Some Practice for R

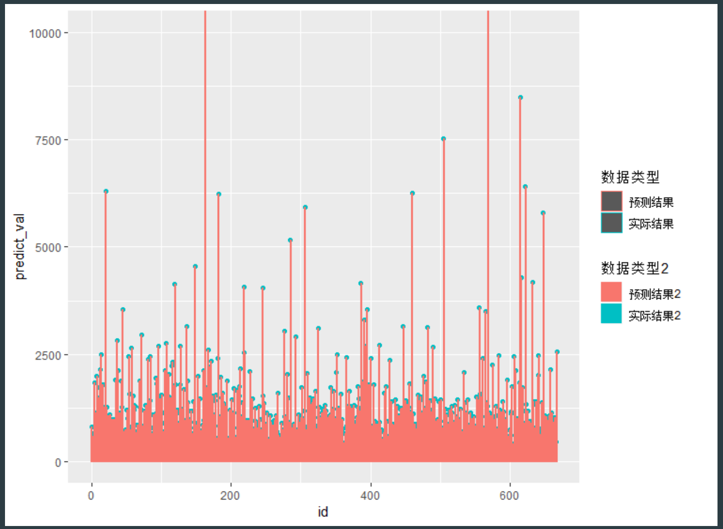
Yiwei Zheng

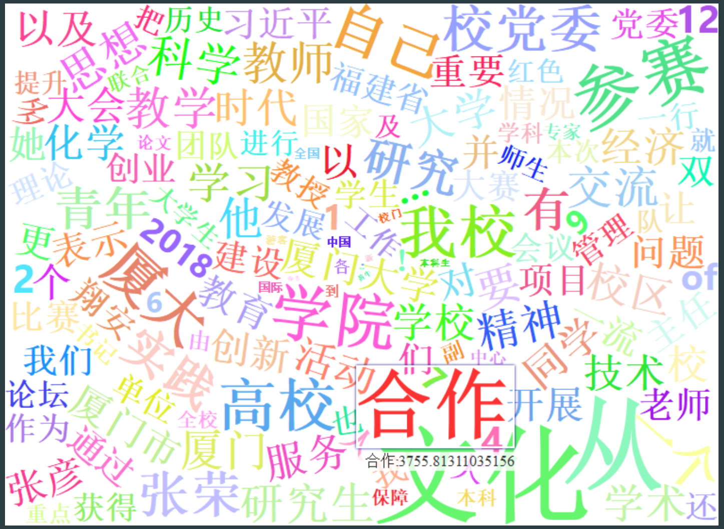
郑逸炜

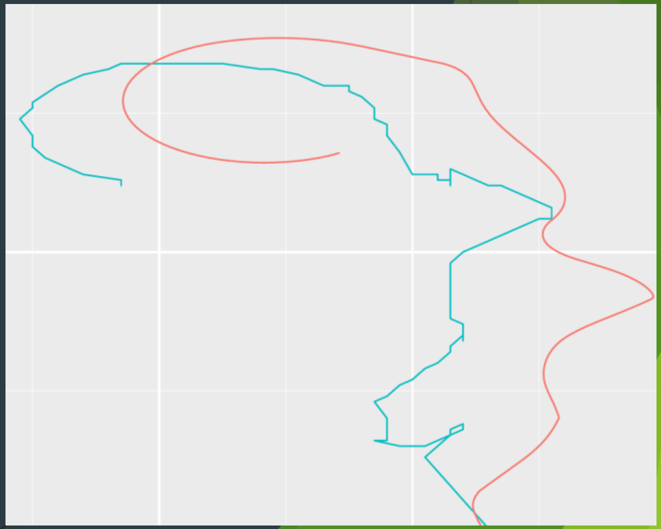
23220181152387

Practice Content & Codes

- Some modification on the homework of the curriculum Bigdata, including:
  - real/predict data for the page view of Xiamen university (XMU) news
  - word cloud of those news
- Plotting the trajectory estimated by the recent work of simultaneous localization and mapping (SLAM)





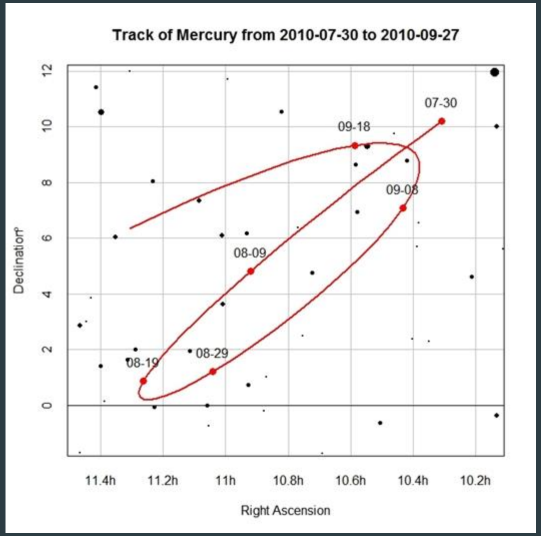


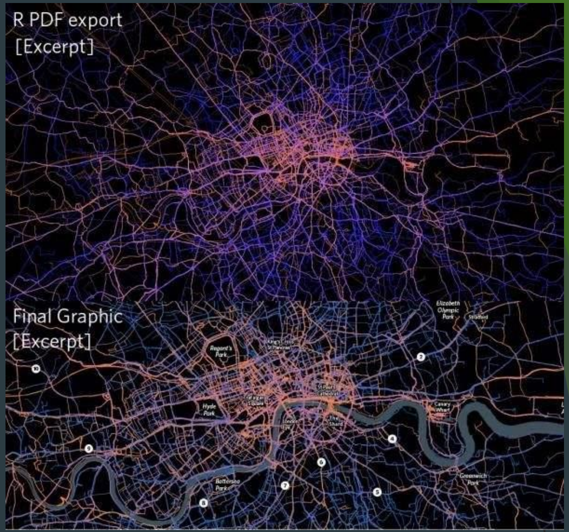
Practice Content & Codes

- Two parts of the codes
- Bigdata & data mining are in
  - https://github.com/trigger1996/R_homework
- SLAM trajectory plotting is in
  - https://github.com/trigger1996/trajectory_plotter_demo

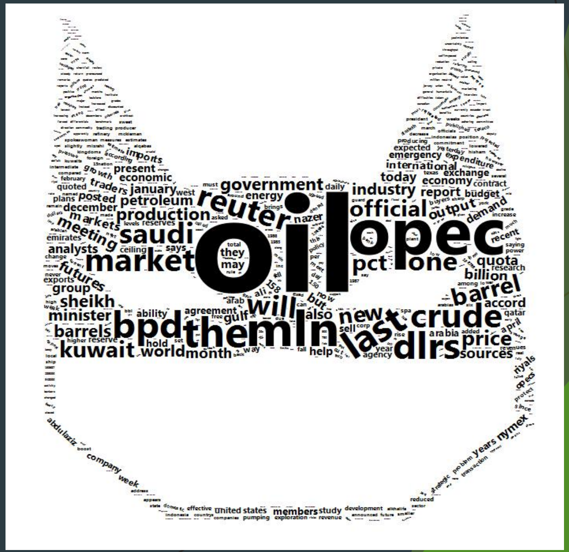
Why R plotting?

- Needs of curriculum
- Indeed beautiful



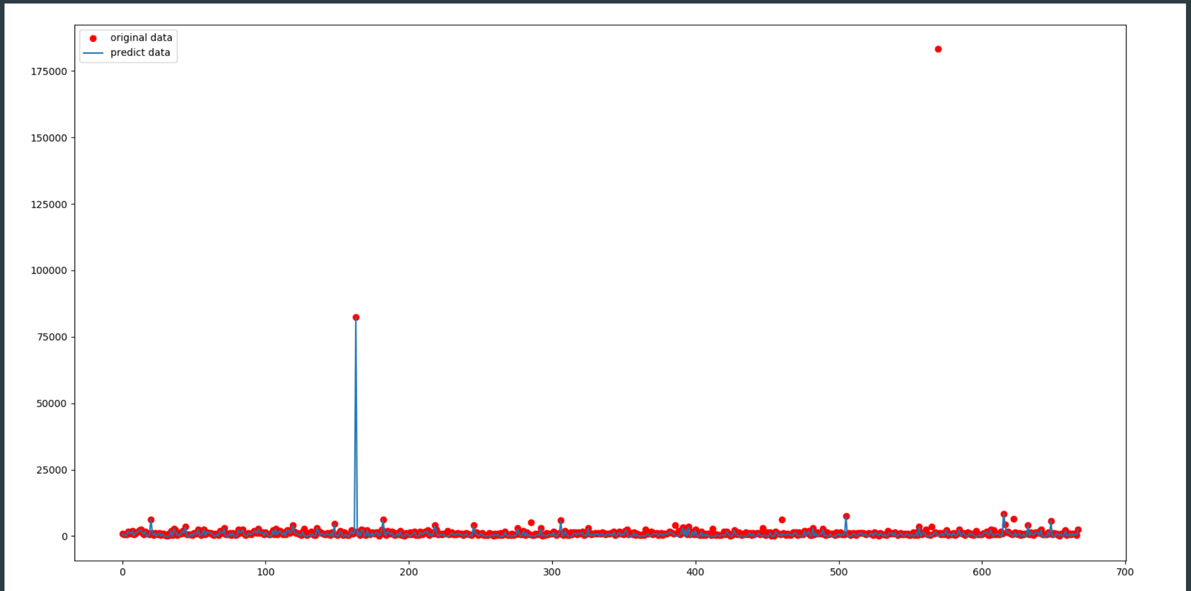

R PDF export
[Excerpt]

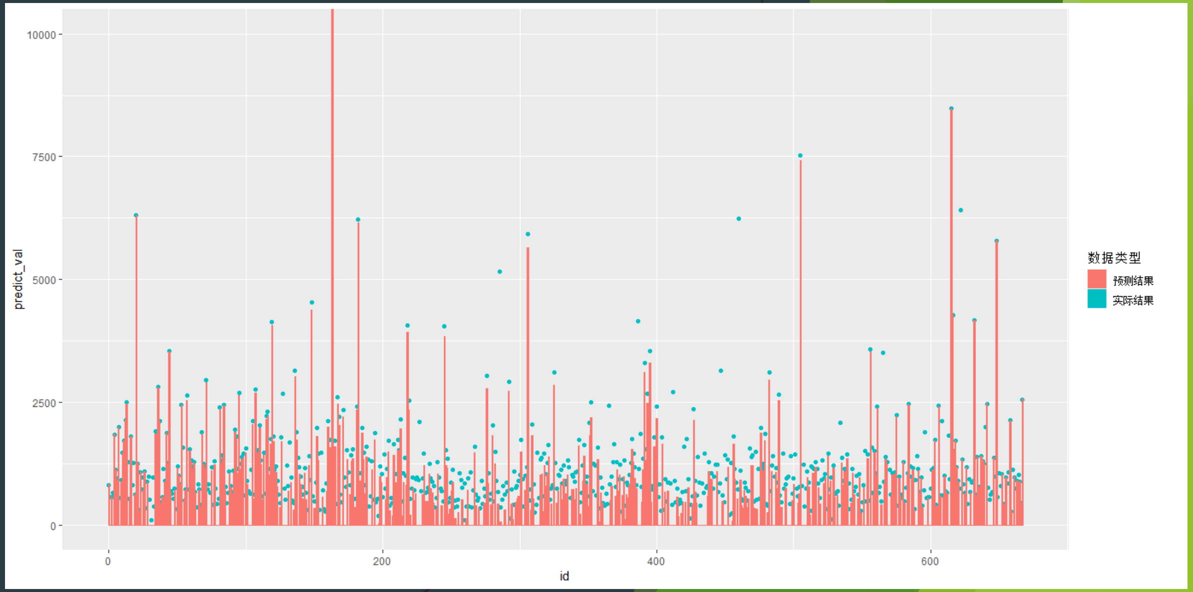
Final Graphic
[Excerpt]



Why R plotting?

