Post 2: Analyzing Literature in r

Alexander Jou

> ## How and why do we analyze literature using r?
>
> While we often don't associate literature with mathematics, there are many tools in r that are helpful for better understanding and analysis of texts. Whether it's counting word frequency or creating a network to further understand relationships, literature can be broken down in many different ways. Doing all of these mathematical analyses help us comprehend the text in a much different way than just reading it. Especially for large journal collections, there is a lot of use for r to better understand the work: https://paginas.fe.up.pt/~prodei/dsie15/web/papers/dsie15_submission_10.pdf. In this project, we will analyze a collection of literature and use r to help us come to new conclusions about the text. Here is a link to the text mining package for more detail: https://cran.r-project.org/web/packages/tm/index.html.

Installing necessary packages

First we must install and load the packages needed so we have all the tools we need for this project.

```
install.packages("tidyverse")
install.packages("stringr")
install.packages("dplyr")
install.packages("tm")
install.packages("wordcloud")
```

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────── tidyverse 1.2.1 ──
```

```
## ✔ ggplot2 2.2.1     ✔ purrr   0.2.4
## ✔ tibble  1.3.4     ✔ dplyr   0.7.4
## ✔ tidyr   0.7.2     ✔ stringr 1.2.0
## ✔ readr   1.1.1     ✔ forcats 0.2.0
```

```
## ── Conflicts ────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(stringr)
library(dplyr)
library(tm)
```

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

```
##
## Attaching package: 'NLP'
```

```
## The following object is masked from 'package:ggplot2':
##
##     annotate
```

```
library(wordcloud)
```

```
## Loading required package: RColorBrewer
```

Loading the book

Now we have to load the clean data. For this project we are using all Psychonomic Society publications between 2004 and 2016. Here is a link to the collection of journals: http://www.psychonomic.org/?page=journals.

```
load(url("https://mvuorre.github.io/data/scopus/scopus-psychonomics-clean.rda"))
```

Learning about the data

The glimpse function allows us to see all the columns in this data set. More about the glimpse function: https://www.rdocumentation.org/packages/dplyr/versions/0.4.3/topics/glimpse.

```
glimpse(d)
```

```
## Observations: 7,614
## Variables: 17
## $ Authors                   <chr> "Button C., Schofield M., Croft J.",...
## $ Title                     <chr> "Distance perception in an open wate...
## $ Year                      <int> 2016, 2016, 2016, 2016, 2016, 2016, ...
## $ Cited_by                  <int> NA, 1, 2, 2, 3, 2, 2, NA, NA, NA, 1,...
## $ DOI                       <chr> "10.3758/s13414-015-1049-4", "10.375...
## $ Affiliations              <chr> "School of Physical Education, Sport...
## $ Authors_with_affiliations <chr> "Button, C., School of Physical Educ...
## $ Abstract                  <chr> "We investigated whether distance es...
## $ Author_Keywords           <chr> "3D perception; Perception and actio...
## $ Index_Keywords            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ References                <chr> "Baird, R.W., Burkhart, S.M., (2000)...
## $ Correspondence_Address    <chr> "Button, C.; School of Physical Educ...
## $ Abbreviated_Source_Title  <chr> "Atten. Percept. Psychophys.", "Atte...
## $ Document_Type             <chr> "Article", "Article", "Article", "Ar...
## $ Pages                     <int> 7, 13, 19, NA, 12, 12, 32, 19, 6, 22...
## $ Publication               <chr> "Attention, Perception & Psychophysi...
## $ Pub_abbr                  <chr> "Atten. Percept. Psychophys.", "Atte...
```

Sorting the Author Keywords

Looking at the Author_Keywords variable, each value is actually a list of grouped keywords separated by semicolons. What we want is every phrase to be considered as its own keyword in the value.

```
d %>%
    select(Author_Keywords) %>%
    mutate(Author_Keywords = str_split(Author_Keywords, "; "))
```

