Homework 8

Peng Xu

2017/10/31

Problem 2

As the data have too many text typos, the cleaning work is done and saved as csv file with Excel. Then my interest is what factors influence the R level. So the people are labelled with three levels. Then the whole table are gathered into pairs of data. Their relationship are drawn with the ggraph package, as shown below.

From the graph, whether people use PC or Mac has no effect on their master levels. But if they have learned some professional software, such as SQL, Minitab, SPSS, people usually have high levels at R. As to their major, stat and finance students do R works better and the average level of masters are higher than those Bachelor's.

```
setwd('D:/Git/STAT_5015_homework/08_text_mining_Rnotebooks_bash_sed_awk')

OriList <- read.table('survey_data.txt', sep="\t")
OriList2 <- read.csv('TidyData.csv', sep=",", header = FALSE)
OriList2
```

```
##                  V1      V2      V3      V4      V5   V6      V7     V8
## 1          beginner      PC    Math      BS    NULL NULL    NULL   NULL
## 2          beginner     Mac    Math      BS    NULL NULL     SAS MATLAB
## 3      intermediate      PC Finance      BS Finance   MS  Matlab    SAS
## 4      intermediate      PC    Stat      BS    NULL NULL Minitab    SAS
## 5      intermediate      PC    Stat      BS    NULL NULL Minitab    SAS
## 6          beginner Surface    Math History    NULL NULL       C   NULL
## 7          beginner      PC    Math    NULL    NULL NULL    NULL   NULL
## 8          beginner      PC    Econ    Math    NULL NULL    Java   NULL
## 9      intermediate     Mac    Econ    Stat    NULL NULL  Python    SAS
## 10     intermediate      PC    Econ    Math    Stat NULL     SAS   NULL
## 11 beg/intermediate      PC    DAAS    NULL    NULL NULL     SAS   NULL
## 12     intermediate     Mac Finance      BS    Stat   MS  Python Matlab
## 13 beg/intermediate      PC    Mech    NULL    NULL NULL  Matlab   Java
## 14     intermediate      PC    Math    Stat    NULL NULL  Matlab   Java
##        V9    V10  V11  V12  V13
## 1    NULL   NULL NULL NULL NULL
## 2    NULL   NULL NULL NULL NULL
## 3     SQL   NULL NULL NULL NULL
## 4  Python   NULL NULL NULL NULL
## 5  Python    SQL  C++    R SPSS
## 6    NULL   NULL NULL NULL NULL
## 7    NULL   NULL NULL NULL NULL
## 8    NULL   NULL NULL NULL NULL
## 9    NULL   NULL NULL NULL NULL
## 10   NULL   NULL NULL NULL NULL
## 11   NULL   NULL NULL NULL NULL
## 12   Java  Linux  C++ NULL NULL
## 13    C++   NULL NULL NULL NULL
## 14    SAS Python NULL NULL NULL
```

```
OriList2$V1 <- sub("beginner", "Level_1", OriList2$V1)
OriList2$V1 <- sub("beg/intermediate", "Level_2", OriList2$V1)
OriList2$V1 <- sub("intermediate", "Level_3", OriList2$V1)

GroupData <- gather(OriList2,Attribute,value,V2:V13)
```

```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```

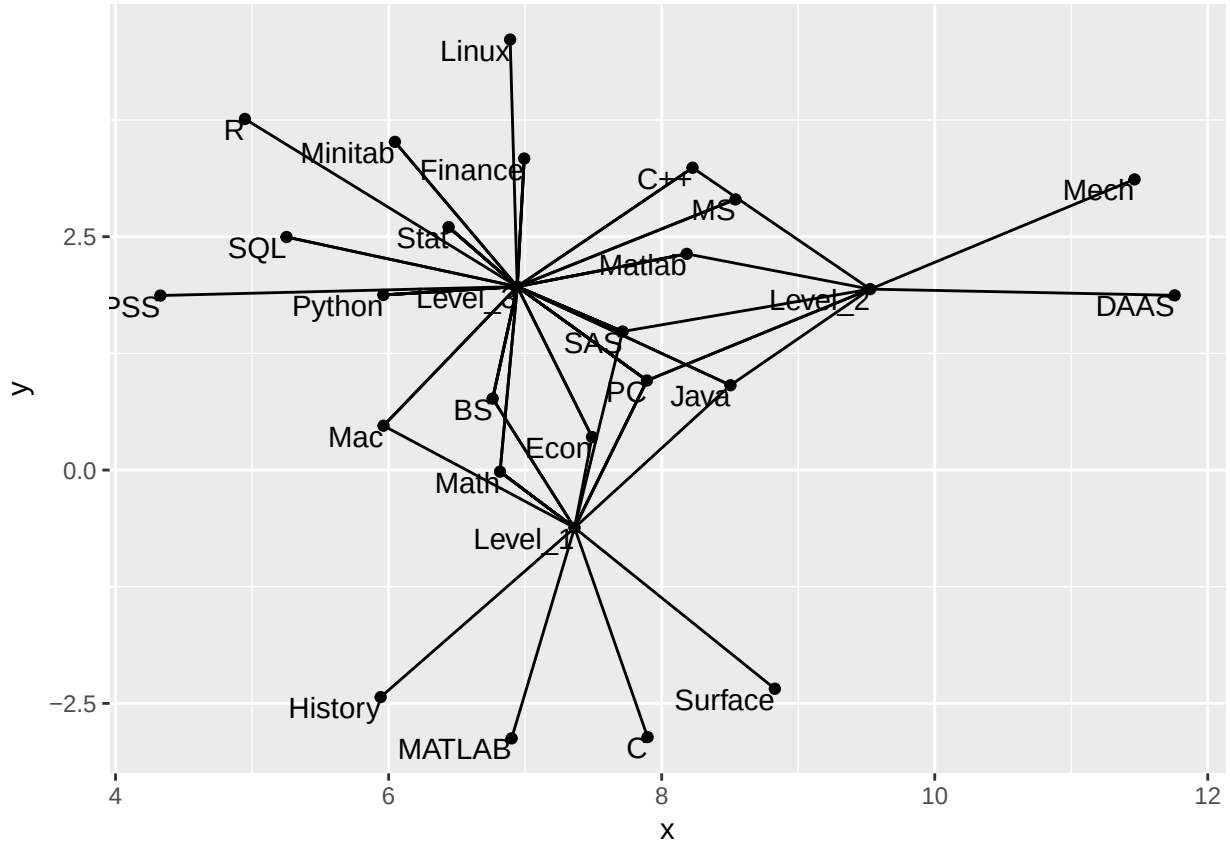
```
GroupData2 <- select(GroupData,-Attribute)
GroupData3 <- filter(GroupData2, GroupData2$value != "NULL")
```

