Assignment 2

Smitesh Patil

2023-03-10

```
knitr::opts_chunk$set(echo = TRUE)
```

```
library(igraph)
```

```
## Warning: package 'igraph' was built under R version 4.1.3
```

```
##
## Attaching package: 'igraph'
```

```
## The following objects are masked from 'package:stats':
##
##     decompose, spectrum
```

```
## The following object is masked from 'package:base':
##
##     union
```

```
library(ggraph)
```

```
## Warning: package 'ggraph' was built under R version 4.1.3
```

```
## Loading required package: ggplot2
```

```
library(ggrepel)
```

```
## Warning: package 'ggrepel' was built under R version 4.1.3
```

```
library(kableExtra)
```

```
## Warning: package 'kableExtra' was built under R version 4.1.3
```

```
library(gt)
```

```
## Warning: package 'gt' was built under R version 4.1.3
```

```
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 4.1.3
```

```
##
## Attaching package: 'tidyr'
```

```
## The following object is masked from 'package:igraph':
##
##     crossing
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.1.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:kableExtra':
##
##     group_rows
```

```
## The following objects are masked from 'package:igraph':
##
##     as_data_frame, groups, union
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# reading the graph
g<- read_graph(file="./WordPairs.txt",format="pajek")
g<- as.undirected(g)
g<- simplify(g)

cues <- read.table("./cue.txt", header = F, sep="\t", skip=4)

V(g)$cue<-cues[[1]]

#checking the diameter value
print(diameter(g, weights = NA))
```

```
## [1] 7
```

```r
check_cue_words <- function(target_node_name1, target_node_name2){
# test if the selected words are cue words

  if(V(g)[target_node_name1]$cue & V(g)[target_node_name2]$cue){
    cat("Both target words are cue words \n")
  }else{
    cat("Both target words are NOT cue words \n")
    cat(target_node_name1, "cue = ", as.logical(V(g)[target_node_name1]$cue ),"\n")
    cat(target_node_name2, "cue = ", as.logical(V(g)[target_node_name2]$cue ),"\n")
  }
}
```

```r
random_walk_topic_network <- function(g,target_node_names, steps, walks, mode, topn){

  vertices <- c()

  for (i in 1:2){
    for (j in 1:walks){
     vertices <- c(vertices, list(random_walk(g, target_node_names[i], steps, mode = mode)))
   }
  }

  frequency_target <- head(sort(table(names(unlist(vertices))), decreasing = TRUE), topn)
  unique_words <- names(frequency_target)

  return(unique_words)
}

centralities = function(word_association_network){
  page_rank <- page_rank(word_association_network)$vector
  page_rank <- na.omit(page_rank[!names(page_rank) %in% c(target_word1, target_word2)])
  page_rank <- sort(page_rank, decreasing = TRUE)[1:5]

  betweenness <- betweenness(word_association_network)
  betweenness <- betweenness[!names(betweenness) %in% c(target_word1, target_word2)]
  betweenness <- sort(betweenness, decreasing = TRUE)[1:5]

  eigen_centrality <- eigen_centrality(word_association_network)$vector
  eigen_centrality <- eigen_centrality[!names(eigen_centrality) %in% c(target_word1, target_word2)]
  eigen_centrality <- sort(eigen_centrality, decreasing = TRUE)[1:5]

  return(list(page_rank, betweenness, eigen_centrality))
}
```

```r
target_word1 <- "BOOK"
target_word2 <- "DICTIONARY"

check_cue_words(target_word1, target_word2)
```

```
## Both target words are cue words
```