- Python result Vs R result
  - Brighter colors
  - Available cutoff

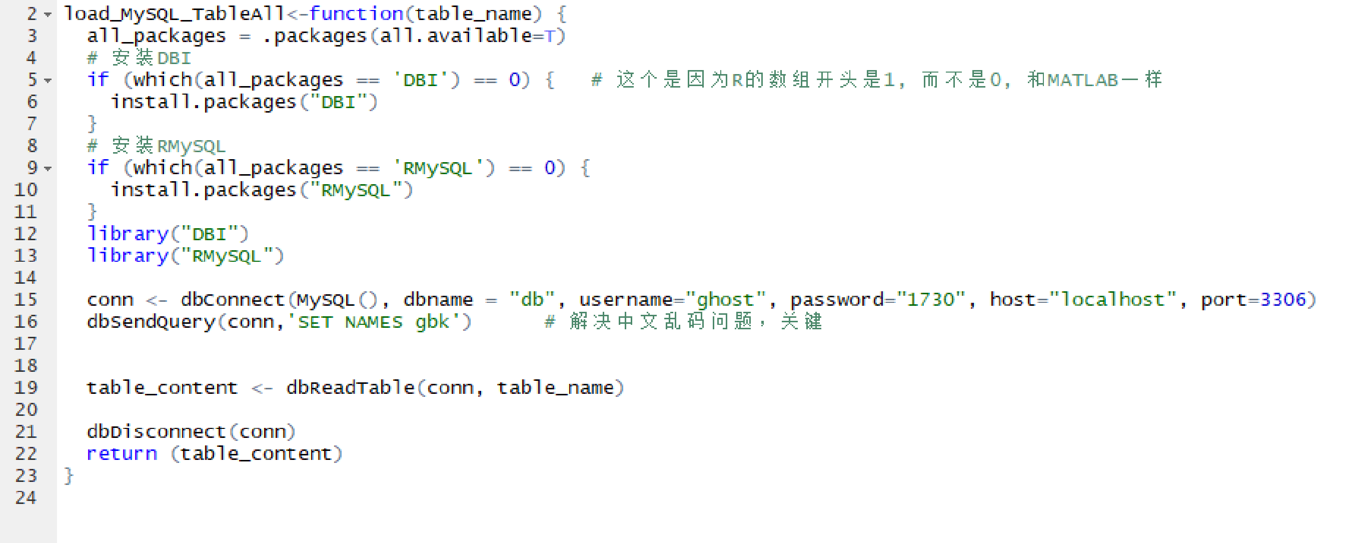




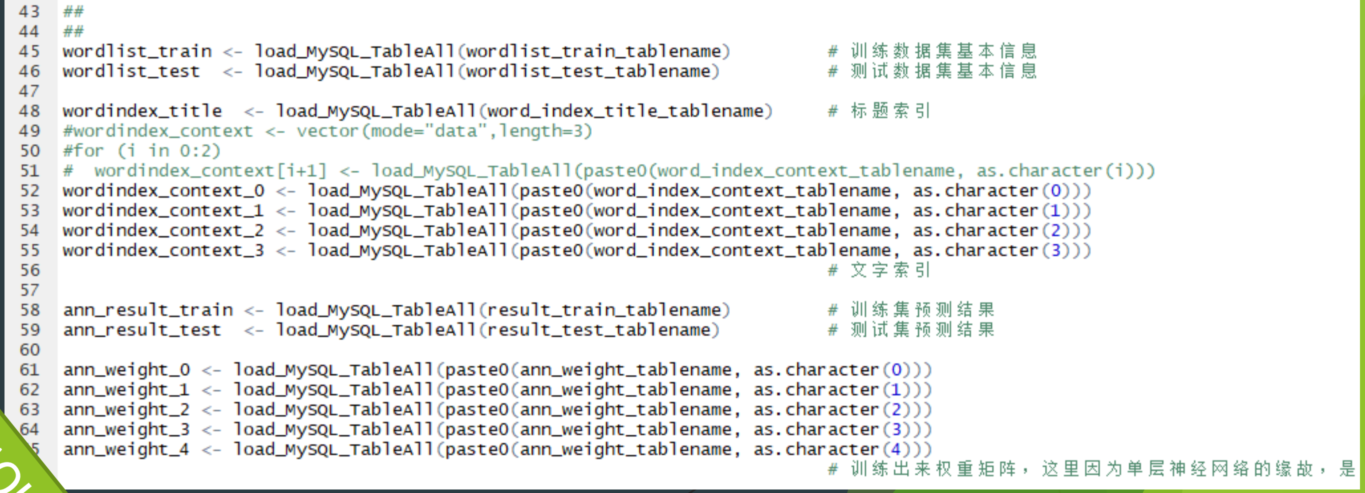
Process of Plotting

- Python & R interaction
  - How to transfer data from python to R?
    - MySQL RMySQL
    - Text document Import data
  - Known issues
    - GBK to UTF8
    - 1000 limits in MySQL

JUST Click your MOUSE!



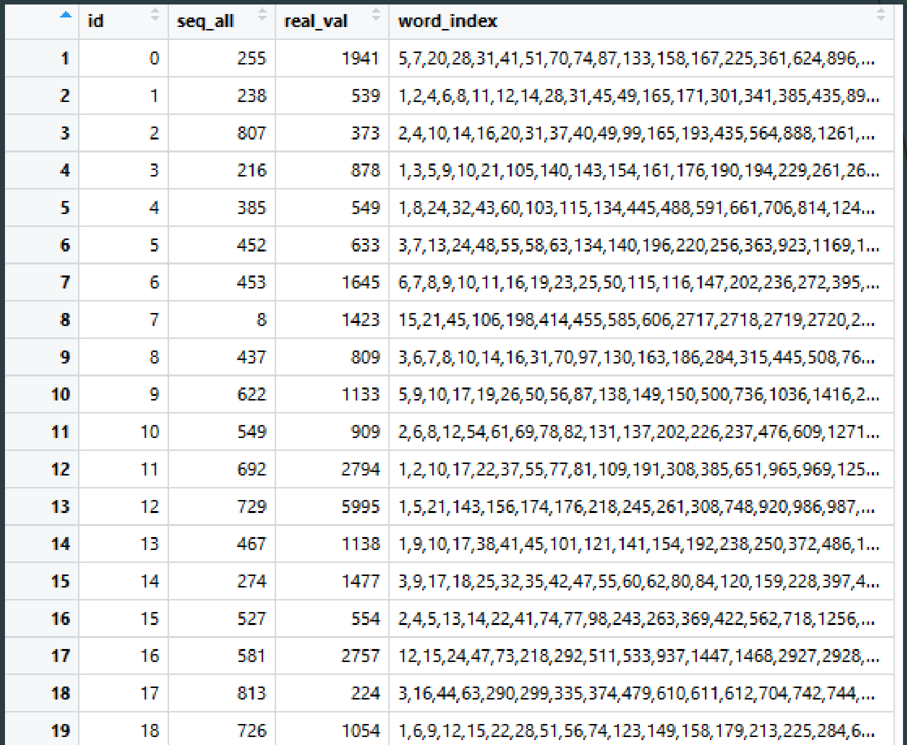
```r
load_MySQL_TableAll<-function(table_name) {
  all_packages = .packages(all.available=T)
  # 安装DBI
  if (which(all_packages == 'DBI') == 0) {   # 这个是因为R的数组开头是1, 而不是0, 和MATLAB一样
    install.packages("DBI")
  }
  # 安装RMySQL
  if (which(all_packages == 'RMySQL') == 0) {
    install.packages("RMySQL")
  }
  library("DBI")
  library("RMySQL")

  conn <- dbConnect(MySQL(), dbname = "db", username="ghost", password="1730", host="localhost", port=3306)
  dbSendQuery(conn,'SET NAMES gbk')      # 解决中文乱码问题, 关键


  table_content <- dbReadTable(conn, table_name)

  dbDisconnect(conn)
  return (table_content)
}

```



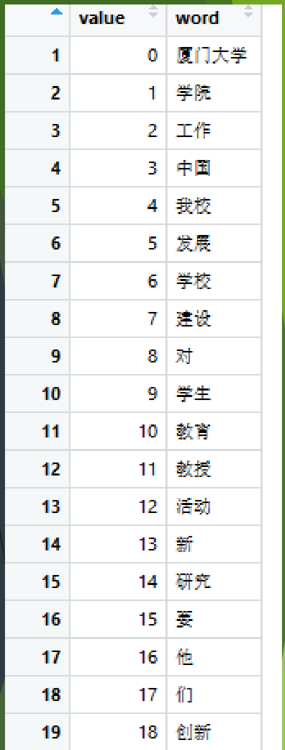
```r
##
##
wordlist_train <- load_MySQL_TableAll(wordlist_train_tablename)         # 训练数据集基本信息
wordlist_test  <- load_MySQL_TableAll(wordlist_test_tablename)          # 测试数据集基本信息

wordindex_title  <- load_MySQL_TableAll(word_index_title_tablename)     # 标题索引
#wordindex_context <- vector(mode="data",length=3)
#for (i in 0:2)
#  wordindex_context[i+1] <- load_MySQL_TableAll(paste0(word_index_context_tablename, as.character(i)))
wordindex_context_0 <- load_MySQL_TableAll(paste0(word_index_context_tablename, as.character(0)))
wordindex_context_1 <- load_MySQL_TableAll(paste0(word_index_context_tablename, as.character(1)))
wordindex_context_2 <- load_MySQL_TableAll(paste0(word_index_context_tablename, as.character(2)))
wordindex_context_3 <- load_MySQL_TableAll(paste0(word_index_context_tablename, as.character(3)))
                                                                        # 文字索引

ann_result_train <- load_MySQL_TableAll(result_train_tablename)         # 训练集预测结果
ann_result_test  <- load_MySQL_TableAll(result_test_tablename)          # 测试集预测结果

ann_weight_0 <- load_MySQL_TableAll(paste0(ann_weight_tablename, as.character(0)))
ann_weight_1 <- load_MySQL_TableAll(paste0(ann_weight_tablename, as.character(1)))
ann_weight_2 <- load_MySQL_TableAll(paste0(ann_weight_tablename, as.character(2)))
ann_weight_3 <- load_MySQL_TableAll(paste0(ann_weight_tablename, as.character(3)))
ann_weight_4 <- load_MySQL_TableAll(paste0(ann_weight_tablename, as.character(4)))
                                                                        # 训练出来权重矩阵, 这里因为单层神经网络的缘故, 是
```

Process of Plotting

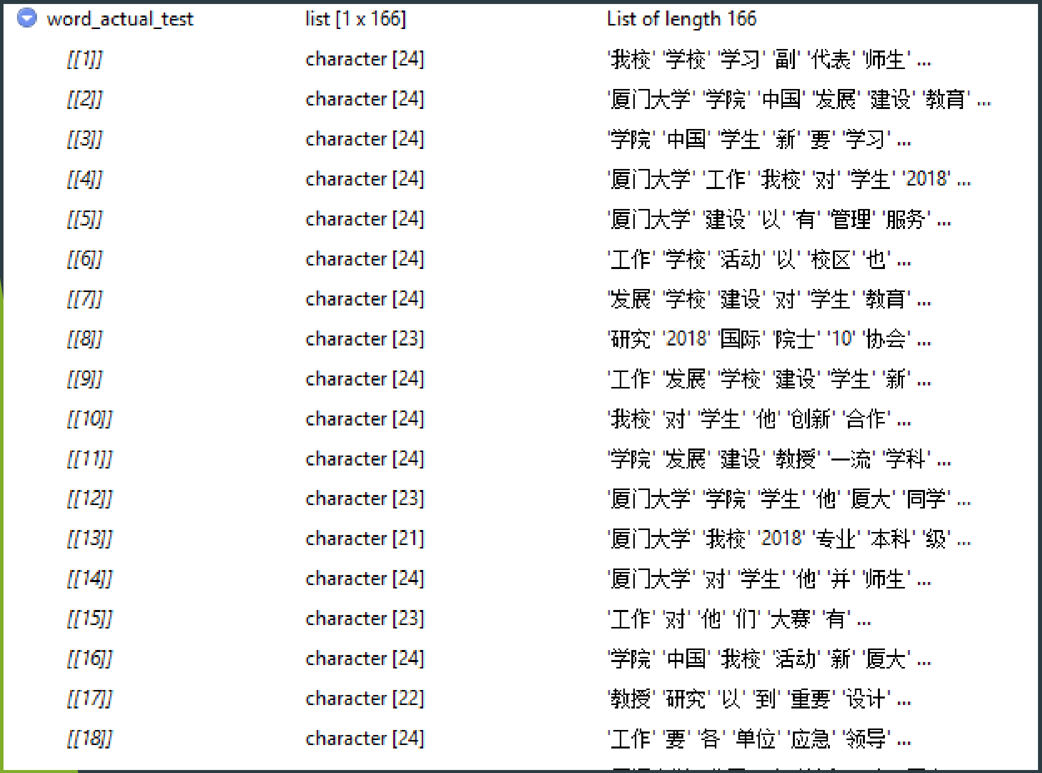
- Data structure transform for words



	id	seq_all	real_val	word_index
1	0	255	1941	5,7,20,28,31,41,51,70,74,87,133,158,167,225,361,624,896,...
2	1	238	539	1,2,4,6,8,11,12,14,28,31,45,49,165,171,301,341,385,435,89...
3	2	807	373	2,4,10,14,16,20,31,37,40,49,99,165,193,435,564,888,1261,...
4	3	216	878	1,3,5,9,10,21,105,140,143,154,161,176,190,194,229,261,26...
5	4	385	549	1,8,24,32,43,60,103,115,134,445,488,591,661,706,814,124...
6	5	452	633	3,7,13,24,48,55,58,63,134,140,196,220,256,363,923,1169,1...
7	6	453	1645	6,7,8,9,10,11,16,19,23,25,50,115,116,147,202,236,272,395,...
8	7	8	1423	15,21,45,106,198,414,455,585,606,2717,2718,2719,2720,2...
9	8	437	809	3,6,7,8,10,14,16,31,70,97,130,163,186,284,315,445,508,76...
10	9	622	1133	5,9,10,17,19,26,50,56,87,138,149,150,500,736,1036,1416,2...
11	10	549	909	2,6,8,12,54,61,69,78,82,131,137,202,226,237,476,609,1271...
12	11	692	2794	1,2,10,17,22,37,55,77,81,109,191,308,385,651,965,969,125...
13	12	729	5995	1,5,21,143,156,174,176,218,245,261,308,748,920,986,987,...
14	13	467	1138	1,9,10,17,38,41,45,101,121,141,154,192,238,250,372,486,1...
15	14	274	1477	3,9,17,18,25,32,35,42,47,55,60,62,80,84,120,159,228,397,4...
16	15	527	554	2,4,5,13,14,22,41,74,77,98,243,263,369,422,562,718,1256,...
17	16	581	2757	12,15,24,47,73,218,292,511,533,937,1447,1468,2927,2928,...
18	17	813	224	3,16,44,63,290,299,335,374,479,610,611,612,704,742,744,...
19	18	726	1054	1,6,9,12,15,22,28,51,56,74,123,149,158,179,213,225,284,6...



