PDF/XML is discouraged, we use the "PMC OA Bulk Download FTP" to obtain research articles from multiple journals. The files are organized by topics alphabetically. We choose the folder named as "Biostatistics" as the one for model development. There are 52 documents in total. We use "LDA" (Latent Dirichlet Allocation) in R package "topicmodels" together with text mining commands in R package "tm".

Method

Exploratory data analysis using "rplos"

We use "rplos" to do some exploratory data analysis for the papers published from year 2007 to year 2016. To use regular expression for further search, we create a dictionary for common statistical methods. We use "The Elements of Statistical Learning - Data Mining, Inference, and Prediction" second edition, by Trevor Hastie, Robert Tibshirani, Jerome Friedman, as a reference.

The key words in the dictionary are searched one by one in the abstract of each article using "searchplos" function. Publication dates are obtained by ascending order. The format of the output is set, so we can use "stringr" package to get the publication date count by month and by year to see the overall trend. The total count is expressed using barplots.

Text mining and topic model

Latent Dirichlet Allocation (LDA) is the common algorithm used for distinguishing topics by text mining. R has several packages based on NLP to solve this kind of problem.

From the bulk downloading data file, we choose "PLoS_Curr" directory as the data for model training. All text files are read in using "DirSource" and "Vcorpus" function to create a corpus. Numbers are removed from the text file. There are 599 documents in total.

We want to create a new list of stopwords which contains English stopwords and the 20 most frequent words from the document. We transform the corpus to a tidy document, unnest the tokens, remove English stopwords and use "count" and "table" to get the words listed by frequency. The 20 most frequent words include "data", "research", "analysis", which are common in research papers but do not give specific information about methods used in the research. We also only keep words from: https://github.com/dwyl/english-words/blob/master/words.txt.zip. Another part of the stopwords list is the difference between the whole vocabulary of the tidy file and this common English words list. The newly created stopwords list is "newstopword".

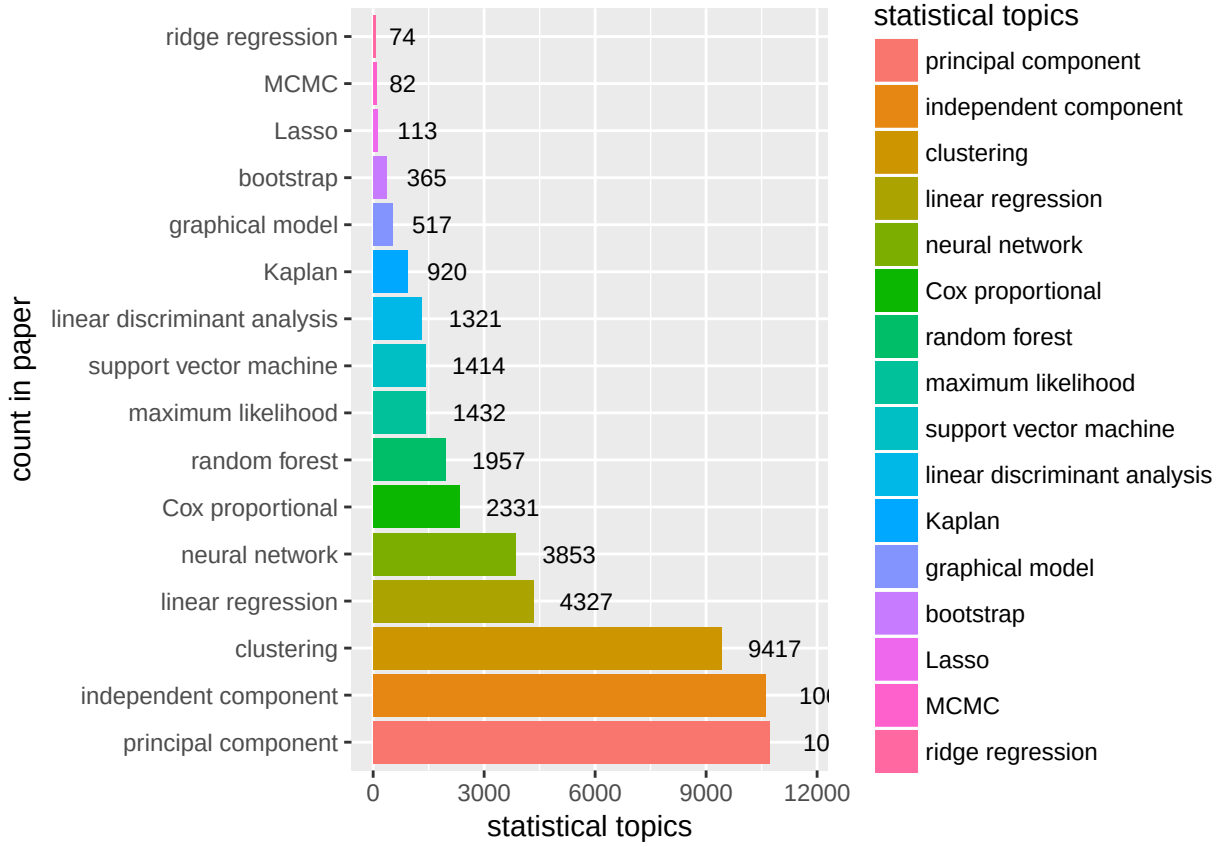
As our only concern is statistical methods, we focus on the "methods" part of the papers. This part usually starts with a line containing three or less words after removing punctuations, and one of the words are "Method", "Methods", "Methodology". The ending indicators are "Results", "Discussion", "Supporting". We use these regular expressions together with the word count of each line to determine the start and end of the paragraphs. We remove all files that do not have a start or end line output as well as the start is greater than the end. Lines with one or no words are removed and there are 337 documents remaining in the corpus.