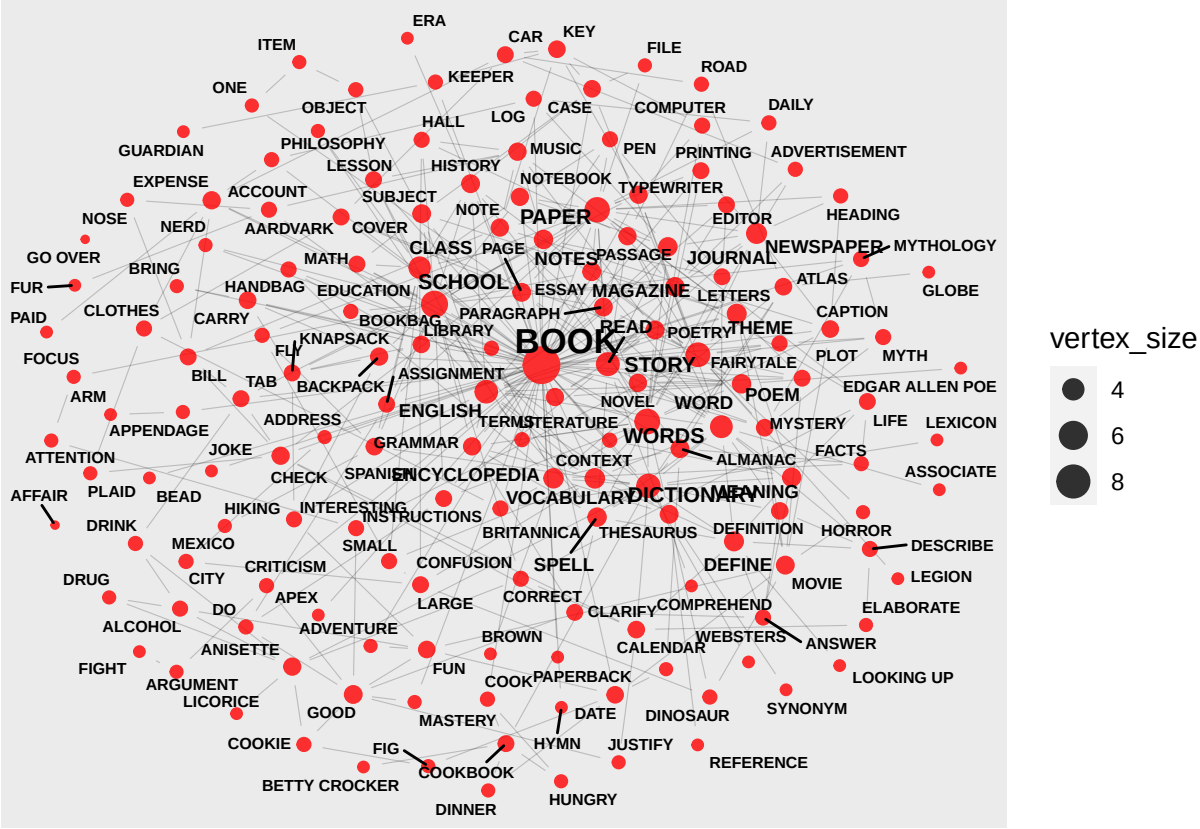
```
out <- random_walk_topic_network(g, c(target_word1, target_word2), 3, 100, "all", 160)

Vertices_in_word_association <- V(g)[name %in% out]

word_association_network1 <- induced.subgraph(g, Vertices_in_word_association)

centrality = centralities(word_association_network1)

df <- tibble(names(centrality[[1]]) , names(centrality[[2]]), names(centrality[[3]]))
colnames(df) <- c("page_rank", "betweeness", "eigen_centrality")

df %>% gt() %>%
  tab_header(paste0("Top 5 words based on centralities based on word association network for
             words ",target_word1 , " and ",target_word2))
```

Top 5 words based on centralities based on word association network for words BOOK and DICTIONARY

page_rank	betweeness	eigen_centrality
WORDS	SCHOOL	READ
PAPER	EXPENSE	LIBRARY
SCHOOL	STORY	NOVEL
STORY	WORDS	LITERATURE
READ	SPANISH	STORY

```
vertex_size <- 2.5 + degree(g)/12
cex_size <-2 + degree(g)/36

#ggraph(word_association_network, layout = "fr")
vertex_size <- 2.5 + degree(word_association_network1)/10
cex_size <-2 + degree(word_association_network1)/30
ggraph(word_association_network1, layout = "fr")+
  geom_edge_link(start_cap = circle(2.5, "mm"),
                 end_cap = circle(2.5, "mm"),
                 edge_width = 0.2,
                 alpha = 0.2)+
  geom_node_point(aes(size = vertex_size),
                  alpha = 0.8,
                  colour = ifelse(V(word_association_network1) %in% c("HEART", "HEAD"), "yellow","red
  geom_node_text(
      aes(label = name),
      fontface = "bold",
      size = cex_size,
      repel = TRUE
    )
```

```
## Warning: Using the ‘size‘ aesthetic in this geom was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ in the ‘default_aes‘ field and elsewhere instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```