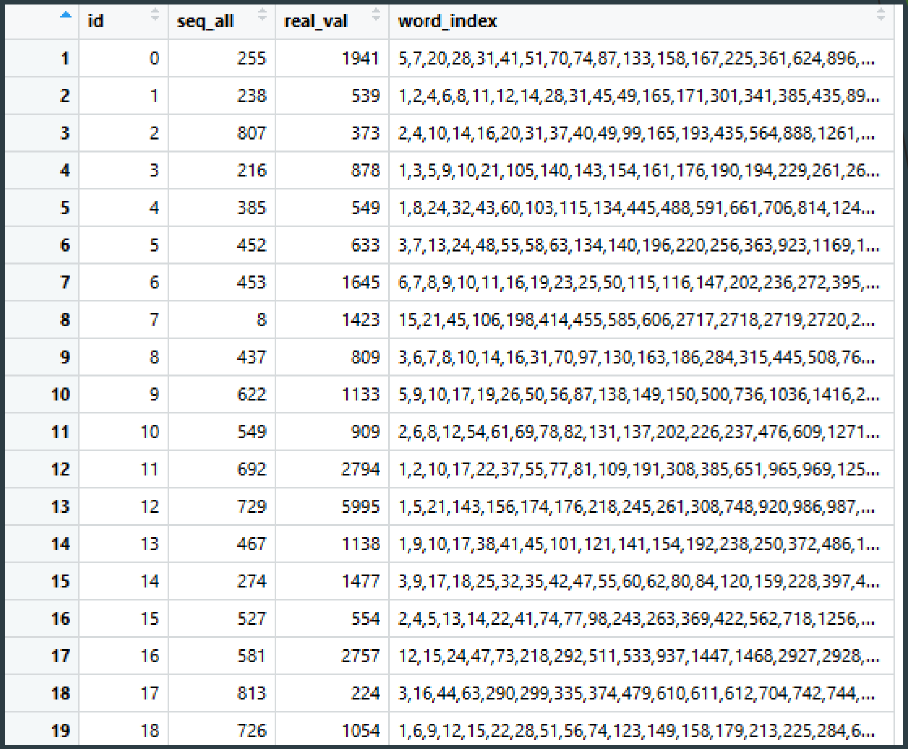
	value	word
1	0	厦门大学
2	1	学院
3	2	工作
4	3	中国
5	4	我校
6	5	发展
7	6	学校
8	7	建设
9	8	对
10	9	学生
11	10	教育
12	11	教授
13	12	活动
14	13	新
15	14	研究
16	15	要
17	16	他
18	17	们
19	18	创新



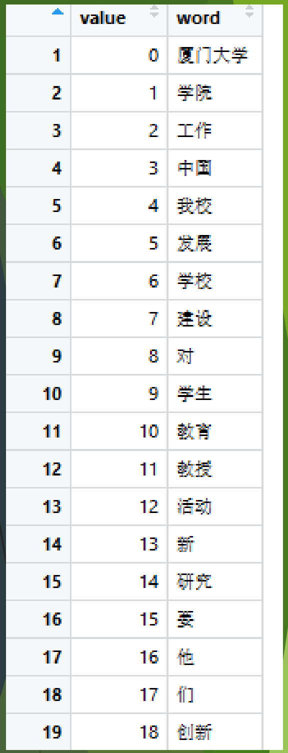
```
word_actual_test      list [1 x 166]      List of length 166
  [[1]]               character [24]      '我校' '学校' '学习' '副' '代表' '师生' ...
  [[2]]               character [24]      '厦门大学' '学院' '中国' '发展' '建设' '教育' ...
  [[3]]               character [24]      '学院' '中国' '学生' '新' '要' '学习' ...
  [[4]]               character [24]      '厦门大学' '工作' '我校' '对' '学生' '2018' ...
  [[5]]               character [24]      '厦门大学' '建设' '以' '有' '管理' '服务' ...
  [[6]]               character [24]      '工作' '学校' '活动' '以' '校区' '也' ...
  [[7]]               character [24]      '发展' '学校' '建设' '对' '学生' '教育' ...
  [[8]]               character [23]      '研究' '2018' '国际' '院士' '10' '协会' ...
  [[9]]               character [24]      '工作' '发展' '学校' '建设' '学生' '新' ...
  [[10]]              character [24]      '我校' '对' '学生' '他' '创新' '合作' ...
  [[11]]              character [24]      '学院' '发展' '建设' '教授' '一流' '学科' ...
  [[12]]              character [23]      '厦门大学' '学院' '学生' '他' '厦大' '同学' ...
  [[13]]              character [21]      '厦门大学' '我校' '2018' '专业' '本科' '级' ...
  [[14]]              character [24]      '厦门大学' '对' '学生' '他' '并' '师生' ...
  [[15]]              character [23]      '工作' '对' '他' '们' '大赛' '有' ...
  [[16]]              character [24]      '学院' '中国' '我校' '活动' '新' '厦大' ...
  [[17]]              character [22]      '教授' '研究' '以' '到' '重要' '设计' ...
  [[18]]              character [24]      '工作' '要' '各' '单位' '应急' '领导' ...
```

Process of Plotting

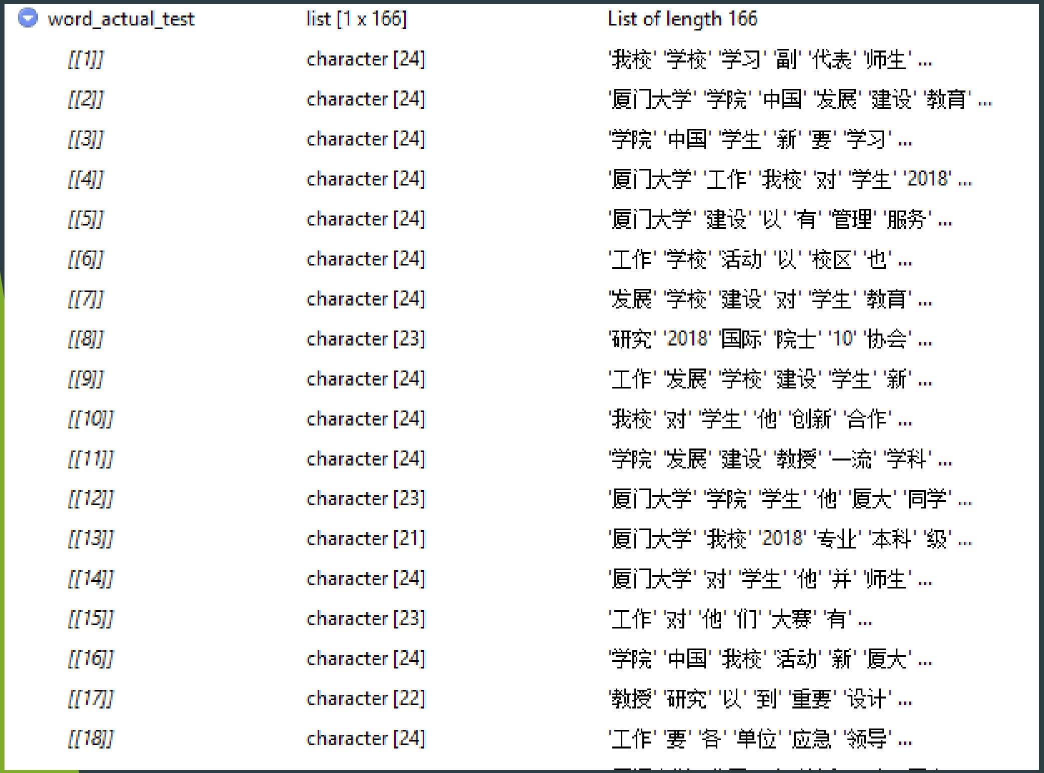
- Data structure transform for words



	id	seq_all	real_val	word_index
1	0	255	1941	5,7,20,28,31,41,51,70,74,87,133,158,167,225,361,624,896,...
2	1	238	539	1,2,4,6,8,11,12,14,28,31,45,49,165,171,301,341,385,435,89...
3	2	807	373	2,4,10,14,16,20,31,37,40,49,99,165,193,435,564,888,1261,...
4	3	216	878	1,3,5,9,10,21,105,140,143,154,161,176,190,194,229,261,26...
5	4	385	549	1,8,24,32,43,60,103,115,134,445,488,591,661,706,814,124...
6	5	452	633	3,7,13,24,48,55,58,63,134,140,196,220,256,363,923,1169,1...
7	6	453	1645	6,7,8,9,10,11,16,19,23,25,50,115,116,147,202,236,272,395,...
8	7	8	1423	15,21,45,106,198,414,455,585,606,2717,2718,2719,2720,2...
9	8	437	809	3,6,7,8,10,14,16,31,70,97,130,163,186,284,315,445,508,76...
10	9	622	1133	5,9,10,17,19,26,50,56,87,138,149,150,500,736,1036,1416,2...
11	10	549	909	2,6,8,12,54,61,69,78,82,131,137,202,226,237,476,609,1271...
12	11	692	2794	1,2,10,17,22,37,55,77,81,109,191,308,385,651,965,969,125...
13	12	729	5995	1,5,21,143,156,174,176,218,245,261,308,748,920,986,987,...
14	13	467	1138	1,9,10,17,38,41,45,101,121,141,154,192,238,250,372,486,1...
15	14	274	1477	3,9,17,18,25,32,35,42,47,55,60,62,80,84,120,159,228,397,4...
16	15	527	554	2,4,5,13,14,22,41,74,77,98,243,263,369,422,562,718,1256,...
17	16	581	2757	12,15,24,47,73,218,292,511,533,937,1447,1468,2927,2928,...
18	17	813	224	3,16,44,63,290,299,335,374,479,610,611,612,704,742,744,...
19	18	726	1054	1,6,9,12,15,22,28,51,56,74,123,149,158,179,213,225,284,6...

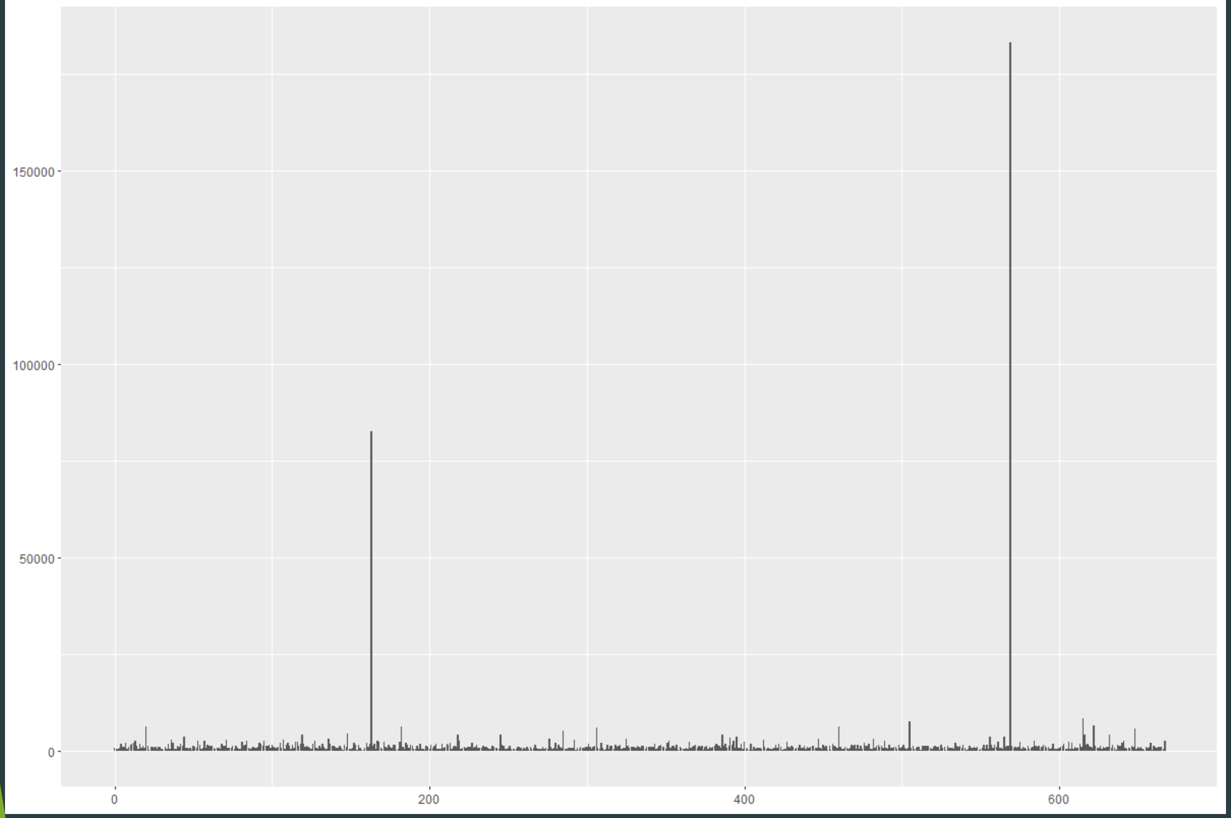


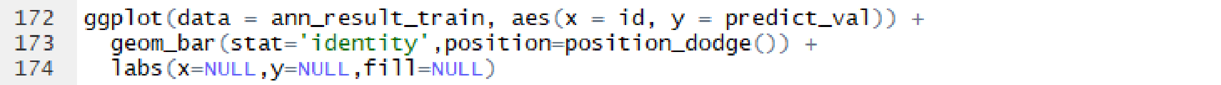
	value	word
1	0	厦门大学
2	1	学院
3	2	工作
4	3	中国
5	4	我校
6	5	发展
7	6	学校
8	7	建设
9	8	对
10	9	学生
11	10	教育
12	11	教授
13	12	活动
14	13	新
15	14	研究
16	15	要
17	16	他
18	17	们
19	18	创新



