**Statistical Programming: IO Portfolio**

**Introduction**

This report examines if the length of the description for each professor on the Rose Hulman website, varies depending on the professor ranking. There are 3 types of professor rankings: assistant professor, associate professor and professor, with increasing ranking respectively. It also gives contact information about alumni working in the academic department.

**Methods**

**Professor Paragraph Lengths**

Professor titles were sorted based on if they contained certain types of RegEx phrases as shown below:

"[[:alnum:]]\*Assistant Professor[[:alnum:]]\*"

"[[:alnum:]]\*Associate Professor[[:alnum:]]\*"

"[[:alnum:]]\*Professor[[:alnum:]]\*"

**Code chunk 1**

Filename information (to obtain the html file of the professor page), was obtained after sorting the titles. The filenames were put into the corresponding vectors for the corresponding type or professor.

After that the paragraphs were obtained from the filenames and put into the corresponding vectors for the different types of professors. Below is the code for associate professors. Similar loops were run for assistant professors and professors.

for(i in 1:length(associate)){

url <- paste(associate[i],".html", sep="")

if(file.exists(url)){

prof.pg <- read\_html(url)

para.df <- html\_nodes(prof.pg, "div.region.content-container-region")

...alumni code section...

ans<-xml\_text(para.df[2], trim=TRUE)

associate.para <- c(associate.para,ans)

}

}

**Code chunk 2**

After that, the paragraph lengths were found for all the paragraphs and stored into the corresponding vector for the specific type of professor as shown below. Similar code was run for assistant professors and professors.

assistant.length <- sapply(strsplit(assistant.para, " "), length)

**Code chunk 3**

Then the 3 vectors for the different professor types were combined into one dataframe and the dataframe was tidied as shown below:

proflen.df <- gather(df2, key=Type, value=Length, c(associate.length,assistant.length,prof.length))

**Code chunk 4**

After that an overlapping density plot was plotted for the three different types of professors.

**Alumni Information**

To obtain alumni information, checks were done in code chunk 2, in the alumni code section, to find if the professors were Rose Hulman alumni. If they were alumni then they were added to the alumni vector. The code for this is below:

alumniInfo <- html\_nodes(prof.pg, "div.accordion-item")

checkvector<-xml\_text(alumniInfo, trim=TRUE)

for(j in 1:length(checkvector)){

if(str\_detect(checkvector[j],"[[:alnum:]]\*Academic Degrees[[:alnum:]]\*")){

if(str\_detect(checkvector[j],"[[:alnum:]]\*Rose-Hulman Institute of Technology[[:alnum:]]\*")){

alumni<-c(associate[i],alumni)

}

...

}

**Code chunk 5 (na and null checks have been excluded)**

After this, all the alumni information was obtained by finding the file for each alumni and getting information from that file. The code chunk for this is below:

alumniInfo<-vector()

for(i in 1:length(alumni)){

thisfilename<-alumni[i]

for(j in 1:length(filename)){

...na and null checks...

if(filename[j]==thisfilename){

alumnus<-users[j,]

info<-paste(alumnus$name,

alumnus$department,

alumnus$email,

alumnus$phone, "\n", sep="\n")

alumniInfo<-c(alumniInfo,info)

}

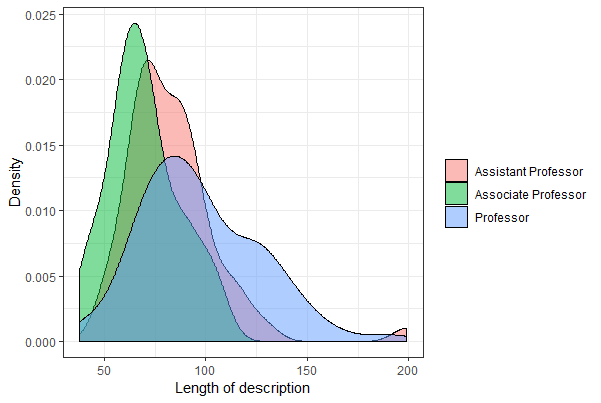
...

}

**Code chunk 6**

**Results**

On the next page, is the density plot for the three different types of professors.



**Figure 1: Density plot of length of description for different types of professors**

In general most of the associate professor description lengths are the lowest (peaking at about 65), while the assistant professor description lengths are higher (peaking at about 70) and professor description lengths are the highest (peaking at 80). Also, most of the graph area region is clearly lowest for associate professors, a little higher for assistant professors, and highest for professors. To conclude it does seem that professors generally get more words in their description due to their ranking but the assistant professors get more words in their description despite their lower ranking compared to associate professors.