```
target_word1 <- "CRIME"
target_word2 <- "TRIAL"

check_cue_words(target_word1, target_word2)
```

```
## Both target words are cue words
```

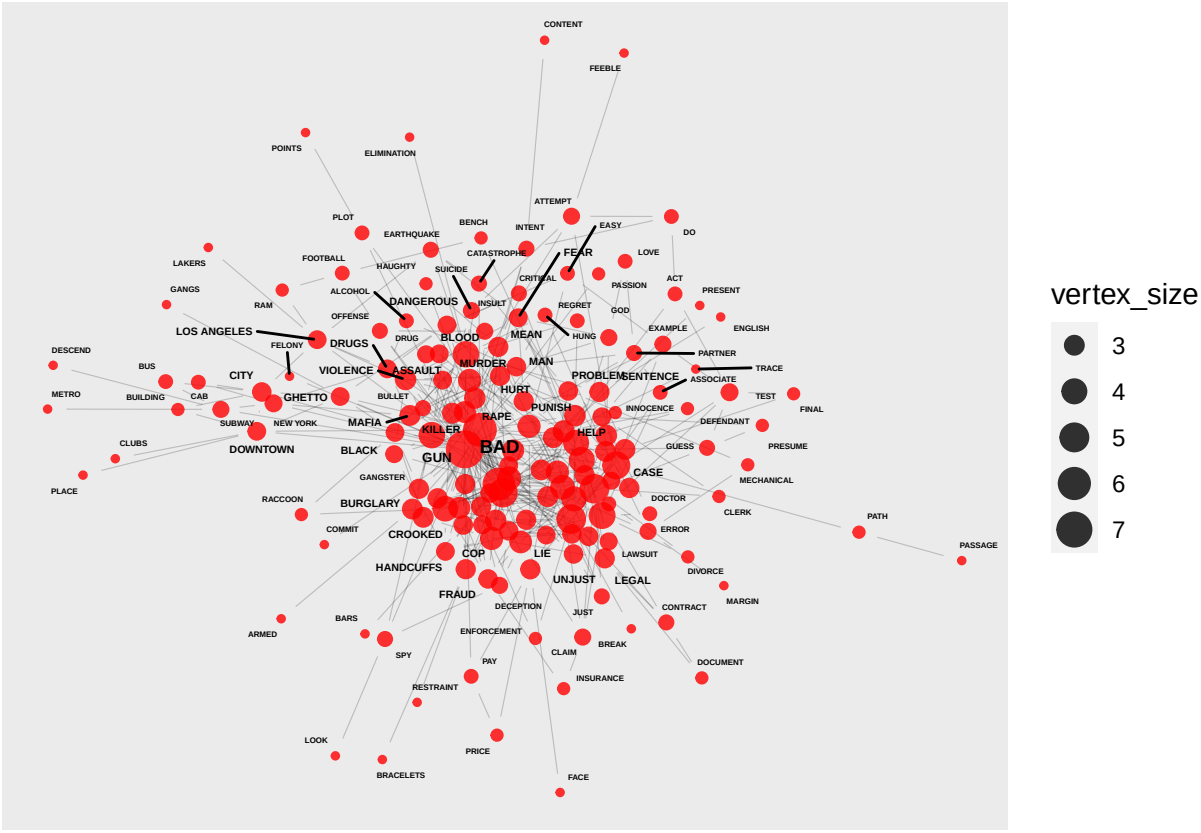
```
out <- random_walk_topic_network(g, c(target_word1, target_word2), 3, 100, "all", 160)