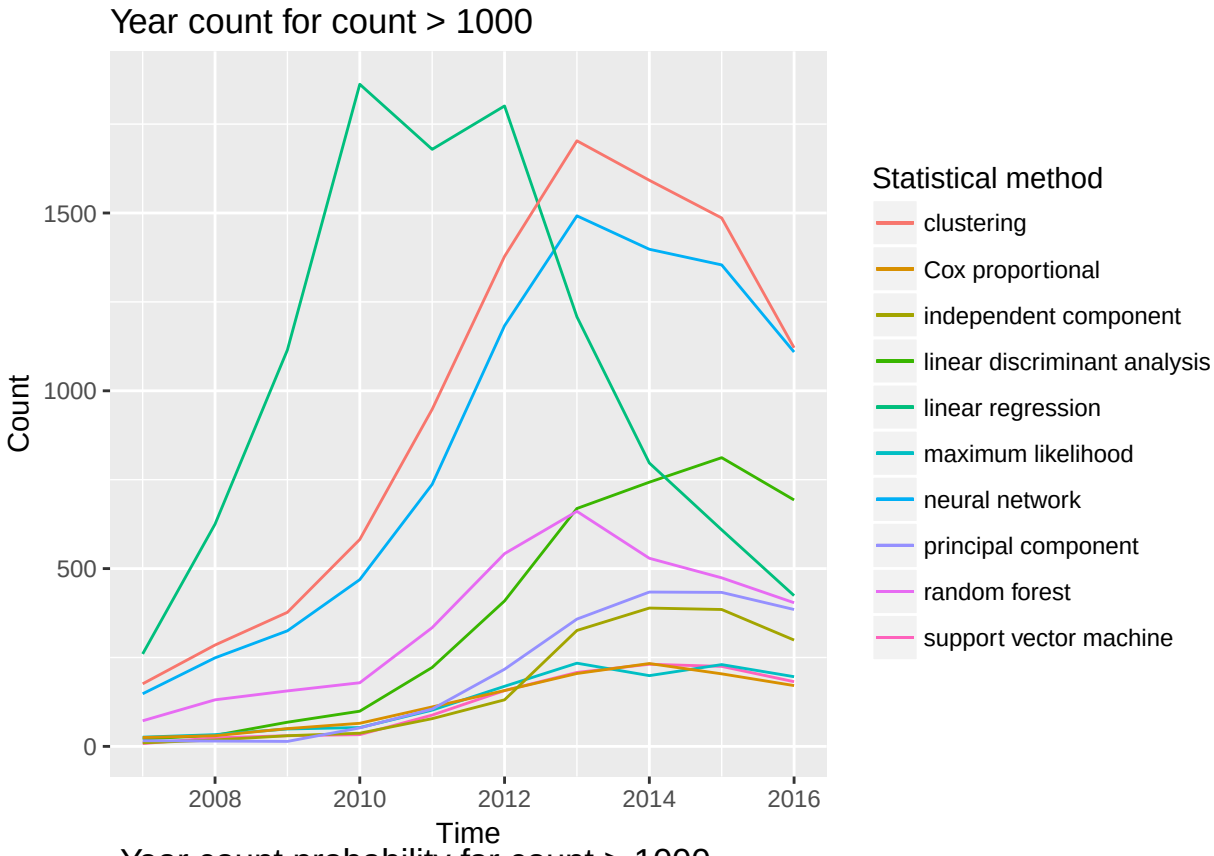
The input of LDA model is a "DocumentTermMatrix". We transform the corpus with only method parts to the required form with "newstopword" specified as the stopwords list. One document is found to have 0 characters. After removing this document, the "DocumentTermMatrix" contains 336 files and we use it for LDA model.