```
## Warning: package 'bindrcpp' was built under R version 3.3.3
```

```
bigram_tf_idf <- GroupData3 %>%
  count(V1, value) %>%
  bind_tf_idf(value, V1, n) %>%
  arrange(desc(tf_idf))

bigram_graph <- GroupData3 %>%
  graph_from_data_frame()

bigram_graph
```

```
## IGRAPH 5e91d7e DN-- 27 77 --
## + attr: name (v/c)
## + edges from 5e91d7e (vertex names):
##  [1] Level_1->PC      Level_1->Mac     Level_3->PC      Level_3->PC
##  [5] Level_3->PC      Level_1->Surface Level_1->PC      Level_1->PC
##  [9] Level_3->Mac     Level_3->PC      Level_2->PC      Level_3->Mac
## [13] Level_2->PC      Level_3->PC      Level_1->Math    Level_1->Math
## [17] Level_3->Finance Level_3->Stat    Level_3->Stat    Level_1->Math
## [21] Level_1->Math    Level_1->Econ    Level_3->Econ    Level_3->Econ
## [25] Level_2->DAAS    Level_3->Finance Level_2->Mech    Level_3->Math
## [29] Level_1->BS      Level_1->BS      Level_3->BS      Level_3->BS
## + ... omitted several edges
```

```
set.seed(2017)

ggraph(bigram_graph, layout = "fr") +
  geom_edge_link() +
  geom_node_point() +
  geom_node_text(aes(label = name), vjust = 1, hjust = 1)
```



Problem 3

For this problem, it hard for me to modify the analysis structure. So I analyze the connection graph and remove some meaningless keywords to get a more concise graph.

```
metadata <- fromJSON("https://data.nasa.gov/data.json")

nasa_title <- data_frame(id = metadata$dataset$`_id`$`$oid`,
                         title = metadata$dataset$title)
#nasa_title
nasa_desc <- data_frame(id = metadata$dataset$`_id`$`$oid`,
                        desc = metadata$dataset$description)

nasa_keyword <- data_frame(id = metadata$dataset$`_id`$`$oid`,
                           keyword = metadata$dataset$keyword) %>%
  unnest(keyword)

#nasa_keyword

nasa_title <- nasa_title %>%
  unnest_tokens(word, title) %>%
  anti_join(stop_words)
```

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