Vertices_in_word_association <- V(g)[name %in% out]

word_association_network2 <- induced.subgraph(g, Vertices_in_word_association)

centrality = centralities(word_association_network2)

df <- tibble(names(centrality[[1]]) , names(centrality[[2]]), names(centrality[[3]]))
colnames(df) <- c("page_rank", "betweeness", "eigen_centrality")

df %>% gt() %>%
  tab_header(paste0("Top 5 words based on centralities based on word association network for
              words ",target_word1 , " and ",target_word2))
```

Top 5 words based on centralities based on word association network for words CRIME and TRIAL

page_rank	betweeness	eigen_centrality
COURT	BAD	COURT
LAWYER	CRIMINAL	JUDGE
CRIMINAL	GUN	LAWYER
BAD	LAW	JURY
CITY	JUDGE	LAW

```
vertex_size <- 2 + degree(word_association_network2)/10
cex_size <- 1 + degree(word_association_network2)/30
ggraph(word_association_network2, layout = "fr")+
  geom_edge_link(start_cap = circle(2.5, "mm"),
                 end_cap = circle(2.5, "mm"),
                 edge_width = 0.2,
                 alpha = 0.2)+
  geom_node_point(aes(size = vertex_size),
                   alpha = 0.8,
                   colour = ifelse(V(word_association_network2) %in% c("HEART", "HEAD"), "yellow","red
  geom_node_text(
      aes(label = name),
      fontface = "bold",
      size = cex_size,
      repel = TRUE
    )
```

```
## Warning: ggrepel: 46 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```




```
target_word1 <- "HEAD"
target_word2 <- "HEART"

check_cue_words(target_word1, target_word2)
```

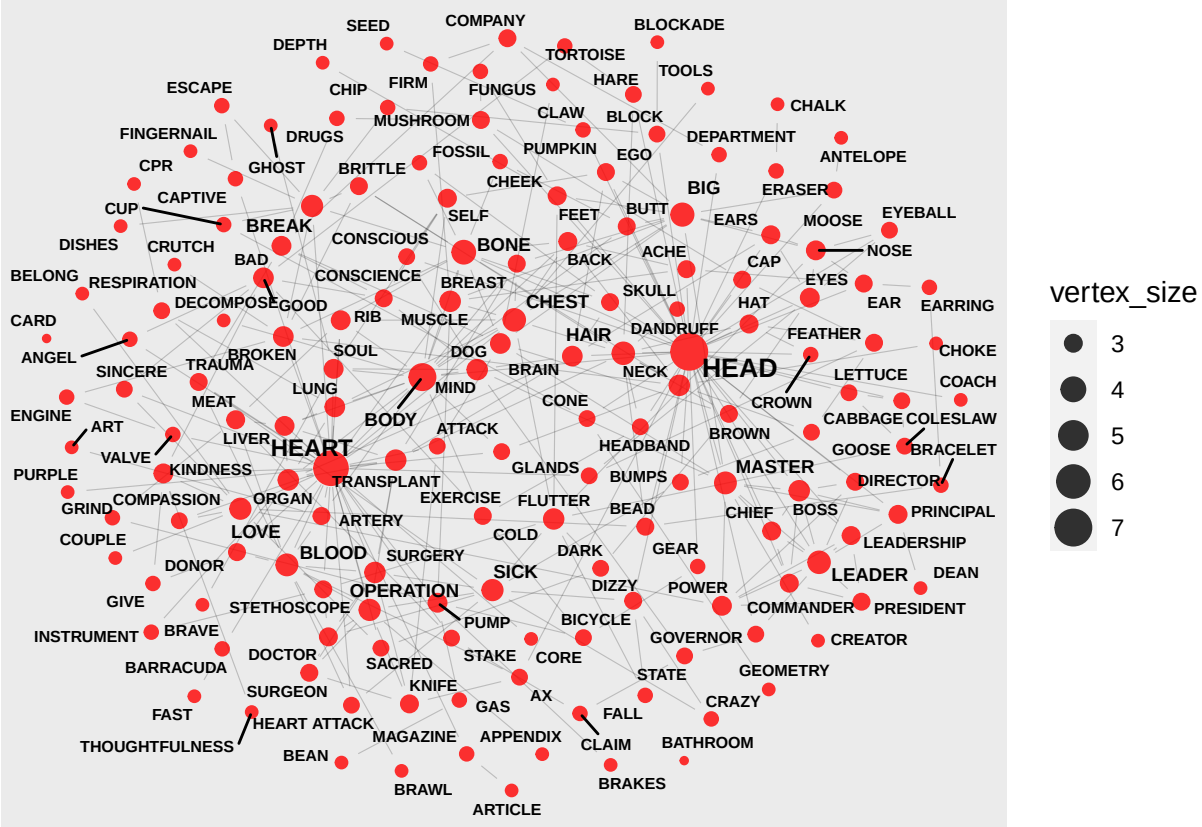
```
## Both target words are cue words
```

```
out <- random_walk_topic_network(g, c(target_word1, target_word2), 3, 100, "all", 160)

Vertices_in_word_association <- V(g)[name %in% out]

word_association_network3 <- induced.subgraph(g, Vertices_in_word_association)

centrality = centralities(word_association_network3)

df <- tibble(names(centrality[[1]]) , names(centrality[[2]]), names(centrality[[3]]))
colnames(df) <- c("page_rank", "betweeness", "eigen_centrality")

df %>% gt() %>%
  tab_header(paste0("Top 5 words based on centralities based on word association network for
             words ",target_word1 , " and ",target_word2))
```