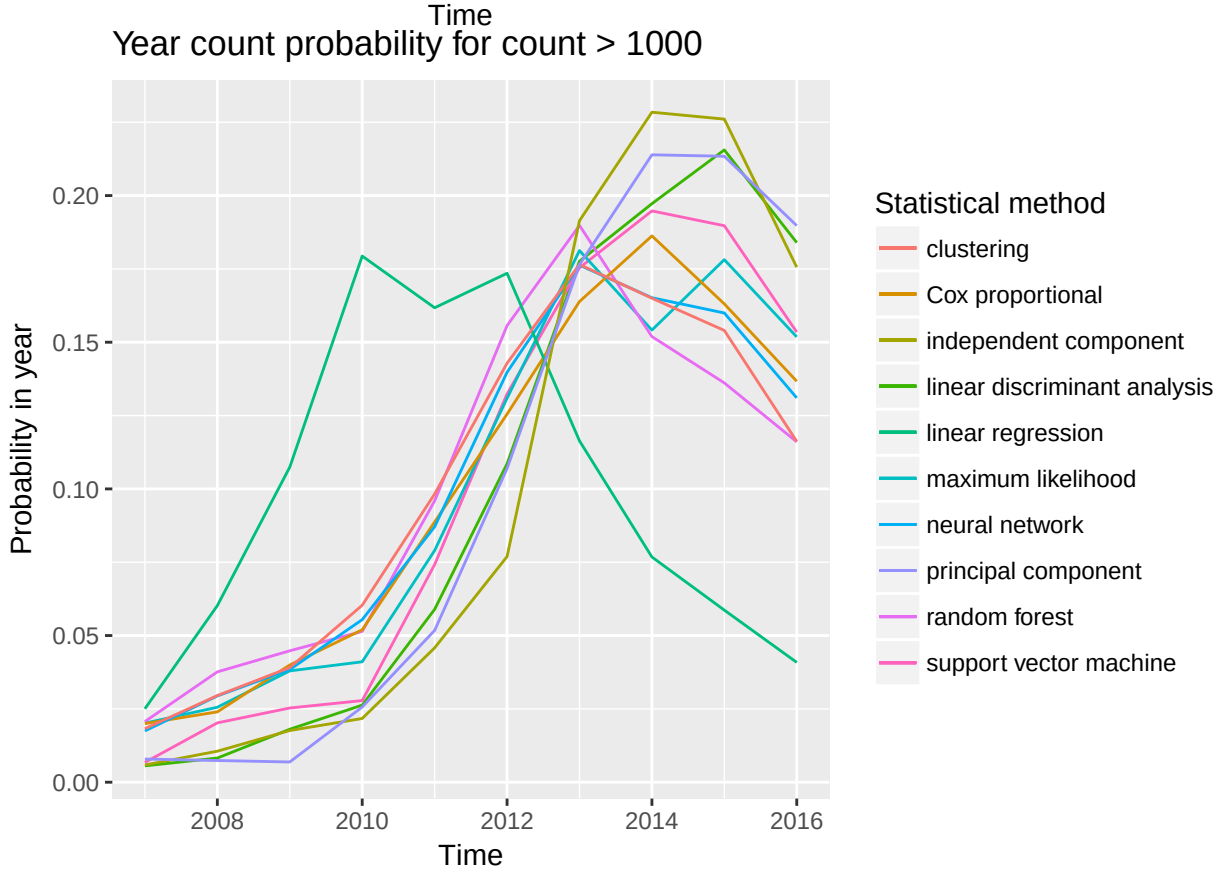
Topics are set to be 3 to 10. The likelihood can be calculated to determine the best number of topics. To extract results from the posterior distribution, we make a wordloud of the top 20 words from each topic of these models. Ideally we can label them by eyeballing.

