Student Epistemics

Tanner Phillips

1/31/2022

Introduction

I’ve included this code in my portfolio not because it’s the most complicated, but because it shows some very basic principles of coding. Functions, efficient looping, and data manipulation. There’s very little tidyverse here, just base R. It’s easy to write this code badly.

The underlying problem is turning a list of items into a frequency matrix based on complex conditions. The data is student chat from a computer-supported collaborative learning environment (i.e., an educational video game). We were attempting to use processes mining to understand the order of types of speech (e.g., question, assertions of fact, social organization). To do this we wanted to get frequency counts of pairs and triplets of types of speach, like a question, followed by statement of fact, followed by another question.

I’m very proud of (a) The efficient speed at which this code ran when applied 10,000 lines of student chat, and (b) figuring out a simple way to visualize the data that allowed us to make interesting inferences about the data. As of February 2022, we are currently in the process of writing this up as a journal article.

Definitions

The definitions of the different codes may help with understanding the ouput: - **K-.** A question or other query for information. Often in speech our questions are implicit, not explicit.

- **K+.** A direct knowledge claim or assertion.
- **Reply.** A reply to previous comment. Would be meaningless outside of context.
- **Reply - Knowledge.** A reply that includes new knowledge.
- **Reply - Hedge.** A reply that “hedges” what is being said (e.g. I’m not sure, but… etc.)
- **Social Organization.** Attempts to organize from introductions like “hello” to more explicit organizaton like “what should our next step be.
- **Other.** Anything else. Often spam or off-topic

Code

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(RColorBrewer)

#Support constant is used to determine the % Cutoff for patterns in student speech. i.e. if less
that support % of speech patterns fall into this category, we don't carry them forward.
support<-.05
```

Custom Functions

```
#Custom Function to calculate marginal percentages for rows or columns of a matrix
margin_prop<-function(x){
  s<-sum(x)
  x/s
}

# Custom function used to select locations in the data where a certain sequence of 2 or 3 codes
 is present
epis_pattern<-function(vector,pattern1,pattern2,pattern3=NA){
  locations<-c()
  if(is.na(pattern3)){
    for(i in 1:(length(vector)-1)){
      if(vector[i]==pattern1 & vector[i+1]==pattern2){
        locations<-c(locations,i)
      }
    }
  }else{
    for(i in 1:(length(vector)-2)){
      if(vector[i]==pattern1 & vector[i+1]==pattern2 & vector[i+2]==pattern3){
        locations<-c(locations,i)
      }
    }
  }
  return(locations+1)
}
```

Load and Clean

```
  codes<-read.csv("EcoJourney_DiscourseData_All_v2.csv")
  codes<-codes[,1:7]
  names(codes)[1]<-"GroupID"

#Split out the epistemic column. This is just for convenience.
  epis<-codes$Epistemics

  Wizard.indecies<-grepl("w",codes$UserID,ignore.case = T)
  Wizard.indecies.IDS<-which(grepl("w",codes$UserID,ignore.case = T))
```

Analysis 1: Epistemic Pairs

```
###Create epistemic pairs matrix
  freq1<-matrix(0,nrow=length(unique(epis)),ncol=length(unique(epis)))
  wizard_freq1<-matrix(0,nrow=length(unique(epis)),ncol=length(unique(epis)))

###Name rows and columns. We'll use these to select cells in the matrices
  rownames(freq1)<-unique(epis)
  colnames(freq1)<-unique(epis)

  rownames(wizard_freq1)<-unique(epis)
  colnames(wizard_freq1)<-unique(epis)

###Comb through coded text and find couplets of speech types and put frequencies in matrix.
  for(i in 1:(length(epis)-1)){
    freq1[epis[i],epis[i+1]]<-(freq1[epis[i],epis[i+1]]+1)
    if(i %in% Wizard.indecies.IDS){
      wizard_freq1[epis[i],epis[i+1]]<-(wizard_freq1[epis[i],epis[i+1]]+1)
    }
  }