Below are the results obtained for alumni contact information:

Thom Adams

Mechanical Engineering

adams1@rose-hulman.edu

872-6089

Zac Chambers

Mechanical Engineering

chambez@rose-hulman.edu

877-8904

Rick Ditteon

Physics & Optical Engineering

ditteon@rose-hulman.edu

877-8247

Rick Stamper

Mechanical Engineering

stamper1@rose-hulman.edu

877-8956

Ashley Bernal

Mechanical Engineering

bernala@rose-hulman.edu

877-8623

Dave Fisher

Computer Science & Software Engineering

fisherds@rose-hulman.edu

877-8619

Kimberly Henthorn

Chemical Engineering

henthorn@rose-hulman.edu

877-8812

Matt Lovell

IRPA

lovellmd@rose-hulman.edu

877-8318

Calvin Lui

Mechanical Engineering

luic@rose-hulman.edu

872-6998

Eric Reyes

Mathematics

reyesem@rose-hulman.edu

877-8287

Sid Stamm

Computer Science & Software Engineering

stammsl@rose-hulman.edu

877-8364

Emma Dosmar

Biology & Biomedical Engineering

dosmare@rose-hulman.edu

877-8821

Megan Heyman

Mathematics

heyman@rose-hulman.edu

877-8557

Greg Neumann

Chemical Engineering

neumann@rose-hulman.edu

877-8026

Amanda Stouder

Computer Science & Software Engineering

stouder@rose-hulman.edu

877-8138

Micah Taylor

Computer Science & Software Engineering

taylormt@rose-hulman.edu

877-8396

**Appendix**

**# Load Necessary Packages**

**pkgs <- c("dplyr", "ggplot2", "httr", "jsonlite", "readr", "rvest",**

**"stringr", "tidyr")**

**for(pkg in pkgs) library(pkg, character.only=TRUE)**

**library("stringr")**

**library("XML")**

**setwd("C:/Users/bharila/Documents/StatisticalProgrammingR/Portfolios")**

**#read faculty directory json**

**jsonfile<-fromJSON(file("facultyDirectory.json"))**

**users<-jsonfile$users[[1]]**

**title<-users$title**

**filename<-users$filename**

**#put filenames in 3 corresponsing vectors**

**#for differnt professor types**

**#get alumnus information**

**associate<-vector()**

**assistant<-vector()**

**prof<-vector()**

**#alumni<-vector()**

**for(i in 1:length(title)){**

**thistitle<-title[i]**

**if(is.na(thistitle)){**

**#do nothing**

**}**

**else if(str\_detect(thistitle,"[[:alnum:]]\*Assistant Professor[[:alnum:]]\*")){**

**associate<-c(filename[i], associate)**

**}**

**else if(str\_detect(thistitle,"[[:alnum:]]\*Associate Professor[[:alnum:]]\*")){**

**assistant<-c(filename[i], assistant)**

**}**

**else if(str\_detect(thistitle,"[[:alnum:]]\*Professor[[:alnum:]]\*")){**

**prof<-c(filename[i], prof)**

**}**

**else{**

**#not a teacher or alum**

**}**

**}**

**#gets paragraphs for each teacher**

**#puts them in 3 different vectors**

**#corresponding to the type of teacher**

**setwd("C:/Users/bharila/Documents/StatisticalProgrammingR/Portfolios/Faculty Portfolios")**

**associate.para<-vector()**

**assistant.para<-vector()**

**prof.para<-vector()**

**alumni<-vector()**

**for(i in 1:length(associate)){**

**url <- paste(associate[i],".html", sep="")**

**if(file.exists(url)){**

**prof.pg <- read\_html(url)**

**para.df <- html\_nodes(prof.pg, "div.region.content-container-region")**

**alumniInfo <- html\_nodes(prof.pg, "div.accordion-item")**

**checkvector<-xml\_text(alumniInfo, trim=TRUE)**

**if(length(checkvector)!