Top 5 words based on centralities based on word association network for words HEAD and HEART

page_rank	betweeness	eigen_centrality
BONE	BIG	GOOD
HAIR	BODY	BAD
HAT	MUSCLE	CONSCIENCE
GOOD	BONE	DRUGS
BAD	MASTER	ANGEL

```
vertex_size <- 2.5 + degree(word_association_network3)/10
cex_size <-2 + degree(word_association_network3)/30
ggraph(word_association_network3, layout = "fr")+
  geom_edge_link(start_cap = circle(2.5, "mm"),
                 end_cap = circle(2.5, "mm"),
                 edge_width = 0.2,
                 alpha = 0.2)+
  geom_node_point(aes(size = vertex_size),
                  alpha = 0.8,
                  colour = ifelse(V(word_association_network3) %in% c("HEART", "HEAD"), "yellow","red
  geom_node_text(
      aes(label = name),
      fontface = "bold",
      size = cex_size,
      repel = TRUE
    )
```




```
#fast greedy community detection
cluster1 <- cluster_louvain(word_association_network1, weights = NA)
#louvain community detection
cluster2 <- cluster_louvain(word_association_network2, weights = NA)
#walktrap community detection
cluster3 <- cluster_louvain(word_association_network3)

create_community_table <- function(clustering_data){
  strings <- c()
  lengths <- c()
  for(i in 1:length(clustering_data)){
    string = ""
    for(word in clustering_data[[i]]){
      string = paste0(string, word)
    }

    strings <- append(strings, string)
    lengths <- append(lengths, length(clustering_data[[i]]))
  }
  df <- tibble(strings, lengths) %>%
    filter(lengths > 1 )
  colnames(df) <- c("Cluster","Size")

  return(df)
}
```

```
community_table1 <- create_community_table(cluster1)
community_table1 %>%
  arrange(desc(Size)) %>%
  kable() %>%
  column_spec(column = 1, width = "5in")
```

Cluster	Size
DOGOODDRUGALCOHOLFIGHTADVENTUREFUNCALENDARARGUMENTDRINKCONFUSIONANISETTELICO	23
ONELOGPAPERNEWSPAPERMAGAZINEADVERTISEMENTLIFEITEMREADDAILYEDITORNOTEPAGEFILECA	21
MUSICCLASSSCHOOLMATHENGLISHERAHALLHISTORYLETTERSOBJECTASSIGNMENTCASENOTEBOOKPE	21
WORDDICTIONARYCORRECTDESCRIBEANSWERGRAMMARVOCABULARYWORDSSPELLASSOCIATECLARI	20
UPELABORATELEXICONWEBSTERSTERMS	
AARDVARKNOSEFURCARACCOUNTCOVERBILLCLOTHESCOMPUTERCHECKBRITANNICATYPEWRITERKE	19
BOOKJOKEARMAPEXAPPENDAGEBACKPACKBOOKBAGCARRYKNAPSACKBREADPAPERBACKBRINGHIKIN	16
PLOTMOVIESTORYNOVELMYSTERYLITERATUREMYTHOLOGYMYTHFAIRYTALEHORRORLEGIONPOETRY	14
ALLEN POE	
LARGEFLYADDRESSCITYSYNONYMSMALLMEXICODINOSAURTHESAURUSREFERENCE	10
ROADALMANACFACTSATLASENCYCLOPEDIAGLOBEBROWNLIBRARYEDUCATION	9
COOKDINNERHUNGRYCOOKBOOKBETTY CROCKER	5

```
community_table2 <- create_community_table(cluster2)
community_table2 %>%
  arrange(desc(Size)) %>%
  kable() %>%
  column_spec(column = 1, width = "5in")
```