###Transform into frequencies.
  freq1_prob<-t(apply(freq1,MARGIN=1,margin_prop))
  rownames(freq1_prob)<-names(freq1_prob)
  matrix_rows = sum(freq1_prob > support)
  support1<-matrix(0,nrow = matrix_rows,ncol=2)

###Select all supported discourse pairs to carry forward into triplet analysis.
  k = 1
  for(i in 1:nrow(freq1_prob)){
    for(j in 1:ncol(freq1_prob)){
      if(freq1_prob[i,j]>support){
        support1[k,]<-c(rownames(freq1)[i],colnames(freq1)[j])
        k = k + 1
      }
    }
  }

freq1
```

```
##                     Other Social organization  K-  K+ Reply Reply-Hedge
## Other                 1103                  90 146 151   213          23
## Social organization     60                 156 104  92   144          24
## K-                     133                  47 254 147   581         131
## K+                     127                  65 208 310   383          39
## Reply                  259                 184 534 393  1440         123
## Reply-Hedge             29                  18  91  35   158          81
## Reply-Knowledge         47                  38 167  52   156          52
##                     Reply-Knowledge
## Other                            32
## Social organization               18
## K-                               211
## K+                                48
## Reply                            142
## Reply-Hedge                       61
## Reply-Knowledge                  214
```

Analysis 2: Epistemic Triplets

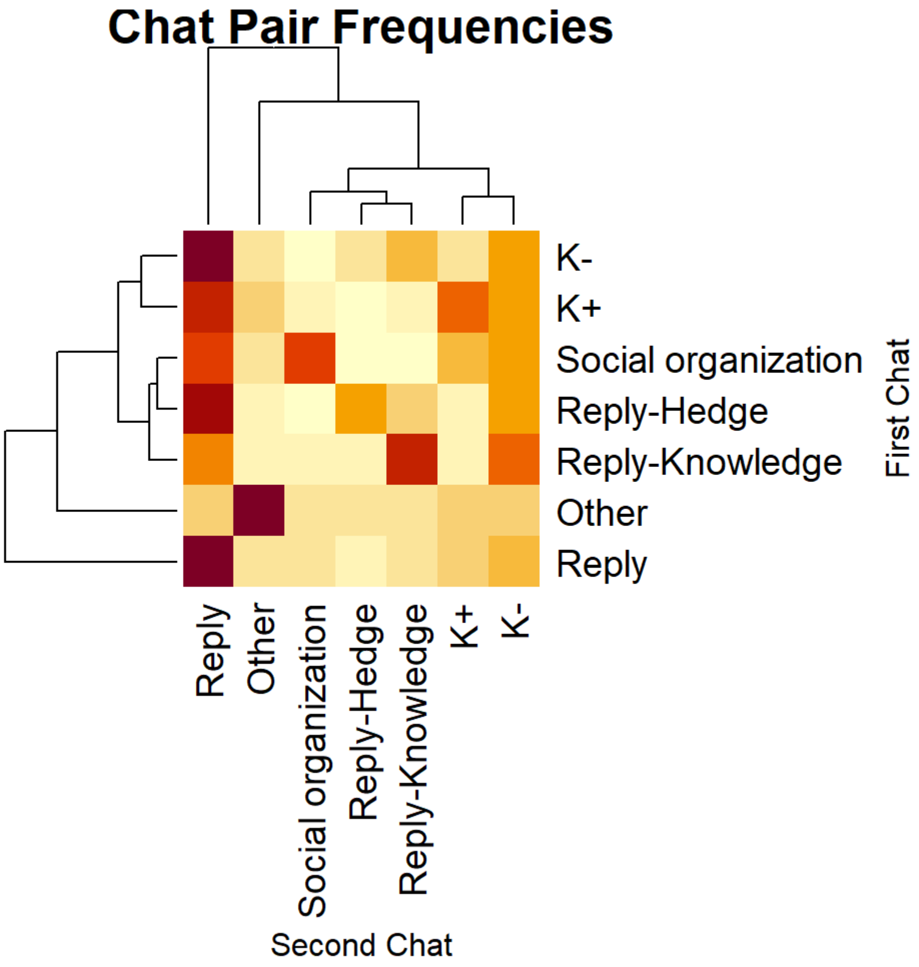
```
###Initiate matrices
  freq2<-matrix(0,nrow=nrow(support1),ncol=nrow(freq1))
  wizard_freq2<-matrix(0,nrow=nrow(support1),ncol=nrow(freq1))