=0){**

**for(j in 1:length(checkvector)){**

**if(!is.null(checkvector[j])&!is.na(checkvector[j])){**

**if(str\_detect(checkvector[j],"[[:alnum:]]\*Academic Degrees[[:alnum:]]\*")){**

**if(str\_detect(checkvector[j],"[[:alnum:]]\*Rose-Hulman Institute of Technology[[:alnum:]]\*")){**

**alumni<-c(associate[i],alumni)**

**}**

**}**

**}**

**}**

**}**

**ans<-xml\_text(para.df[2], trim=TRUE)**

**associate.para <- c(associate.para,ans)**

**}**

**}**

**for(i in 1:length(assistant)){**

**url <- paste(assistant[i],".html", sep="")**

**if(file.exists(url)){**

**prof.pg <- read\_html(url)**

**para.df <- html\_nodes(prof.pg, "div.region.content-container-region")**

**alumniInfo <- html\_nodes(prof.pg, "div.accordion-item")**

**checkvector<-xml\_text(alumniInfo, trim=TRUE)**

**if(length(checkvector)!=0){**

**for(j in 1:length(checkvector)){**

**if(str\_detect(checkvector[j],"[[:alnum:]]\*Academic Degrees[[:alnum:]]\*")){**

**if(str\_detect(checkvector[j],"[[:alnum:]]\*Rose-Hulman Institute of Technology[[:alnum:]]\*")){**

**alumni<-c(assistant[i],alumni)**

**}**

**}**

**}**

**}**

**ans<-xml\_text(para.df[2], trim=TRUE)**

**assistant.para <- c(assistant.para,ans)**

**}**

**}**

**for(i in 1:length(prof)){**

**url <- paste(prof[i],".html", sep="")**

**if(file.exists(url)){**

**prof.pg <- read\_html(url)**

**para.df <- html\_nodes(prof.pg, "div.region.content-container-region")**

**alumniInfo <- html\_nodes(prof.pg, "div.accordion-item")**

**checkvector<-xml\_text(alumniInfo, trim=TRUE)**

**if(length(checkvector)!=0){**

**for(j in 1:length(checkvector)){**

**if(!is.null(checkvector[j])&!is.na(checkvector[j])){**

**if(str\_detect(checkvector[j],"[[:alnum:]]\*Academic Degrees[[:alnum:]]\*")){**

**if(str\_detect(checkvector[j],"[[:alnum:]]\*Rose-Hulman Institute of Technology[[:alnum:]]\*")){**

**alumni<-c(prof[i],alumni)**

**}**

**}**

**}**

**}**

**}**

**ans<-xml\_text(para.df[2], trim=TRUE)**

**prof.para <- c(prof.para,ans)**

**}**

**}**

**#put all alumni contact information in one vector**

**alumniInfo<-vector()**

**for(i in 1:length(alumni)){**

**thisfilename<-alumni[i]**

**for(j in 1:length(filename)){**

**if(!is.null(thisfilename)){**

**if(!is.na(thisfilename)){**

**if(!is.na(filename[j])){**

**if(filename[j]==thisfilename){**

**alumnus<-users[j,]**

**info<-paste(alumnus$name,**

**alumnus$department,**

**alumnus$email,**

**alumnus$phone, "\n", sep="\n")**

**alumniInfo<-c(alumniInfo,info)**

**}**

**}**

**}**

**}**

**}**

**}**

**#get paragraph lengths**

**assistant.length <- sapply(strsplit(assistant.para, " "), length)**

**associate.length <- sapply(strsplit(associate.para, " "), length)**

**prof.length <- sapply(strsplit(prof.para, " "), length)**

**#tidy data dataframe**

**max.len = max(length(associate.length), length(assistant.length), length(prof.length))**

**associate.length = c(associate.length, rep(NA, max.len - length(associate.length)))**

**assistant.length = c(assistant.length, rep(NA, max.len - length(assistant.length)))**

**prof.length = c(prof.length, rep(NA, max.len - length(prof.length)))**

**df2 <- data.frame(associate.length,assistant.length,prof.length)**

**proflen.df <- gather(df2,**

**key=Type, value=Length,**

**c(associate.length,assistant.length,prof.length))**

**proflen.df$Type<-as.factor(proflen.df$Type)**

**library(plyr)**

**proflen.df$Type<-mapvalues(proflen.df$Type, from = c("associate.length", "assistant.length", "prof.length"),**

**to = c("Associate Professor", "Assistant Professor", "Professor"))**

**proflen.df<-na.omit(proflen.df)**

**#plot graphic**

**ggplot(data=proflen.df, mapping=aes(x=proflen.df$Length, fill=proflen.df$Type)) +**

**geom\_density(alpha=0.5) +**

**labs(x="Length of description", y="Density", fill="") +**

**theme\_bw()**

**#Print alumni info**

**cat(alumniInfo)**