```
##########################################
### TextMining Lecture 09              #####
### Subject: Text Network Analysis     #####
### Developed by. KKIM                 #####
##########################################

# remove.packages('rlang')
# install.packages('rlang')
library('dplyr')
library('magrittr')
library('stringr')
library('tm')
library('textstem')
library('tidyr')
library('igraph')

# library('textnets')

#### Harry Porter Movies
# ID_number: the unique ID number of each line of dialogue
# scene: the scene number as stated on the DVD/Blu-ray
# character_name: name of the character speaking the line of
dialogue
# dialogue: the dialogue of the character

hp.script <-
  read.csv(file="R file/R file_LEC09/hp_script_ed.csv")
hp.script %>% head(2)
hp.script %>% nrow
hp.script$ID_number %>% unique %>% length
hp.script$scene %>% unique %>% length
hp.script$character_name %>% unique %>% length
hp.script$dialogue %>% unique %>% length

# Based on the frequentist approach,
# who are the top 3 main characters in the Harry Porter movie?
hp.script %>%
  group_by(character_name) %>%
  summarize(count=length(ID_number)) %>%
  arrange(desc(count))

# Based on the frequentist approach,
# find the main characters in each scene.
```

```
hp.script %>%
  group_by(scene, character_name) %>%
  summarize(count=length(ID_number)) %>%
  arrange(desc(count)) %>%
  slice(1)

### Create a matrix
hp.mat <- hp.script %>%
  select(scene, character_name) %>%
  table %>%
  crossprod
diag(hp.mat) <- 0
hp.mat

### Node table
hp.mat.node <- hp.mat %>%
  as.data.frame %>%
  mutate(character_name = rownames(.),
         freq = rowSums(.)) %>%
  select(character_name, freq)
hp.mat.node %>% head

### Edge table
hp.mat.edge <- hp.mat %>%
  as.data.frame %>%
  mutate(from = rownames(.)) %>%
  gather(to, Frequency,
         'Albus Dumbledore':Voldemort) %>%
  mutate(Frequency =
           ifelse(Frequency == 0, NA, Frequency))
hp.mat.edge %>% head

### Create igraph object
library(igraph)
hp.graph <-
  graph_from_data_frame(
    d=hp.mat.edge %>% filter(is.na(Frequency)==FALSE),
    vertices=hp.mat.node,
    directed = FALSE)
hp.graph

library(tidygraph)
hp.tgraph <-
```

```
  tidygraph::as_tbl_graph(hp.graph) %>%
  tidygraph::activate(nodes) %>%
  dplyr::mutate(label=name)
hp.tgraph

### Network indicator
hp.graph %>% degree
hp.graph %>% betweenness
hp.graph %>% closeness

library(igraph)
hp.graph.tbl <-
  data.frame(
    node = V(hp.graph) %>% names,
    deg = hp.graph %>% degree,
    bet = hp.graph %>% betweenness,
    clo = hp.graph %>% closeness
  )
library(ggplot2)
library(ggrepel)
hp.graph.tbl %>%
  mutate(deg=log(deg+1)/max(log(deg+1)),
         bet=log(bet+1)/max(log(bet+1))) %>%
  ggplot(aes(x=deg, y=bet)) +
  geom_vline(xintercept=0.5, linetype='dashed', color='red') +
  geom_hline(yintercept=0.5, linetype='dashed', color='red') +
  geom_point() +
  geom_label_repel(aes(label=node))

### Visualize
# set seed
set.seed(1)
# edge size shows frequency of co-occurrence
library(ggraph)
hp.graph %>%
  ggraph(layout = "auto") +
  geom_edge_arc(colour= "gray50",
                strength = .2, alpha = .2) +
  geom_node_text(aes(label = name),
                 repel = TRUE,
                 colour="gray10") +
  theme_graph(background = "white")
```

```
node.size <-
  V(hp.graph)$freq
node.size
E(hp.graph)$weight <-
  E(hp.graph)$Frequency
E(hp.graph)$weight

hp.graph %>%
  ggraph(layout = "auto") +
  geom_edge_arc(colour= "gray50",
                strength = .1, alpha = .1) +
  geom_node_point(size=log(node.size)*2) +
  geom_node_text(aes(label = name),
                 repel = TRUE,
                 point.padding = unit(0.2, "lines"),
                 size=sqrt(node.size),
                 colour="gray10") +
  scale_edge_width(range = c(0, 2.5)) +
  scale_edge_alpha(range = c(0, .3)) +
  theme_graph(background = "white") +
  guides(edge_width = FALSE,
         edge_alpha = FALSE)

# Visualization in Gephi
head(hp.mat.edge)
head(hp.mat.node)
write.csv(hp.mat.node %>%
            rename(Id=character_name) %>% mutate(Label=Id) %>%
            select(Id, Label, freq),
          file="R file/R file_LEC09/hp.mat.node.csv",
          row.names=FALSE)

write.csv(hp.mat.edge %>% filter(is.na(Frequency)==FALSE) %>%
            rename(Source=from, Target=to) %>%
            rename(Weight=Frequency) %>%
            select(Source, Target, Weight),
          file="R file/R file_LEC09/hp.mat.edge.csv",
          row.names=FALSE)
```