###name rows and columns
  rownames(freq2)<-paste(support1[,1],support1[,2],sep = "->")
  colnames(freq2)<-colnames(freq1)

  rownames(wizard_freq2)<-paste(support1[,1],support1[,2],sep = "->")
  colnames(wizard_freq2)<-colnames(freq1)

###Grab Epistemic Triplets. Essentialy same as for couplets.
  for(i in 1:(length(epis)-3)){
    pattern = paste(epis[i],epis[i+1],sep = "->")
    if(pattern %in% rownames(freq2)){
      freq2[pattern,epis[i+2]]<-freq2[pattern,epis[i+2]]+1
      if(i %in% Wizard.indecies.IDS){
        wizard_freq2[pattern,epis[i+2]]<- wizard_freq2[pattern,epis[i+2]]+1
      }
    }
  }

###Save percentage frequencies by row
  freq2_prob<-as.data.frame(t(apply(freq2,MARGIN=1,margin_prop)))
  colnames(freq2_prob)<-colnames(freq2)
head(freq2)
```

```
##                            Other Social organization K- K+ Reply Reply-Hedge
## Other->Other                 828                  47 61 83    69           8
## Other->Social organization    18                  30 10 15    15           2
## Other->K-                     37                   3 27 13    43           6
## Other->K+                     46                   8 23 36    32           3
## Other->Reply                  51                   5 30 22    90           7
## Social organization->Other    31                   8  9  6     5           0
##                            Reply-Knowledge
## Other->Other                             5
## Other->Social organization               0
## Other->K-                               17
## Other->K+                                3
## Other->Reply                             8
## Social organization->Other               1
```

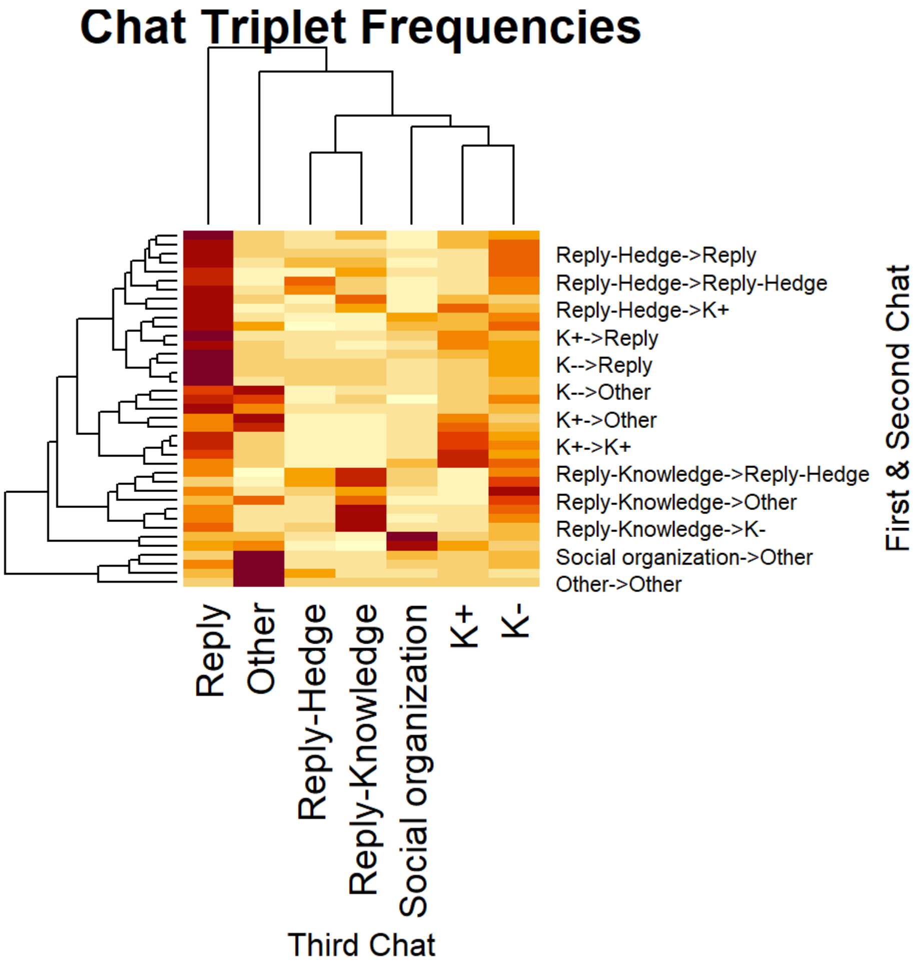
Results

Unpacking these results is the topic of the paper we are currently writing and would make this document a bit unruly. See the “Student Discourse Results” powerpoint if you’re interest in a quick overview.

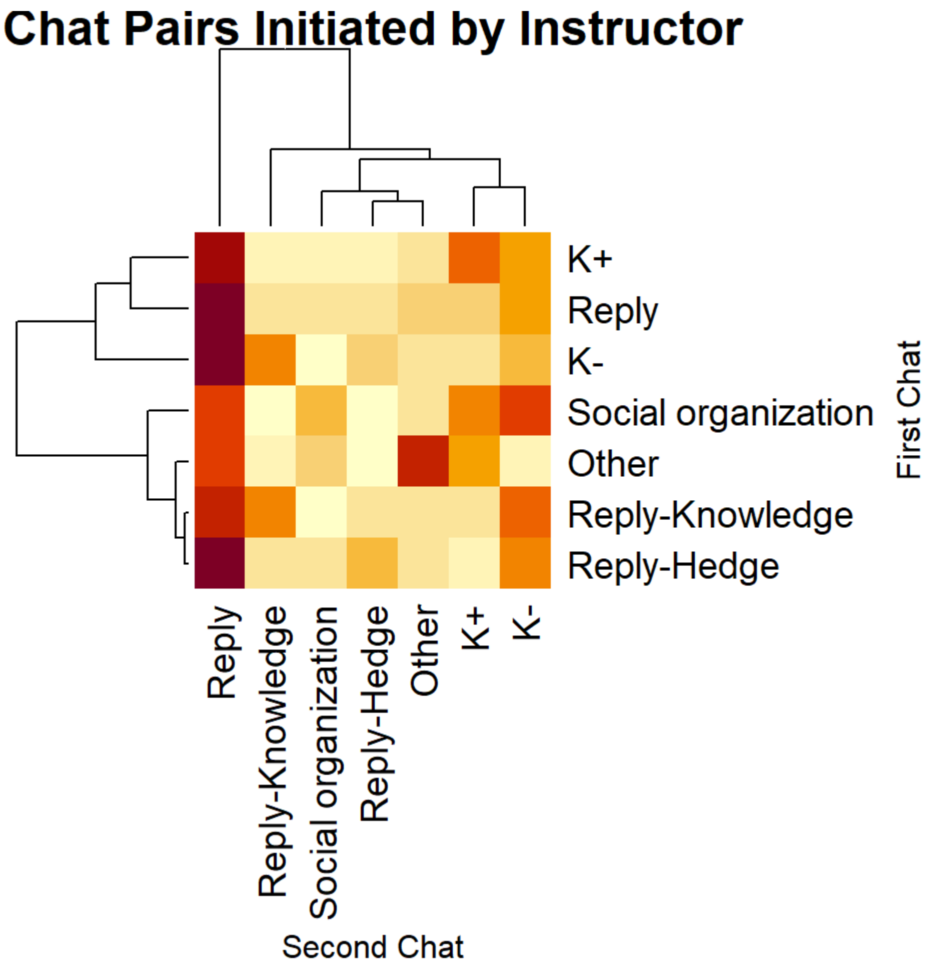