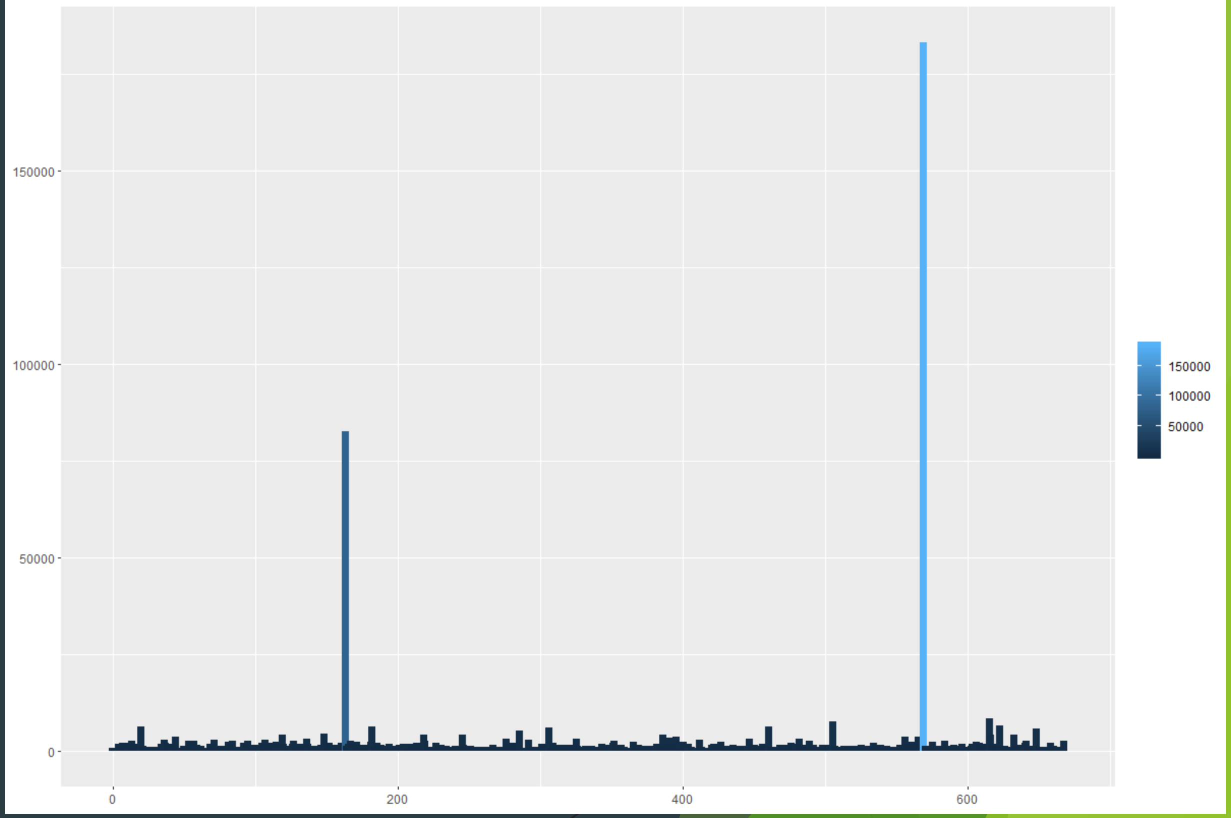
word_actual_test	list [1 x 166]	List of length 166
[[1]]	character [24]	'我校' '学校' '学习' '副' '代表' '师生' ...
[[2]]	character [24]	'厦门大学' '学院' '中国' '发展' '建设' '教育' ...
[[3]]	character [24]	'学院' '中国' '学生' '新' '要' '学习' ...
[[4]]	character [24]	'厦门大学' '工作' '我校' '对' '学生' '2018' ...
[[5]]	character [24]	'厦门大学' '建设' '以' '有' '管理' '服务' ...
[[6]]	character [24]	'工作' '学校' '活动' '以' '校区' '也' ...
[[7]]	character [24]	'发展' '学校' '建设' '对' '学生' '教育' ...
[[8]]	character [23]	'研究' '2018' '国际' '院士' '10' '协会' ...
[[9]]	character [24]	'工作' '发展' '学校' '建设' '学生' '新' ...
[[10]]	character [24]	'我校' '对' '学生' '他' '创新' '合作' ...
[[11]]	character [24]	'学院' '发展' '建设' '教授' '一流' '学科' ...
[[12]]	character [23]	'厦门大学' '学院' '学生' '他' '厦大' '同学' ...
[[13]]	character [21]	'厦门大学' '我校' '2018' '专业' '本科' '级' ...
[[14]]	character [24]	'厦门大学' '对' '学生' '他' '并' '师生' ...
[[15]]	character [23]	'工作' '对' '他' '们' '大赛' '有' ...
[[16]]	character [24]	'学院' '中国' '我校' '活动' '新' '厦大' ...
[[17]]	character [22]	'教授' '研究' '以' '到' '重要' '设计' ...
[[18]]	character [24]	'工作' '要' '各' '单位' '应急' '领导' ...

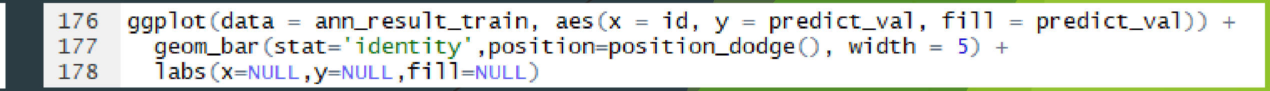
Process of Plotting





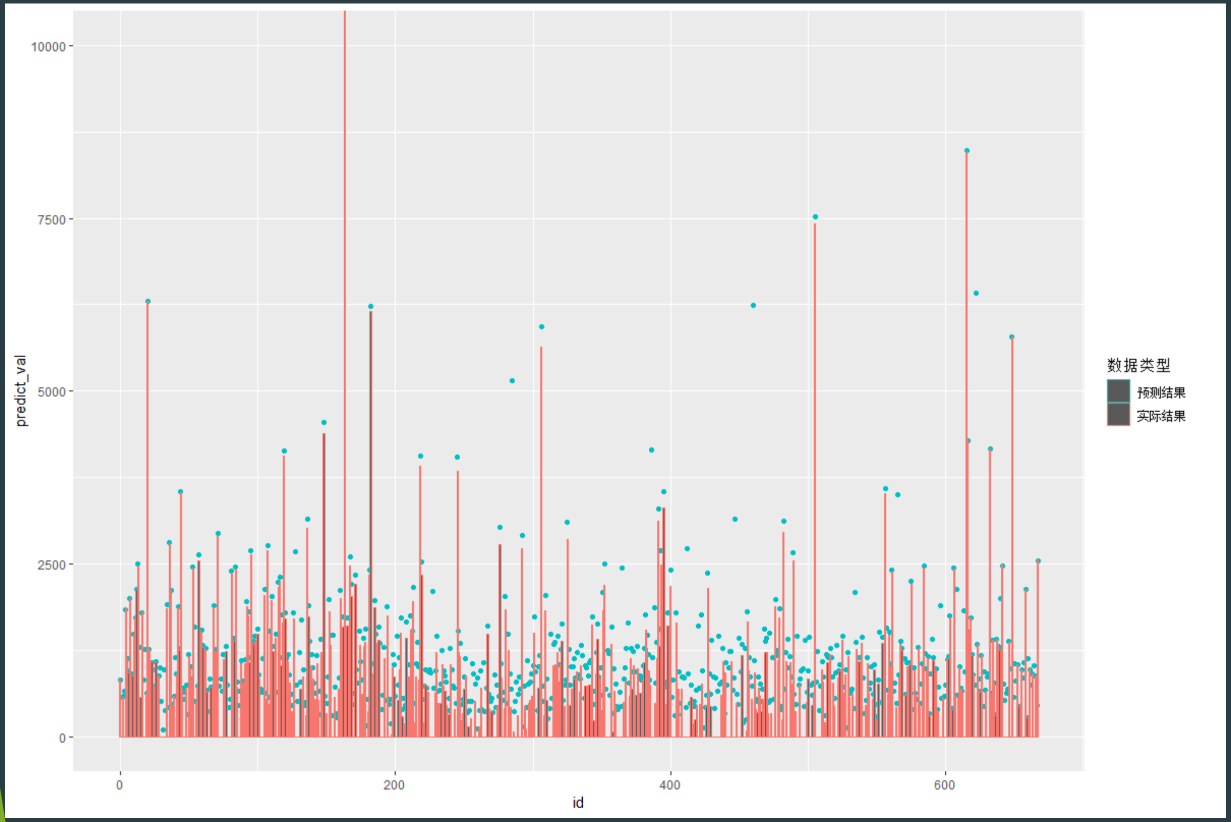
```
ggplot(data = ann_result_train, aes(x = id, y = predict_val)) +
  geom_bar(stat='identity',position=position_dodge()) +
  labs(x=NULL,y=NULL,fill=NULL)
```

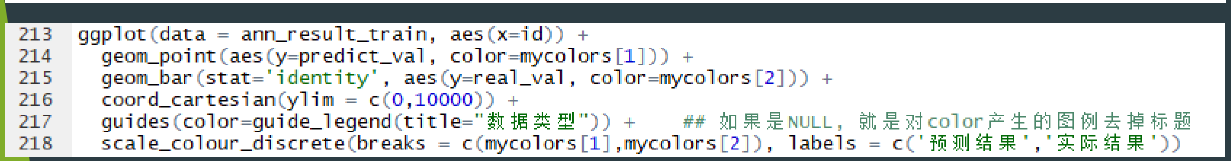




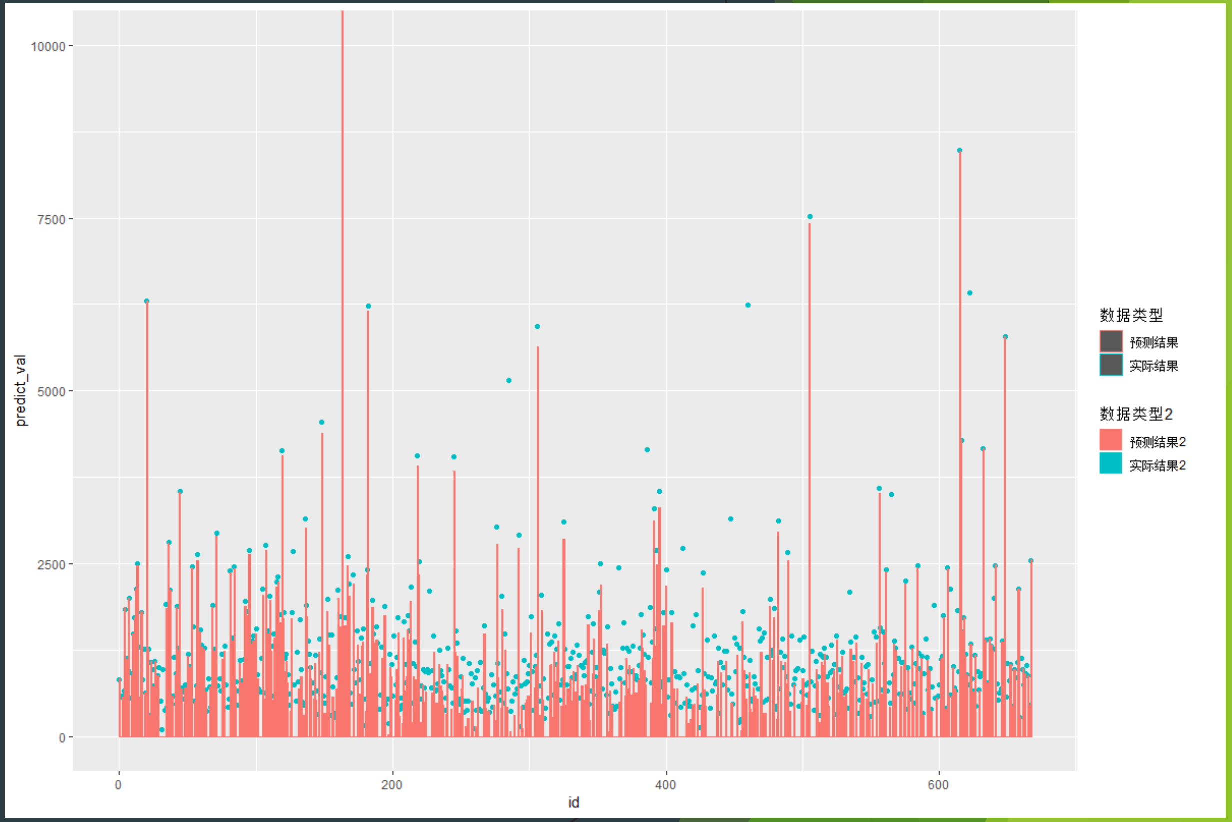
```
ggplot(data = ann_result_train, aes(x = id, y = predict_val, fill = predict_val)) +
  geom_bar(stat='identity',position=position_dodge(), width = 5) +
  labs(x=NULL,y=NULL,fill=NULL)
```