Cluster	Size
PRESENTACCUSEGUILTYPROSECUTECOURTLAWYERDEFENDJUDGETRIALCLAIMENGLISHDIVORCECONT	37
STEALBLACKCRIMELIEPLOTMANLOVETRUTHINSURANCESPYLOOKGODPARTNERLIARTHIEFARMEDBUR	33
HURTBADGUILTDANGEROUSDRUGSFEARASSAULTMEANDEATHGUNBLOODREGRETHAUGHTYRAPEVIOL	31
EASYDRUGALCOHOLPROBLEMWRONGTESTPAYPRICEHELPFACETROUBLEGUESSPRESUMEPUNISHMENT	22
CRIMINALARRESTJAILCOPHANDCUFFSOUTLAWBARSCROOKEDRESTRAINTPRISONERCONVICTCORRUPT	18
CITYNEW YORKPLACEBUILDINGBUSRAMFOOTBALLDESCENDBENCH-DOWNTOWNCABLOS ANGELESOFFENSECLUBSLAKERSSUBWAYMETRO	17
DOACTCONTENTINTENT	4

```
community_table3 <- create_community_table(cluster3)
community_table3 %>%
  arrange(desc(Size)) %>%
  kable() %>%
  column_spec(column = 1, width = "5in")
```

Cluster	Size
ORGANARTFASTGIVELOVEBRAVEGHOSTENGINEBLOODHEARTCOREARTERYBARRACUDAPURPLEBELON	32
MINDGEOMETRYACHEHEADSEEDCROWNBUTTDEPARTMENTCONSCIOUSBRAINNECKHATCAPFEATHERSK	24
BACKATTACKDOGBONEBREAKMEATBLOCKBLOCKADECUPCLAWGRINDCAPTIVEBRITTLEDISHESESCAPE	20
BODYMUSCLEEXERCISEDARKCOMPANYFIRMHAIRBROWNHEADBANDBICYCLECHESTBREASTDECOMPOS	17
POWERCLAIMSTATECREATORDIRECTORBOSSPRINCIPALPRESIDENTMASTERCHIEFLEADERSTAKECOMM	16
CRAZYSICKFALLCOLDDOCTOROPERATIONKNIFEAXINSTRUMENTDIZZYSURGERYSURGEONHEART ATTACKAPPENDIXSTETHOSCOPE	15
NOSEBIGSELFEGOBRACELETANTELOPEMOOSEEYESBEADEARRINGEAREARSEYEBALL	13
GOODBADANGELSINCERETRAUMA	5
GASBEANPUMPBRAKES	4
DRUGSFUNGUSMUSHROOM	3
CABBAGELETTUCECOLESLAW	3
MAGAZINEARTICLE	2
CHALKERASER	2
HARETORTOISE	2

```
#loop on the communities

define_community_labels <- function(cluster, dataframe){
  community_label = c()
  for (i in 1:length(cluster)){

    if (length(cluster[[i]]) > 1){
      vertices_in_community <- V(g)[name %in% cluster[[i]]]
      community_graph <- induced.subgraph(g, vertices_in_community)
      page_rank <- page_rank(community_graph)$vector
      label = sort(page_rank, decreasing = TRUE)[1]
      community_label <- c(community_label, label)
    }
  }
  dataframe<- dataframe %>% mutate(community_label = names(community_label))

  return(dataframe)
}

community_with_label<- define_community_labels(cluster1, community_table1)

community_with_label %>%
  arrange(desc(Size)) %>%
  kable() %>%
  column_spec(column = 1, width = "5in")
```