```
###Full Results
  heatmap(freq1,
          main="Chat Pair Frequencies",
          margins = c(12,12),
          ylab="First Chat",
          xlab="Second Chat")
```



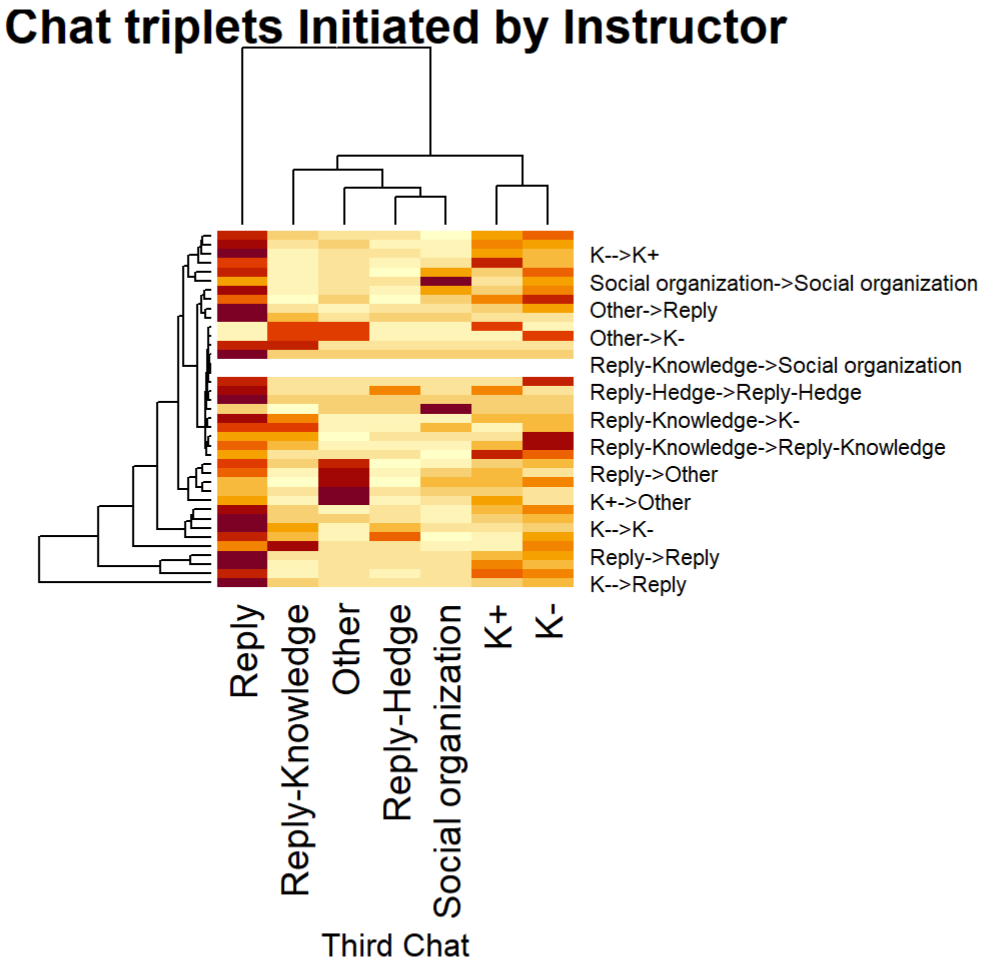
```
heatmap(as.matrix(freq2_prob),
        margins=c(12,12),
        main="Chat Triplet Frequencies",
        ylab="First & Second Chat",
        xlab="Third Chat")
```



```
###Wizard Results
  heatmap(as.matrix(wizard_freq1),
          main="Chat Pairs Initiated by Instructor",
          margins=c(12,12),
          ylab="First Chat",
          xlab="Second Chat")
```



```
heatmap(as.matrix(wizard_freq2),
        main="Chat triplets Initiated by Instructor",
        xlab="Third Chat",
        margins=c(12,12))
```



```
####Select Lines for qualitative review
  lines<-epis_pattern(epis,"Reply-Knowledge","K-","K-")
  lines
```

```
##  [1]  247 1522 1993 2309 2412 4980 5226 5663 5677 5685 5732 6764 7496 7515 7528
## [16] 7694 8389 8846 8921
```