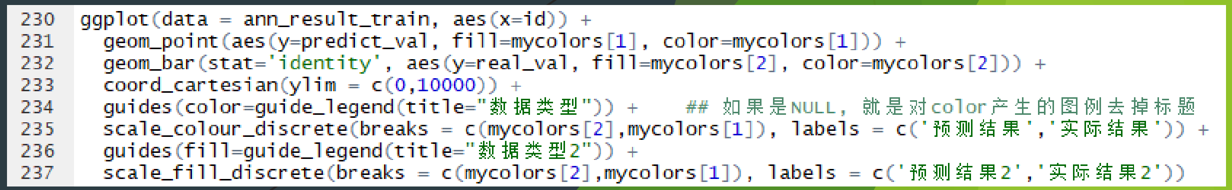
Process of Plotting





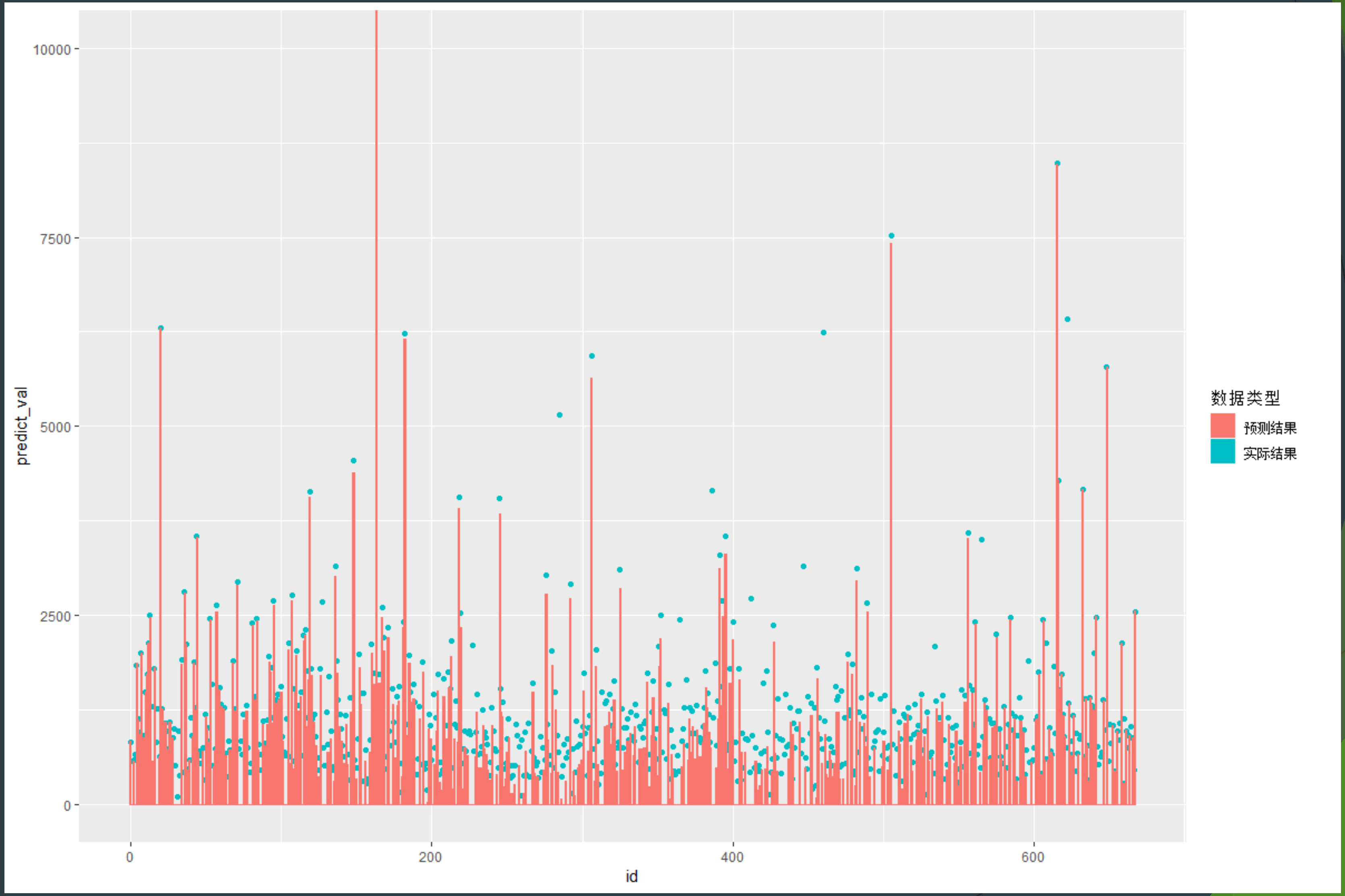
```
ggplot(data = ann_result_train, aes(x=id)) +
  geom_point(aes(y=predict_val, color=mycolors[1])) +
  geom_bar(stat='identity', aes(y=real_val, color=mycolors[2])) +
  coord_cartesian(ylim = c(0,10000)) +
  guides(color=guide_legend(title="数据类型")) +    ## 如果是NULL，就是对color产生的图例去掉标题
  scale_colour_discrete(breaks = c(mycolors[1],mycolors[2]), labels = c('预测结果','实际结果'))
```





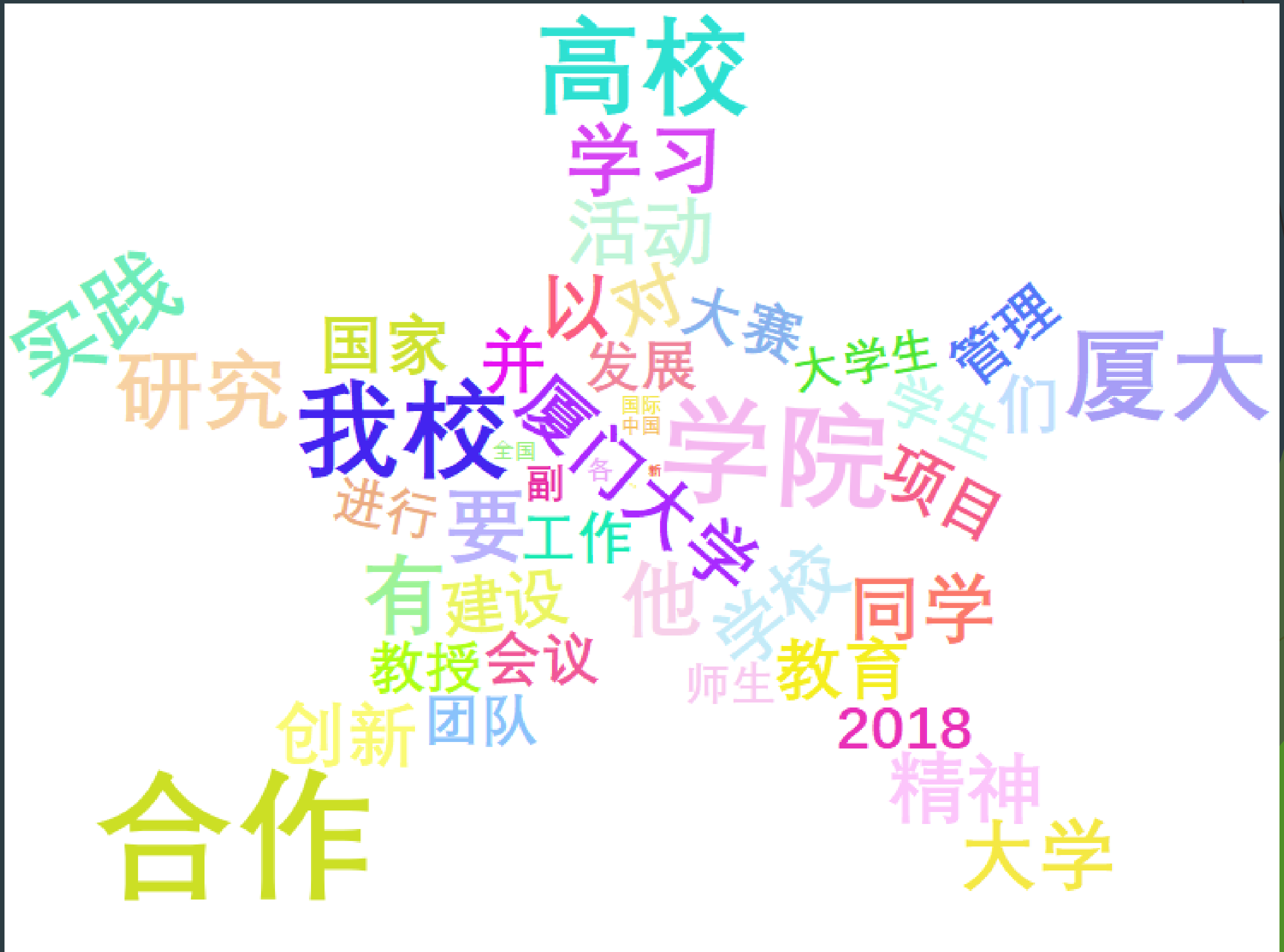
```
ggplot(data = ann_result_train, aes(x=id)) +
  geom_point(aes(y=predict_val, fill=mycolors[1], color=mycolors[1])) +
  geom_bar(stat='identity', aes(y=real_val, fill=mycolors[2], color=mycolors[2])) +
  coord_cartesian(ylim = c(0,10000)) +
  guides(color=guide_legend(title="数据类型")) +    ## 如果是NULL，就是对color产生的图例去掉标题
  scale_colour_discrete(breaks = c(mycolors[2],mycolors[1]), labels = c('预测结果','实际结果')) +
  guides(fill=guide_legend(title="数据类型2")) +
  scale_fill_discrete(breaks = c(mycolors[2],mycolors[1]), labels = c('预测结果2','实际结果2'))
```