```
## # A tibble: 7,614 x 1
##    Author_Keywords
##             <list>
##  1       <chr [4]>
##  2       <chr [6]>
##  3       <chr [4]>
##  4       <chr [5]>
##  5       <chr [5]>
##  6       <chr [3]>
##  7       <chr [3]>
##  8       <chr [6]>
##  9       <chr [3]>
## 10       <chr [3]>
## # ... with 7,604 more rows
```

Now we have to separate all of those individual keywords within one value into their own values.

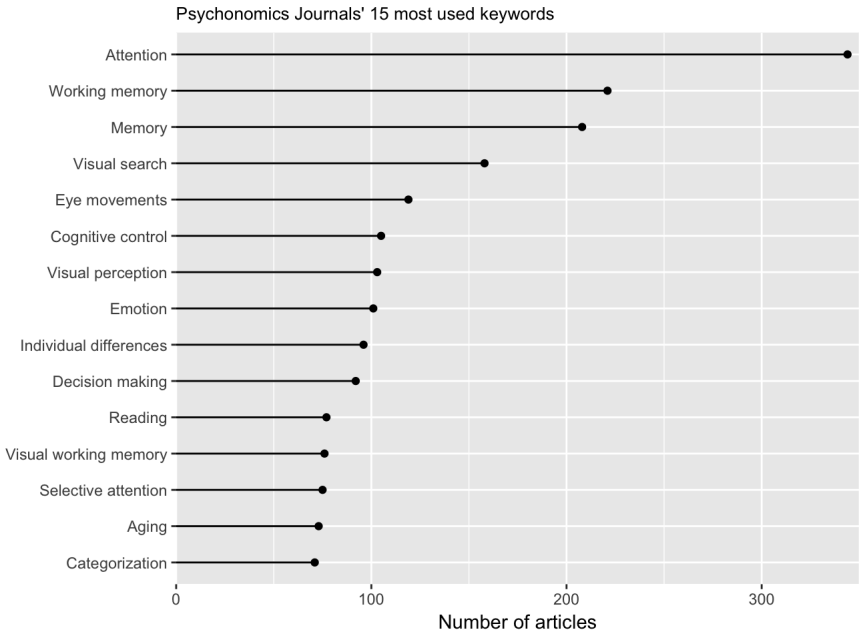
```
d %>%
  select(Author_Keywords) %>%
  mutate(Author_Keywords =   str_split(Author_Keywords, "; ")) %>%
  unnest(Author_Keywords)
```

Visualizing Data

Now that we've separated the variables, we can use this the modified data to find the most used keywords in this collection of journals.

```
d %>%
    select(Author_Keywords) %>%
    mutate(Author_Keywords = str_split(Author_Keywords, "; ")) %>%
    unnest(Author_Keywords) %>%
    filter(!is.na(Author_Keywords)) %>%
    group_by(Author_Keywords) %>%
    count() %>%  #Counting the amount of each keyword
    ungroup() %>%
    top_n(15) %>%  #Getting the 15 top tags
    mutate(Author_Keywords = reorder(Author_Keywords, n)) %>%  #
    ggplot(aes(n, Author_Keywords)) +
    geom_point() +
    geom_segment(aes(xend = 0, yend=Author_Keywords)) +
    scale_x_continuous(limits = c(0, 350), expand = c(0, 0)) +
    labs(x="Number of articles", y="",
         subtitle = "Psychonomics Journals' 15 most used keywords")
```

```
## Selecting by n
```



Takeaways

From this plot we can see that attention is the most used keyword by a significant margin. The prevalance of memory in these journals can also be seen with three different variations of memory in the top 15. This list of keywords gives great insight into the kinds of articles that are in this journal and also a look into the field of psychonomics. Looking at this plot you can see that this study is clearly concerned with cognitive ability, sensory perception, attention, and memory.

Page Analysis

```
summary(d$Pages)
```

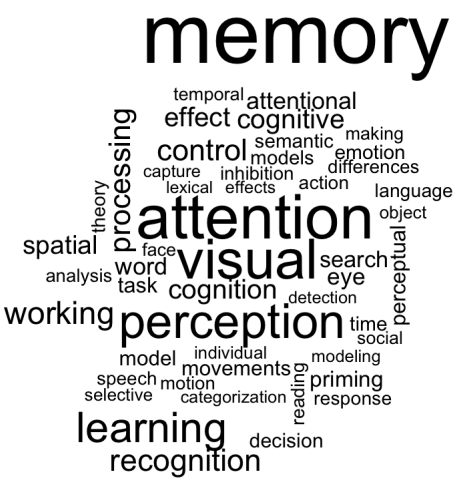
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00    6.00    9.00   10.17   13.00 1006.00     201
```

Invoking the summary function allows us to analyze the lengths of these journals. While the majority of them are less than 10 pages and the first quartile and third quartile are relatively close, it can be seen based on the max that some of the texts in this collection are very extensive and more than just short articles.

Wordcloud

Another useful visual feature is wordlcoud. This function creates a collection of the most frequently used words and increases the size of the font the more the word is used. Link to the wordcloud package: https://cran.r-project.org/web/packages/wordcloud/wordcloud.pdf.

```
m <- as.matrix(d$Author_Keywords)
wordcloud(m, min.freq = 100)
```



Conclusion

As seen from the work we've done above, r can be used in many different ways to visualize a text. Whether it's creating a wordcloud or just looking at the summary stats for pages, many conclusions can be drawn on this extensive collection of work. These tools allow us to get a general sense of what the journal is about withouth having to individuall go through each text. This shows the power of r to analyze bulk amounts of literature efficiently. For in depth look at text analysis use these resources: http://tidytextmining.com/ and https://eight2late.wordpress.com/2015/05/27/a-gentle-introduction-to-text-mining-using-r/.