```
nasa_desc <- nasa_desc %>%
  unnest_tokens(word, desc) %>%
  anti_join(stop_words)
```

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

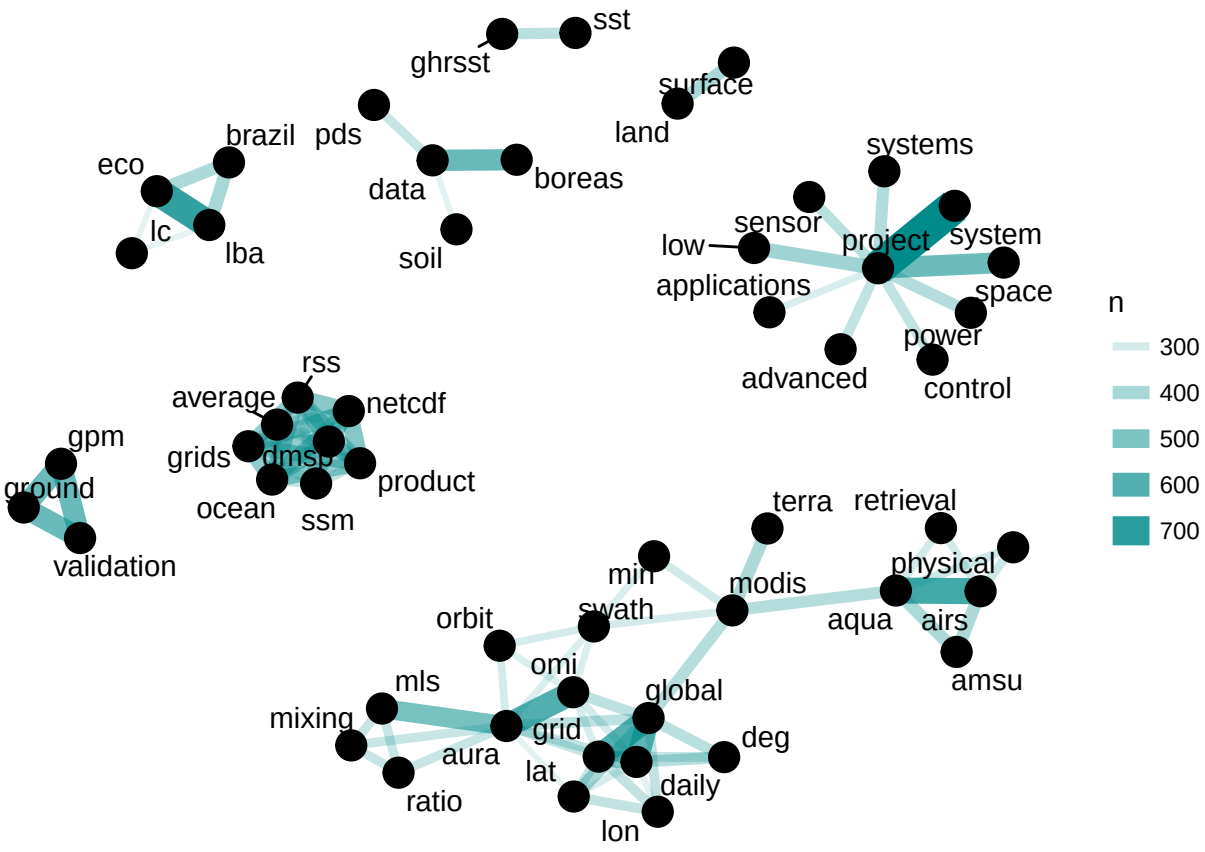
```
#nasa_title

my_stopwords <- data_frame(word = c(as.character(1:11),
                                    "v1", "v03", "l2", "l3", "l4", "v5.2.0",
                                    "v003", "v004", "v005", "v006", "v7",
                                    "2000","total","level","based","degree"))

nasa_title <- nasa_title %>%
  anti_join(my_stopwords)
```

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

```
nasa_desc <- nasa_desc %>%
  anti_join(my_stopwords)
```

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

```
nasa_keyword <- nasa_keyword %>%
  mutate(keyword = toupper(keyword))


title_word_pairs <- nasa_title %>%
  pairwise_count(word, id, sort = TRUE, upper = FALSE)

#title_word_pairs

desc_word_pairs <- nasa_desc %>%
  pairwise_count(word, id, sort = TRUE, upper = FALSE)

#desc_word_pairs


set.seed(1234)
title_word_pairs %>%
  filter(n >= 250) %>%
  graph_from_data_frame() %>%
  ggraph(layout = "fr") +
  geom_edge_link(aes(edge_alpha = n, edge_width = n), edge_colour = "cyan4") +
  geom_node_point(size = 5) +
  geom_node_text(aes(label = name), repel = TRUE,
                 point.padding = unit(0.2, "lines")) +
  theme_void()
```


sst
ghrsst
surface
land
brazil
pds
eco
data
boreas
lc
lba
soil
systems
sensor
project
system
low
applications
space
power
advanced
control
rss
average
netcdf
gpm
grids
dmsp
product
ground
ocean
ssm
validation
terra
retrieval
min
modis
physical
orbit
swath
aqua
airs
omi
amsu
mls
global
mixing
grid
deg
aura
lat
daily
ratio
lon
n
300
400
500
600
700