Process of Plotting



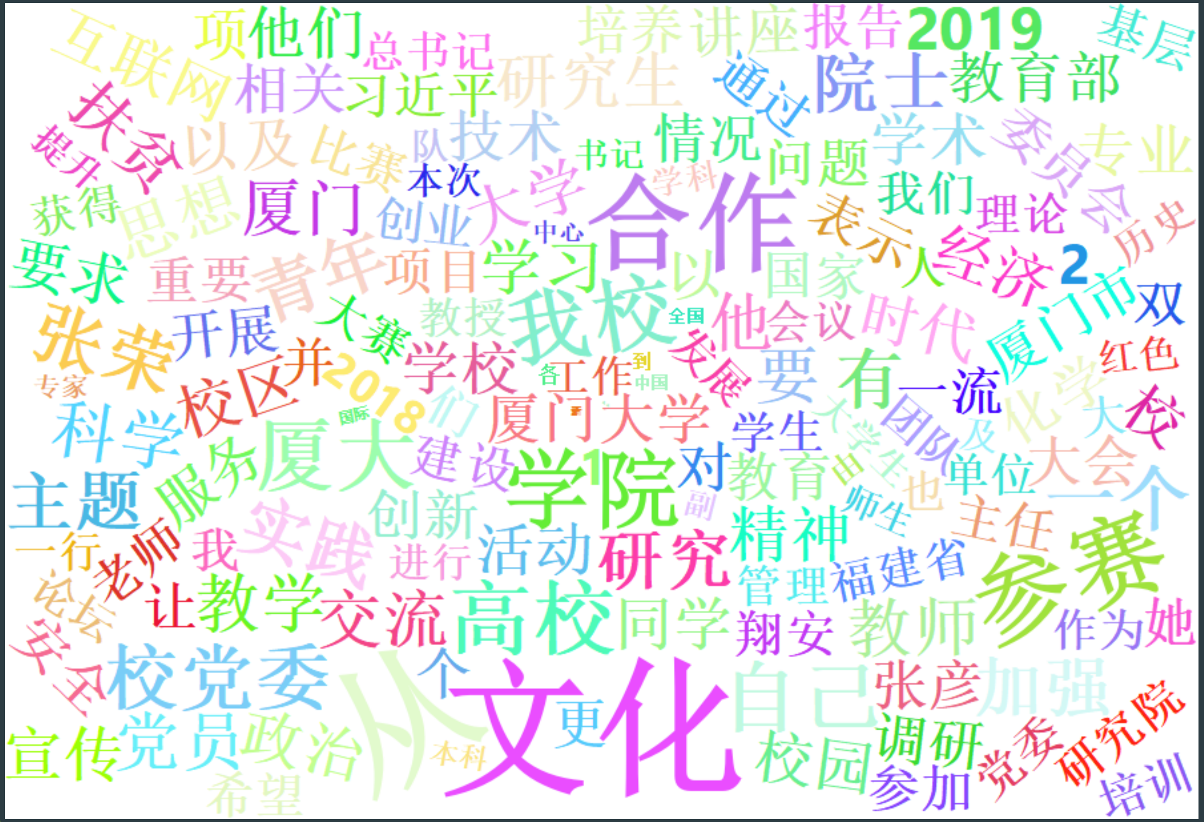
Word Cloud

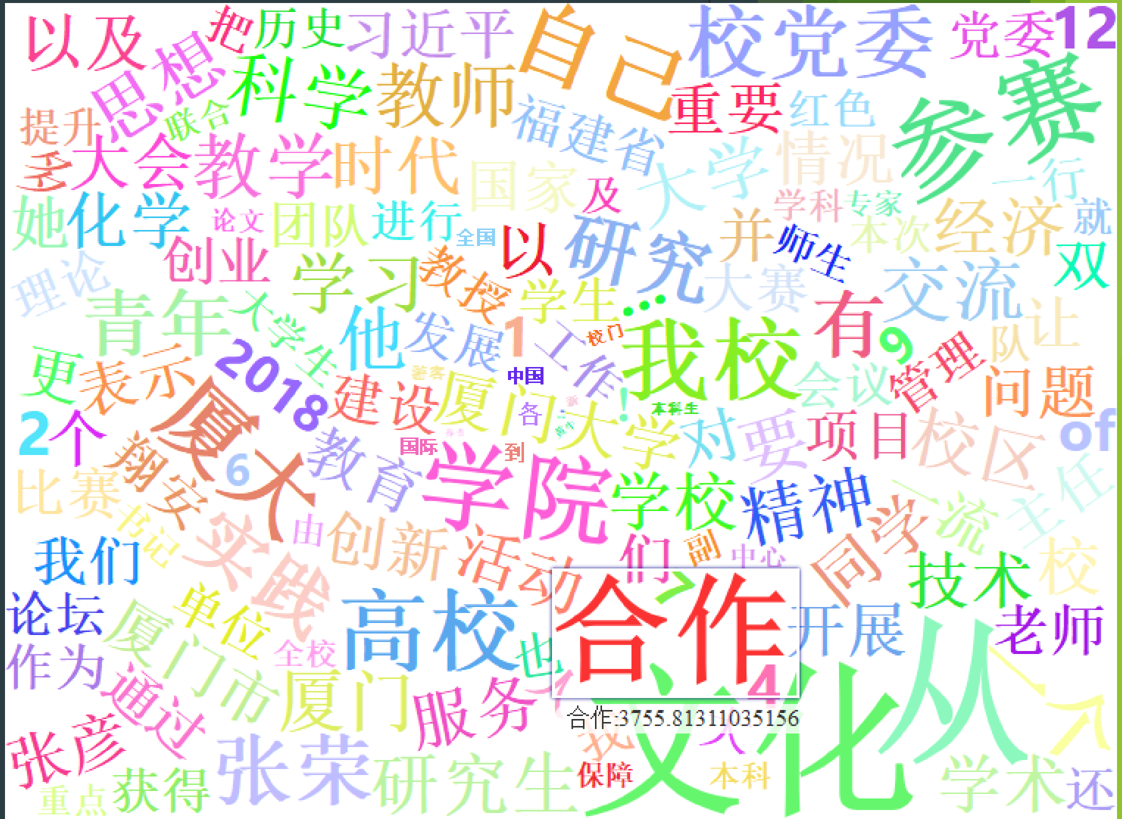
- Content of Word Cloud
  - Word
  - Word frequency



Word Cloud

- Individual word examinable





合作:3755.81311035156

Word Cloud

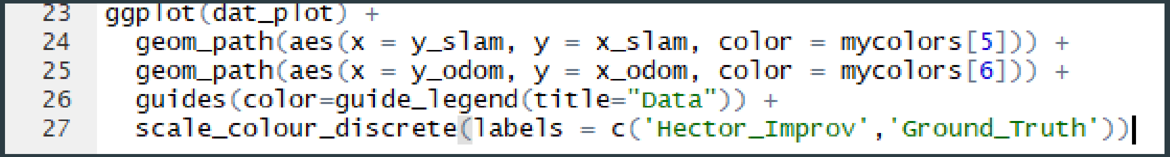
- Shape & size & font customizable

高校
学习
活动
实践
研究
国家
以
对
大赛
大学生
管理
厦大
并
发展
学生
们
我校
厦门大学
学院
项目
国际
中国
全国
副
各
新
进行
要
工作
有
建设
他
学校
同学
教授
会议
师生
教育
创新
团队
2018
精神
合作
大学

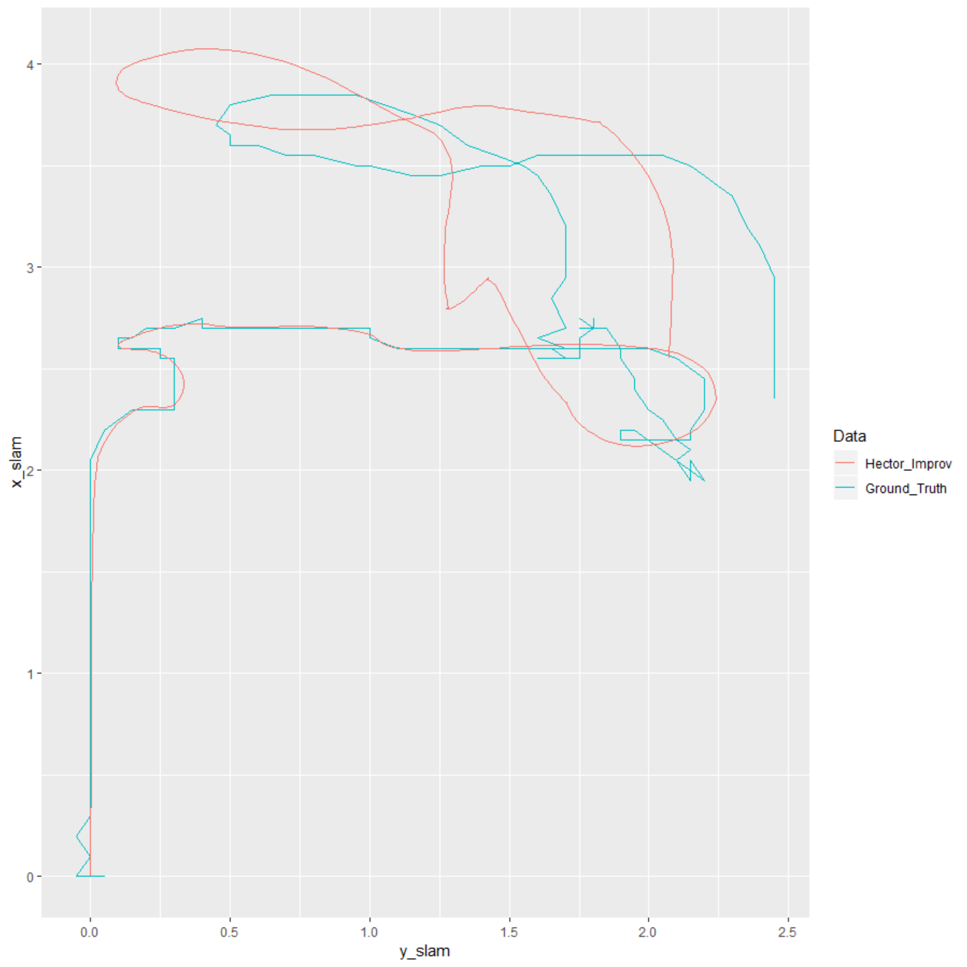
坚持
希望
办学
青春
做好
总书记
社会
报告
大会
更
我
书记
厦门大学
2018
国家
我们
作为
她
10
基层
现场
双方
双
通过
福建省
并
们
工作
发展
以
单位
张荣
党委
级
产业
学员
政治
化学
学科
厦大
学生
学院
1
教授
有
创业
校区
让
校
都
名
院长
参加
也
大赛
我校
要
会议
师生
比赛
校园
领域
了解
老师
表示
项目
教育
学校
建设
活动
队
问题
大学生
管理
提升
知识
举办
院士
重要
翔安
学习
他
研究
创新
实践
教师
思想
互联网
向
邓
改革开放
论坛
厦门
一流
大学
进行
精神
合作
高校
及
开展
厦门市
以及
一行
作品
好
介绍
培养
人
个
教学
同学
团队
时代
科研
办公室
集团
改革
历史
主题
科学
服务
青年
交流
本次
文化
研究生
扶贫
副校长
推进
人才
项
宣传
相关
理论
大
经济
技术
情况
2
习近平
要求
教育部
重点
推动
孔子
该
研究院
红色
参赛
自己
本科
主任
7
张彦
加强
竞赛
海洋
来自
培训
校友
他们
委员会
学术
校党委
一个
讲座
专业
2019
仪式
特色
获得
调研
党员
队伍
科技
分校
方面
实验室
组织
党支部
安全
马来西亚
优秀
世界
马克思主义
队员
12
座谈会
企业
艺术
招生
会
化工学院

Trajectory Plotting

- My CURRENT work
- What is primarily needed?
  - A clear comparison between SLAM trajectory and ground truth

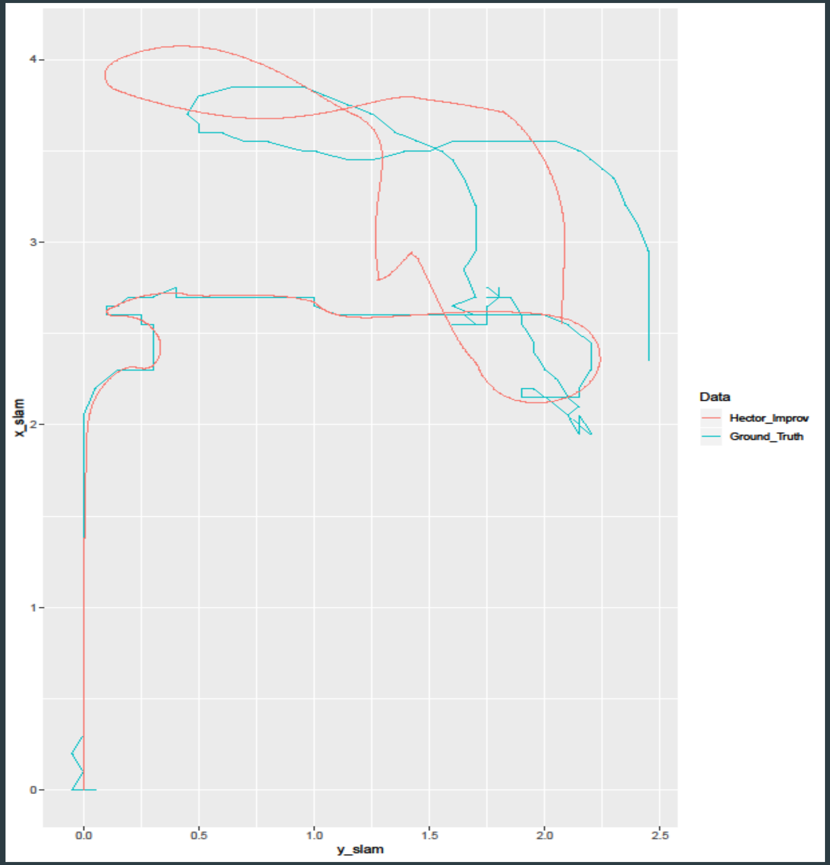


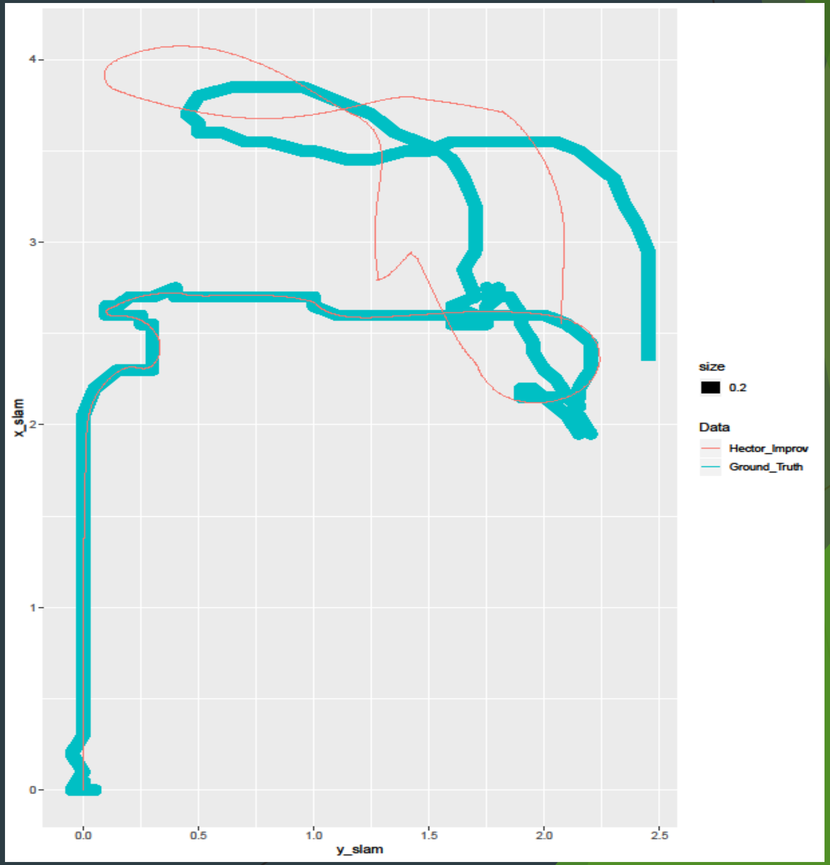
```
ggplot(dat_plot) +
  geom_path(aes(x = y_slam, y = x_slam, color = mycolors[5])) +
  geom_path(aes(x = y_odom, y = x_odom, color = mycolors[6])) +
  guides(color=guide_legend(title="Data")) +
  scale_colour_discrete(labels = c('Hector_Improv','Ground_Truth'))
```