Result

The total count shows below with sixteen common statistical methods. To make the time trend more clearly, we choose the methods with more than one thousand count. The time trends stay similar when using raw count and normalized probability of each method.




Year count for count > 1000
Count
1500
1000
500
0
2008
2010
2012
2014
2016
Time
Statistical method
clustering
Cox proportional
independent component
linear discriminant analysis
linear regression
maximum likelihood
neural network
principal component
random forest
support vector machine


Year count probability for count > 1000
Probability in year
0.20
0.15
0.10
0.05
0.00
2008
2010
2012
2014
2016
Time
Statistical method
clustering
Cox proportional
independent component
linear discriminant analysis
linear regression
maximum likelihood
neural network
principal component
random forest
support vector machine

From the barplots we observe that the most common statistical methods are principal component analysis, independent component analysis and clustering. Each of them has more than 9000 count. Ridge regression, MCMC and Lasso are mentioned little, with around or less than 100 count for each.

As for the time trend, linear regression has an increasing trend to 2010, then fluctuates and decreasing to nearly the starting level. All the other methods increase from 2017 to around 2013, then decrease by 5%.

The words related with first 3 out of 5 topics are shown below. It is not easy to get a summary word from them.

```
##                  1            2                    3
## word 1      samples    community           population
## word 2        cells  information documentclassptarticle
## word 3    performed        local         transmission
## word 4      animals   management          individuals
## word 5       muscle    emergency                 days
## word 6      protein       survey                 rate
## word 7         days     response            sequences
## word 8       figure     included             epidemic
## word 9      control participants               period
## word 10      medium       people         distribution
## word 11    training   interviews             outbreak
## word 12   collected   government            estimates
## word 13        cell        focus            estimated
## word 14      tested   collection            incidence
## word 15       water  communities           infectious
## word 16  previously    conducted            infection
## word 17      animal       social            parameter
## word 18         cag        level           parameters
## word 19 significant       floods               models
## word 20       human    including           individual
```

Discussion

Statistical models can be applied to view the relationship of time and topics. However, the only feature is time, so regression analysis is hard to conduct. Plotting is the most straight forward way. There is no function in "rplos" that can search the key words in method parts, but the appearance in abstract can be representative.

We need to conduct analysis by fields besides time trend.

The number of frequent words can be adjusted by the total number of unique words in the model. We need to consider whether to obtain these words by the full text files or only the method part.

As not all papers are statistics papers, the topics are possibly to be field specific, but the modeling procedure is intriguing.