Cluster	Size	community_label
DOGOODDRUGALCOHOLFIGHTADVENTUREFUNCALENDARARGUMENTDRINKCANCHOICE	23	ANCHOI
ONELOGPAPERNEWSPAPERMAGAZINEADVERTISEMENTLIFEITEMREADDAILYEDITOR	21	EDITOR
MUSICCLASSSCHOOLMATHENGLISHERAHALLHISTORYLETTERSOBJECTASSIGNMENT	21	CLASS
WORDDICTIONARYCORRECTDESCRIBEANSWERGRAMMARVOCABULARYWORDSPHRASE UPELABORATELEXICONWEBSTERSTERMS	20	DESCRIBE
AARDVARKNOSEFURCARACCOUNTCOVERBILLCLOTHESCOMPUTERCHECKBRITANNICA	19	BRITANNICA
BOOKJOKEARMAPEXAPPENDAGEBACKPACKBOOKBAGCARRYKNAPSACKHEADPOCKET	16	POCKET
PLOTMOVIESTORYNOVELMYSTERYLITERATUREMYTHOLOGYMYTHFAIRYTALESHORT ALLEN POE	14	SHORT
LARGEFLYADDRESSCITYSYNONYMSMALLMEXICODINOSAURTHESAURUSREFERENCE	10	REFERENCE
ROADALMANACFACTSATLASENCYCLOPEDIAGLOBEBROWNLIBRARYEDUCATION	9	ATLAS
COOKDINNERHUNGRYCOOKBOOKBETTY CROCKER	5	COOKBOOK

```
define_community_labels(cluster2, community_table2) %>%
  arrange(desc(Size)) %>%
  kable() %>%
  column_spec(column = 1, width = "5in")
```

Cluster	Size	community_label
PRESENTACCUSEGUILTYPROSECUTECOURTLAWYERDEFENDJUDGETRIALCLAIMENGLISHDIVORCECONT	37	ENGLISH
STEALBLACKCRIMELIEPLOTMANLOVETRUTHINSURANCESPYLOOKGODPARTNER	31	LIE
HURTBADGUILTDANGEROUSDRUGSFEARASSAULTMEANDEATHGUNBLOODREGRETHEALTH	31	DEATH
EASYDRUGALCOHOLPROBLEMWRONGTESTPAYPRICEHELPFACETROUBLEQUESTIONSUE	22	SUE
CRIMINALARRESTJAILCOPHANDCUFFSOUTLAWBARSCROOKEDRESTRAINTPRISONERCONVICTCORRUPT	18	CONVICT
CITYNEW YORKPLACEBUILDINGBUSRAMFOOTBALLDESCENDBENCH-DOWNTOWNCABLOS ANGELESOFFENSECLUBSLAKERSSUBWAYMETRO	17	CITY
DOACTCONTENTINTENT	4	DO

```
out <-define_community_labels(cluster3, community_table3) %>%
  arrange(desc(Size))


kable(as.data.frame(out), booktabs = TRUE) %>%
  kable_styling(full_width = TRUE)
```