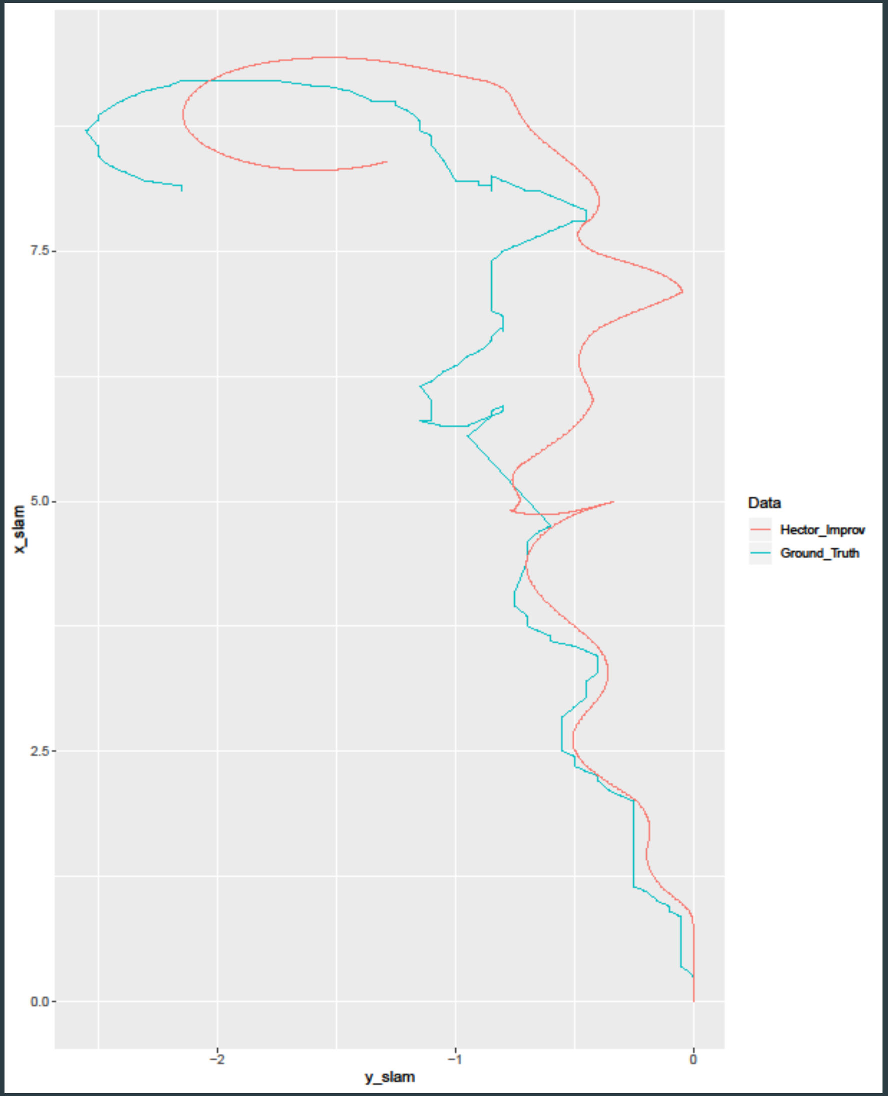
Trajectory Plotting

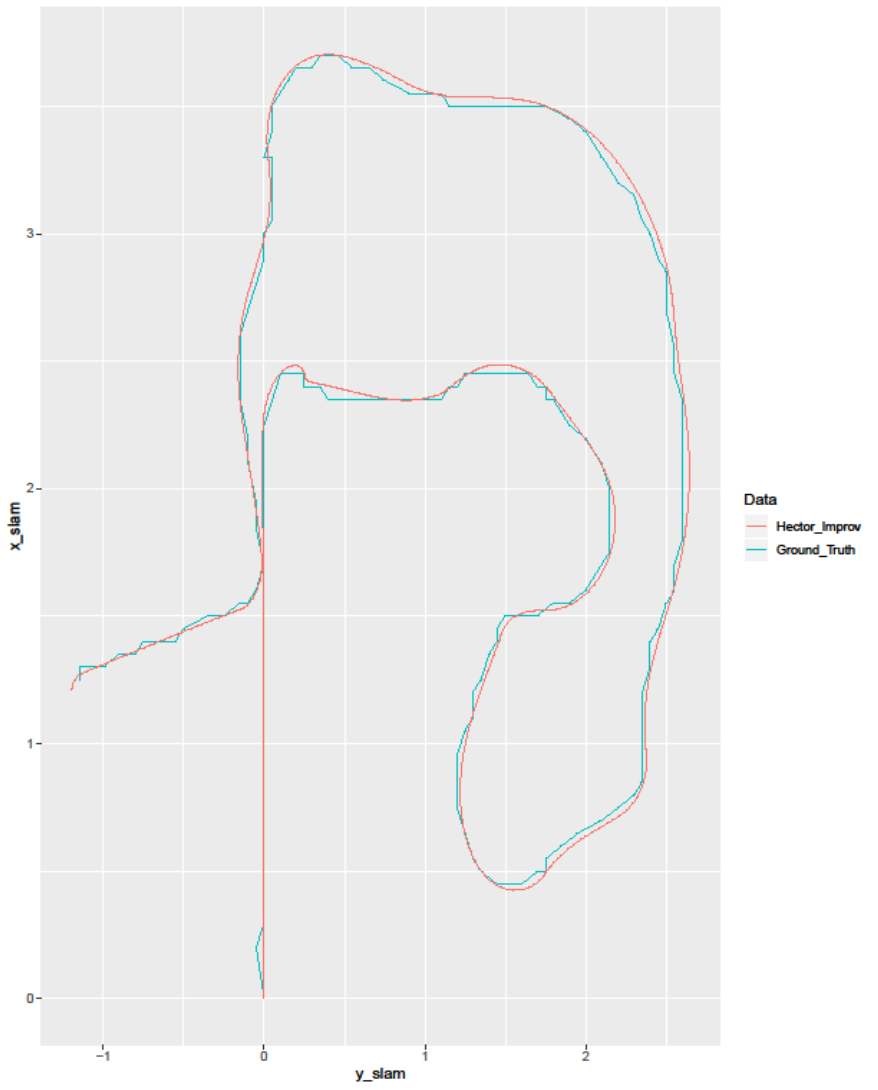
- Good representation but too slim





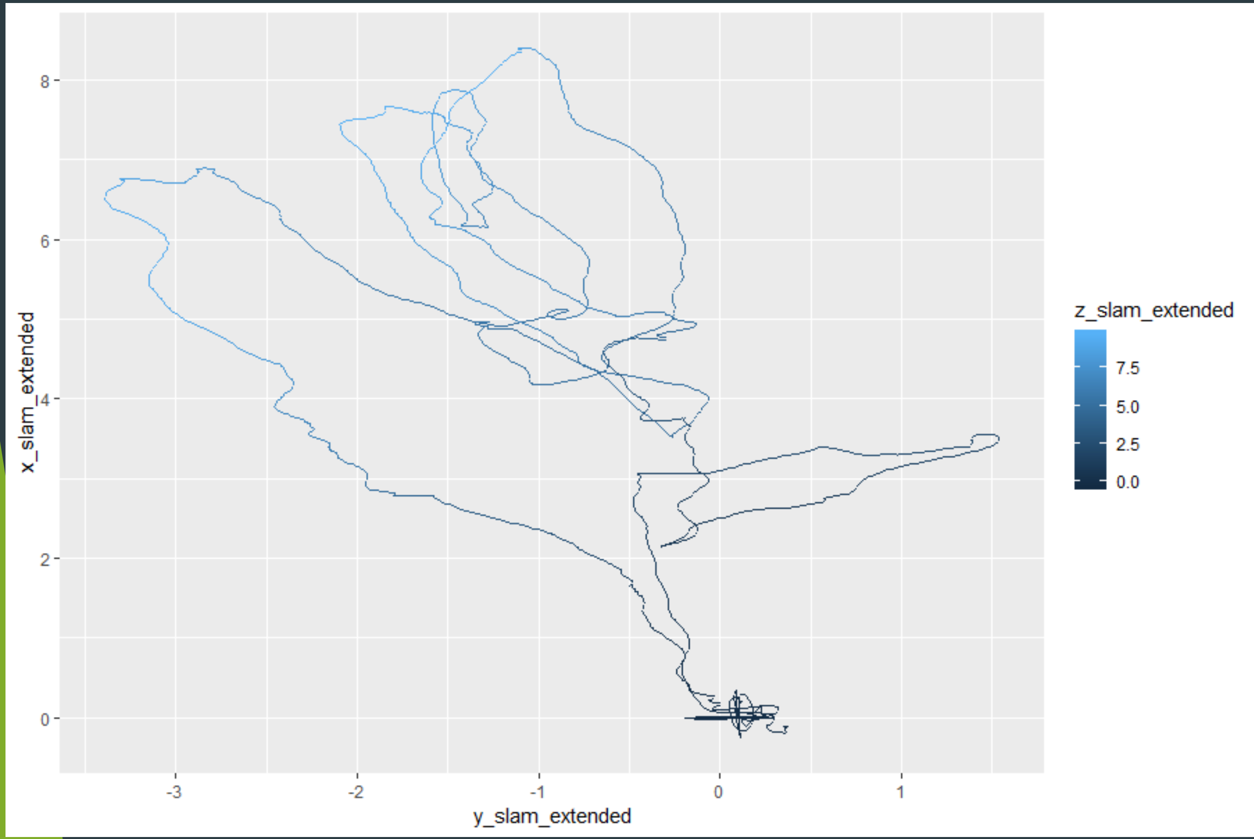
Trajectory Plotting

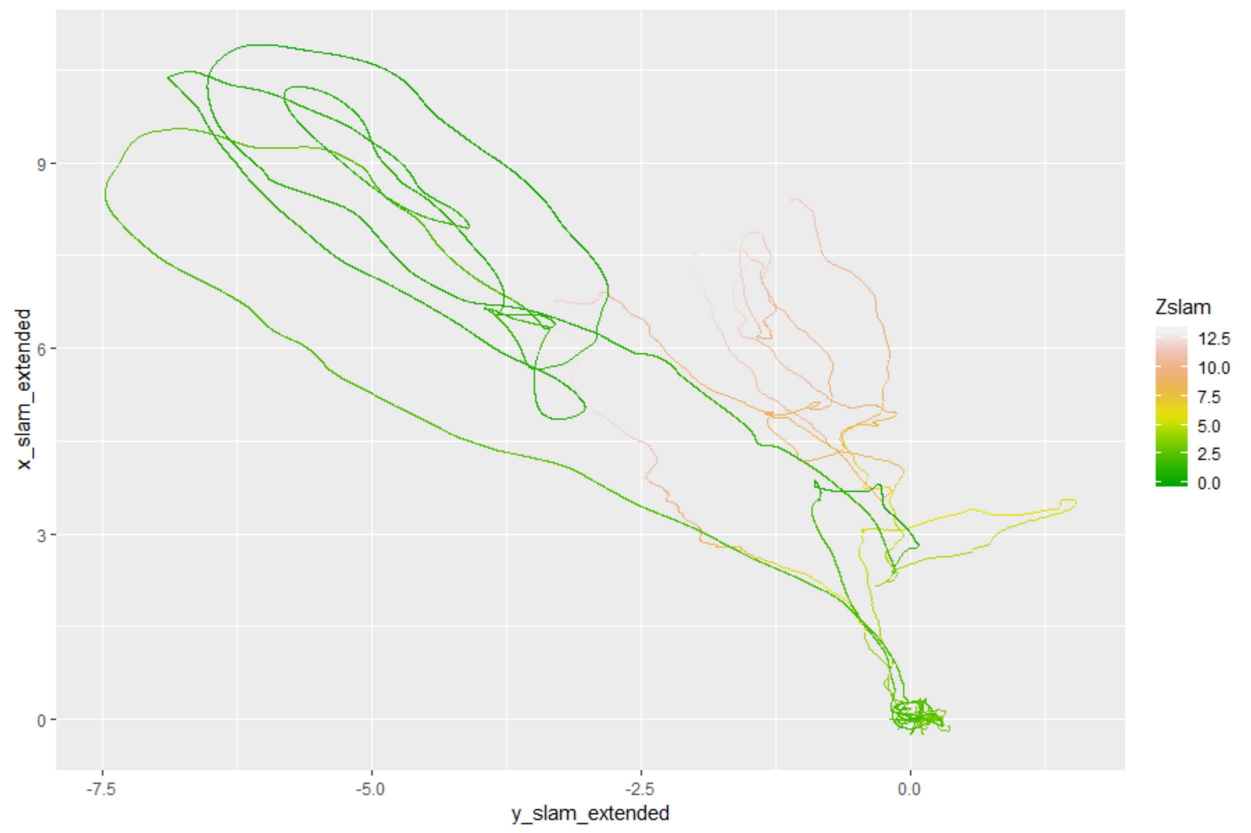




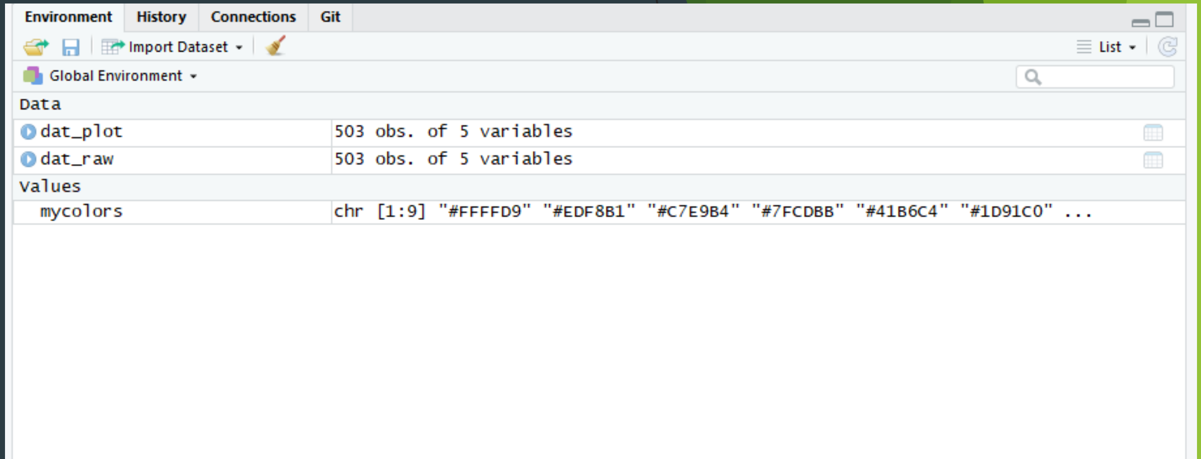
Trajectory Plotting

- Demonstrate flight altitude for micro aerial vehicle (MAV) SLAM



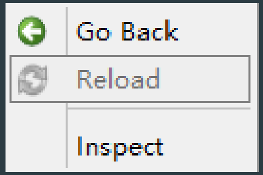


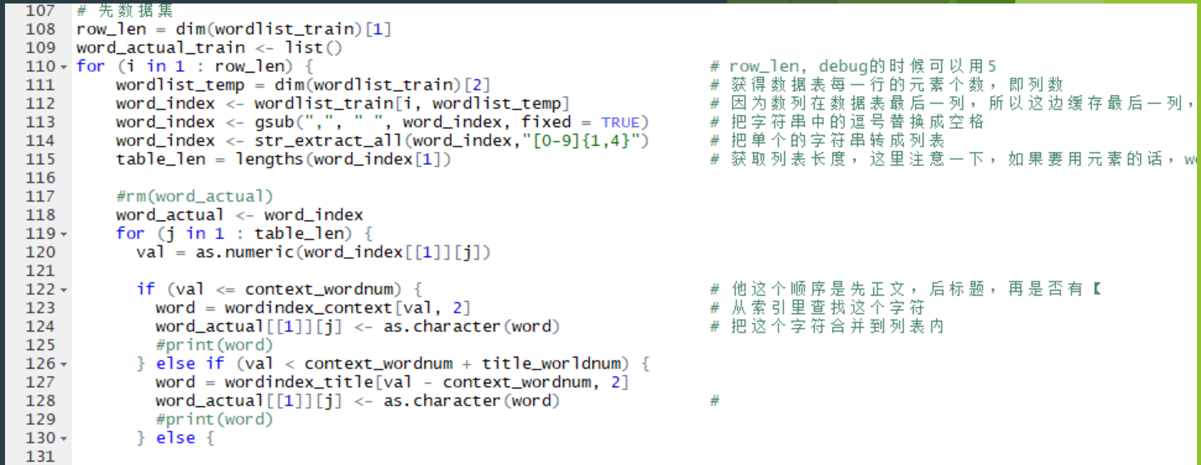
Feeling & Deep Thinking



```
Data
dat_plot      503 obs. of 5 variables
dat_raw       503 obs. of 5 variables
Values
mycolors      chr [1:9] "#FFFFD9" "#EDF8B1" "#C7E9B4" "#7FCDBB" "#41B6C4" "#1D91C0" ...
```

- Advantages
  - Beautiful plotting
  - Good statistical functions
  - "Strange" interpreter, still running after encountering errors
- Disadvantages
  - Lack of string manipulation methods, especially EXTRACTION method
  - Strange RStudio
  - Unclear variable type
  - Not as flexible as C/C++/Python/Java

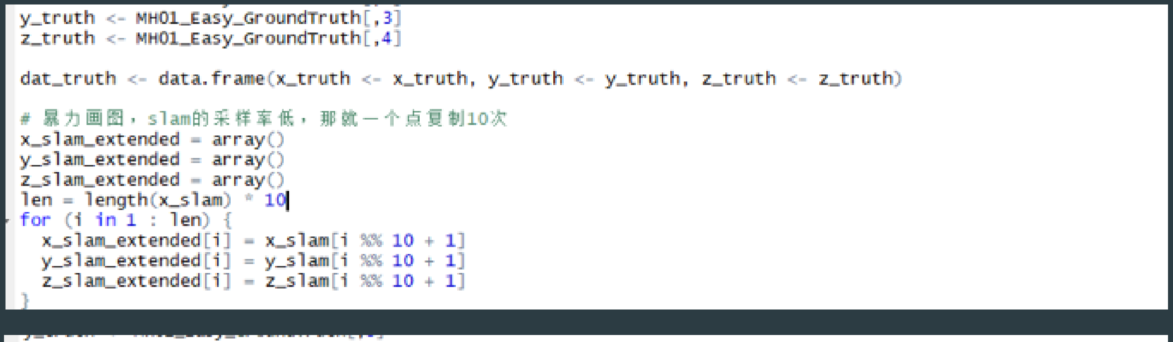




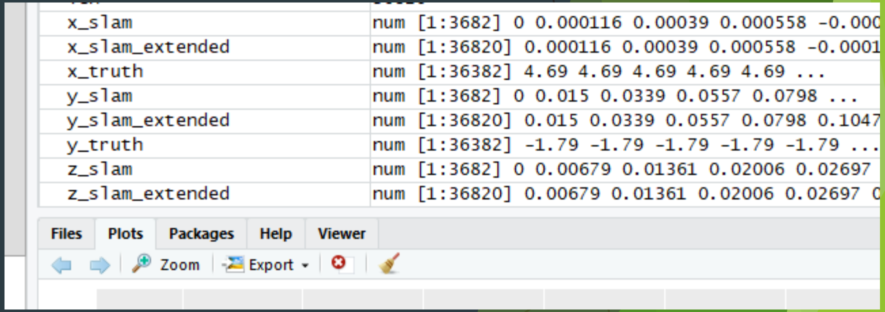
```r
# 先数据集
row_len = dim(wordlist_train)[1]
word_actual_train <- list()
for (i in 1 : row_len) {                                        # row_len, debug的时候可以用5
    wordlist_temp = dim(wordlist_train)[2]                      # 获得数据表每一行的元素个数，即列数
    word_index <- wordlist_train[i, wordlist_temp]              # 因为数列在数据表最后一列，所以这边缓存最后一列，
    word_index <- gsub(",", " ", word_index, fixed = TRUE)      # 把字符串中的逗号替换成空格
    word_index <- str_extract_all(word_index,"[0-9]{1,4}")      # 把单个的字符串转成列表
    table_len = lengths(word_index[1])                          # 获取列表长度，这里注意一下，如果要用元素的话，w

    #rm(word_actual)
    word_actual <- word_index
    for (j in 1 : table_len) {
      val = as.numeric(word_index[[1]][j])

      if (val <= context_wordnum) {                             # 他这个顺序是先正文，后标题，再是否有【
        word = wordindex_context[val, 2]                        # 从索引里查找这个字符
        word_actual[[1]][j] <- as.character(word)               # 把这个字符合并到列表内
        #print(word)
      } else if (val < context_wordnum + title_worldnum) {
        word = wordindex_title[val - context_wordnum, 2]
        word_actual[[1]][j] <- as.character(word)               #
        #print(word)
      } else {

```

Feeling & Deep Thinking