Cluster	Size	community_label
ORGANARTFASTGIVELOVEBRAVEGHOSTENGINEBLOODHEARTCORE	32	HEARTERYBARRACUDAPURPLEBELON
MINDGEOMETRYACHEHEADSEEDCROWNBUTTDEPARTMENTGENES	21	GENESCIOUSBRAINNECKHATCAPFEATHERSK
BACKATTACKDOGBONEBREAKMEATBLOCKBLOCKADECAPCAVEGRINDCAPTIVEBRITTLEDISHESESCAPE	20	
BODYMUSCLEEXERCISEDARKCOMPANYFIRMHAIRBROWNHEARTBRAINBANDBICYCLECHESTBREASTDECOMPOS	17	
POWERCLAIMSTATECREATORDIRECTORBOSSPRINCIPALPRESIDENTMASTERCHIEFLEADERSTAKECOMM	16	
CRAZYSICKFALLCOLDDOCTOROPERATIONKNIFEAXINSTRUMENTDOCTORCRAZYSURGERYSURGEONHEART ATTACKAPPENDIXSTETHO-SCOPE	15	
NOSEBIGSELFEGOBRACELETANTELOPEMOOSEEYESBEARDEARNOSEGEAREARSEYEBALL	13	
GOODBADANGELSINCERETRAUMA	5	GOOD
GASBEANPUMPBRAKES	4	GAS
DRUGSFUNGUSMUSHROOM	3	MUSHROOM
CABBAGELETTUCECOLESLAW	3	CABBAGE
MAGAZINEARTICLE	2	MAGAZINE
CHALKERASER	2	CHALK
HARETORTOISE	2	HARE

```
  #column_spec(column = 1, width = "5in")

dt <- tibble(
   Items = c("Item 1", "Item 2", "Item 3"),
   Text_1 = c("Lorem ipsum dolor sit amet, consectetur adipiscing elit. Proin vehicula tempor ex. Morbi
   Text_2 = c("Duis posuere placerat magna, ac aliquam lorem viverra non. Ut ultrices tempus eros, quis
)

kable(dt, "latex", booktabs = T,
col.names = c("Item", "Short Title", "Very Very Very Very Very Very Long Title")) %>%
column_spec(2:3, width = "5cm")
```

Item	Short Title	Very Very Very Very Very Very Long Title
Item 1	Lorem ipsum dolor sit amet, consectetur adipiscing elit. Proin vehicula tempor ex. Morbi malesuada sagittis turpis, at venenatis nisl luctus a.	Duis posuere placerat magna, ac aliquam lorem viverra non. Ut ultrices tempus eros, quis sodales libero commodo non. In non neque ut lacus vestibulum dictum a quis ipsum.
Item 2	In eu urna at magna luctus rhoncus quis in nisl. Fusce in velit varius, posuere risus et, cursus augue. Duis eleifend aliquam ante, a aliquet ex tincidunt in.	Aenean ut justo interdum, laoreet enim nec, viverra eros. Donec vel pharetra nunc. Suspendisse vel ipsum ac lectus semper aliquam ac a orci. Suspendisse libero mauris, egestas semper auctor sit amet, tempor et orci.
Item 3	Vivamus venenatis egestas eros ut tempus. Vivamus id est nisi. Aliquam molestie erat et sollicitudin venenatis. In ac lacus at velit scelerisque mattis.	Phasellus quis neque aliquet, finibus nunc eget, lacinia neque. Sed auctor lectus vel ex scelerisque commodo.

```
t1w <-out %>% mutate(cl = paste0('"',Cluster,'")) %>% select(cl)

kable( t1w, "latex", booktabs = T) %>%
  column_spec(1, width = "10cm")
```

cl
"ORGANARTFASTGIVELOVEBRAVEGHOSTENGINEBLOODHEARTCOREARTERYBARRACUDAPURPLEBELON
"MINDGEOMETRYACHEHEADSEEDCROWNBUTTDEPARTMENTCONSCIOUSBRAINNECKHATCAPFEATHERS
"BACKATTACKDOGBONEBREAKMEATBLOCKBLOCKADECUPCLAWGRINDCAPTIVEBRITTLEDISHESESCAP
"BODYMUSCLEEXERCISEDARKCOMPANYFIRMHAIRBROWNHEADBANDBICYCLECHESTBREASTDECOMPO
"POWERCLAIMSTATECREATORDIRECTORBOSSPRINCIPALPRESIDENTMASTERCHIEFLEADERSTAKECOMM
"CRAZYSICKFALLCOLDDOCTOROPERATIONKNIFEAXINSTRUMENTDIZZYSURGERYSURGEONHEART ATTACKAPPENDIXSTETHOSCOPE"
"NOSEBIGSELFEGOBRACELETANTELOPEMOOSEEYESBEADEARRINGEAREARSEYEBALL"
"GOODBADANGELSINCERETRAUMA"
"GASBEANPUMPBRAKES"
"DRUGSFUNGUSMUSHROOM"
"CABBAGELETTUCECOLESLAW"
"MAGAZINEARTICLE"
"CHALKERASER"
"HARETORTOISE"