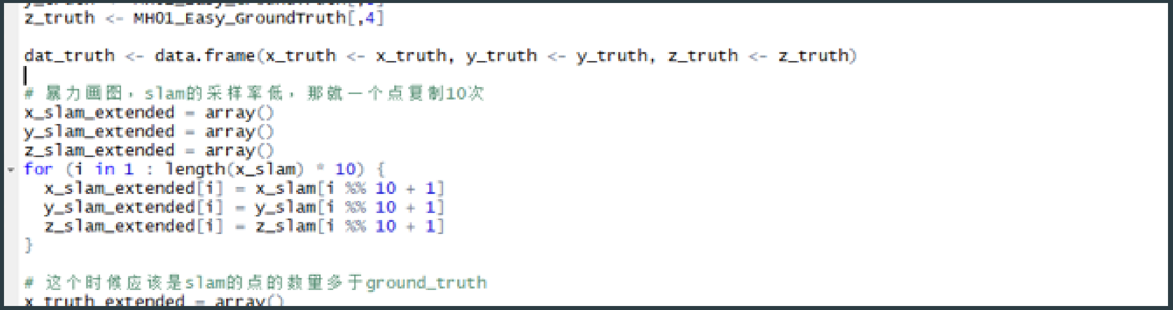
- Feelings
  - In R, data is much more important than the project itself.
  - Never try saving the unfinished data before.
  - A good tool for after-experiment processing.



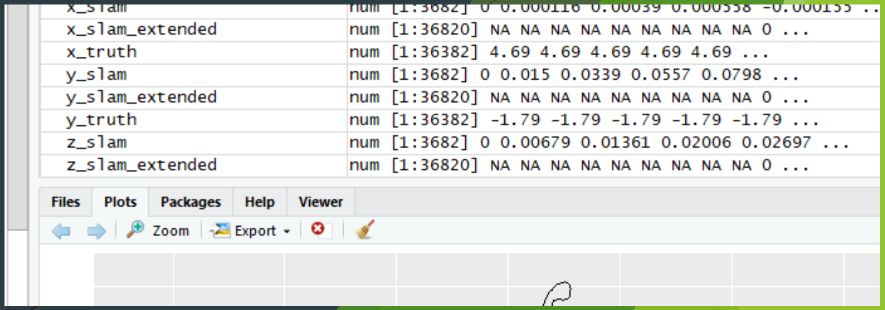
```
y_truth <- MH01_Easy_GroundTruth[,3]
z_truth <- MH01_Easy_GroundTruth[,4]

dat_truth <- data.frame(x_truth <- x_truth, y_truth <- y_truth, z_truth <- z_truth)

# 暴力画图，slam的采样率低，那就一个点复制10次
x_slam_extended = array()
y_slam_extended = array()
z_slam_extended = array()
len = length(x_slam) * 10
for (i in 1 : len) {
  x_slam_extended[i] = x_slam[i %% 10 + 1]
  y_slam_extended[i] = y_slam[i %% 10 + 1]
  z_slam_extended[i] = z_slam[i %% 10 + 1]
}
```



```
x_slam             num [1:3682] 0 0.000116 0.00039 0.000558 -0.000
x_slam_extended    num [1:36820] 0.000116 0.00039 0.000558 -0.0001
x_truth            num [1:36382] 4.69 4.69 4.69 4.69 4.69 ...
y_slam             num [1:3682] 0 0.015 0.0339 0.0557 0.0798 ...
y_slam_extended    num [1:36820] 0.015 0.0339 0.0557 0.0798 0.1047
y_truth            num [1:36382] -1.79 -1.79 -1.79 -1.79 -1.79 ...
z_slam             num [1:3682] 0 0.00679 0.01361 0.02006 0.02697
z_slam_extended    num [1:36820] 0.00679 0.01361 0.02006 0.02697 0
```



```
z_truth <- MH01_Easy_GroundTruth[,4]

dat_truth <- data.frame(x_truth <- x_truth, y_truth <- y_truth, z_truth <- z_truth)

# 暴力画图，slam的采样率低，那就一个点复制10次
x_slam_extended = array()
y_slam_extended = array()
z_slam_extended = array()
for (i in 1 : length(x_slam) * 10) {
  x_slam_extended[i] = x_slam[i %% 10 + 1]
  y_slam_extended[i] = y_slam[i %% 10 + 1]
  z_slam_extended[i] = z_slam[i %% 10 + 1]
}

# 这个时候应该是slam的点的数量多于ground_truth
x_truth_extended = array()
```



```
x_slam             num [1:3682] 0 0.000116 0.00039 0.000558 -0.000155 ...
x_slam_extended    num [1:36820] NA NA NA NA NA NA NA NA NA 0 ...
x_truth            num [1:36382] 4.69 4.69 4.69 4.69 4.69 ...
y_slam             num [1:3682] 0 0.015 0.0339 0.0557 0.0798 ...
y_slam_extended    num [1:36820] NA NA NA NA NA NA NA NA NA 0 ...
y_truth            num [1:36382] -1.79 -1.79 -1.79 -1.79 -1.79 ...
z_slam             num [1:3682] 0 0.00679 0.01361 0.02006 0.02697 ...
z_slam_extended    num [1:36820] NA NA NA NA NA NA NA NA NA 0 ...
```

Thanks for listening!