PSYC 6783 Lab Notebook

Huda Akef

Running To Do List:

- test R and LIWC with Arabic text
  - If Arabic text will work
    - test whether files with both Arabic and English text will work –> *this works in R, also works in LIWC*
    - need to figure out what delimiters to use, probably after deciding on analytical approach
    - think about small edits in transcripts that would make it easier to identify pronouns, negations…etc that would normally be attached to a word rather than being a separate word (ماء، ش، ف), also can an R program be written to make these changes?
    - ~~edit transcripts according to previous point~~
    - ~~look into building dictionary through iterations~~
    - ~~look into building Arabic dictionary, potentially using Dedoose code excerpts?~~
    - work on dictionary specific to religion/morality to come up with analysis relevant to completed Dedoose coding
    - ~~identify types of analyses that would make sense for Arabic text, considering the lack of a dictionary~~
    - **remove my questions from transcript text** *ongoing*
  - ~~If Arabic text will not work~~
    - ~~translate morality portions of all interviews~~
    - ~~review codes of interest that may be used for analysis~~
- ~~read more about language analysis conceptually rather than technically, create table to compare methods~~
- identifying religion related words, review possible categories from 2/25/22

Date Friday 4/8/22

started using Git here!

Current Tasks:

- Removing questions for transcript text, doing this manually - trancripts not consistent enough for attempting automation
- Testing with regular expressions or stringr to identify everyday phrases and convert them into single word (by placing dashes in the middle)

```
#testing withe str_replace_all

test_text <-  "أيوة أنا إن شاء الله جاية و بعدين هروح إن شاء الله "
new_text <- test_text
isa <- "إن شاء الله"
isa_dash <- "إن-شاء-الله"
str_replace_all(new_text,isa,isa_dash)
```

```
## [1] "أيوة أنا إن-شاء-الله جاية و بعدين هروح إن-شاء-الله "
```

```
#success!
```

next: need to convert all double spaces to single spaces

Date Thursday 3/31/22

Hayes: relate whether parent separates to mention of religion and top frequent words? go with the easier option for the purposes of this class.

Kaya: this could be just a methods thing to see if it's feasible, also thinks the separator/combiner analysis is interesting and enough for this class.

maybe focus on methodology of identifying religious salience in Arabic text;

- identifying religious every day phrases
- identifying explicit mentions of religion
- regular expressions to identify phrases, and to remove questions

Date Friday 3/25/22

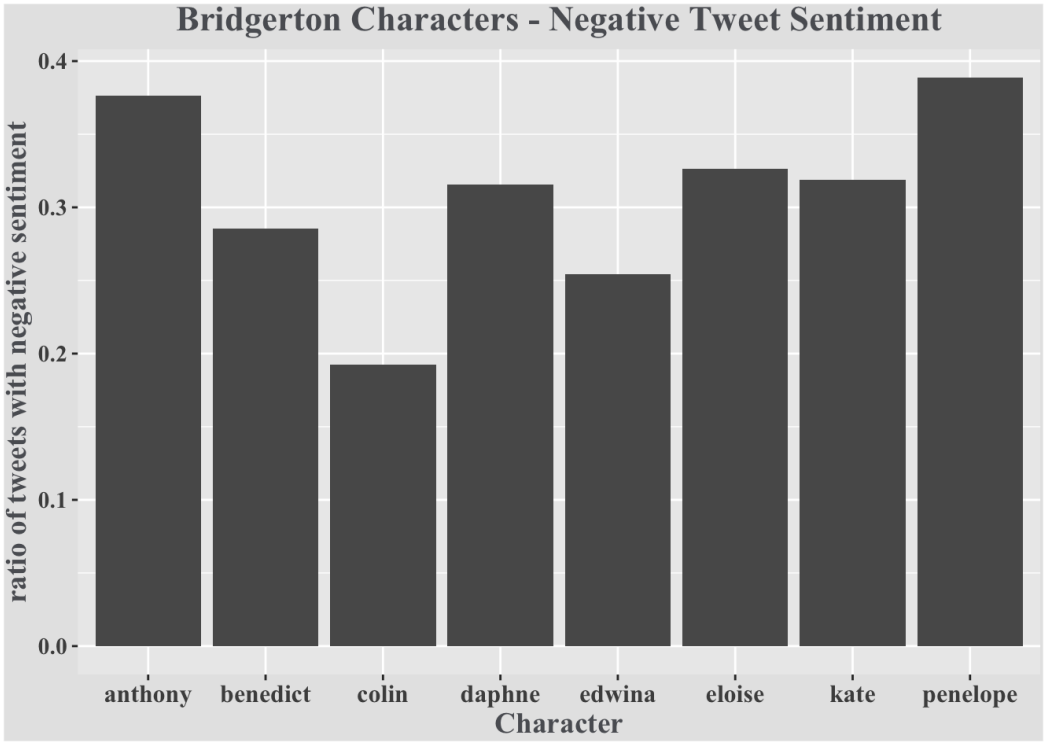
Analysis of Tweets about the TV Show Bridgerton on the premiere of its second season (which was today 3/25)

- Data was obtained using the TAGS tool with #bridgerton in search field
- The idea is to perform a sentiment analysis using Vader to identify negative and positive sentiments and how they associate with different characters.
- Analysis:
    - Spreadsheet downloaded into csv file and read into R.
    - Retweets were removed from data frame, this took the # of tweets down from 2,521 to 859.
    - Data frame column with tweet text was written into text file, and used to run an analysis on MEH.
    - Frequency list from MEH used to identify top mentioned characters.
    - Using Vader, sentiment analysis was conducted on list of tweets.
    - Data frame created to list character names along with:
        - number of tweets where negative sentiment >0
        - number of tweets where positive sentiment >0
        - ratio variable for each
- This is a very simplistic analysis and the raw data probably needed more cleaning beforehand.
- The charts show how the characters compare on the ratios of tweets with positive/negative sentiment, and the first thing that pops out is that the same character (Penelope) had the highest score on both. However, she is mentioned with the least frequency, so that may be a factor. Another factor could be that most tweets that mention her actually have a score >0 on the positive/negative sentiments. Since a tweet can mention more than one character, it's possible that all tweets mentioning Penelope also mentioed another character, but it affected Penelope's ratios because she had the lowest frequency.
- I would want to run an analysis like this with a larger number of tweets, which was limited by using TAGS. I couldn't get rtweet to work for me yet, unfortunately.

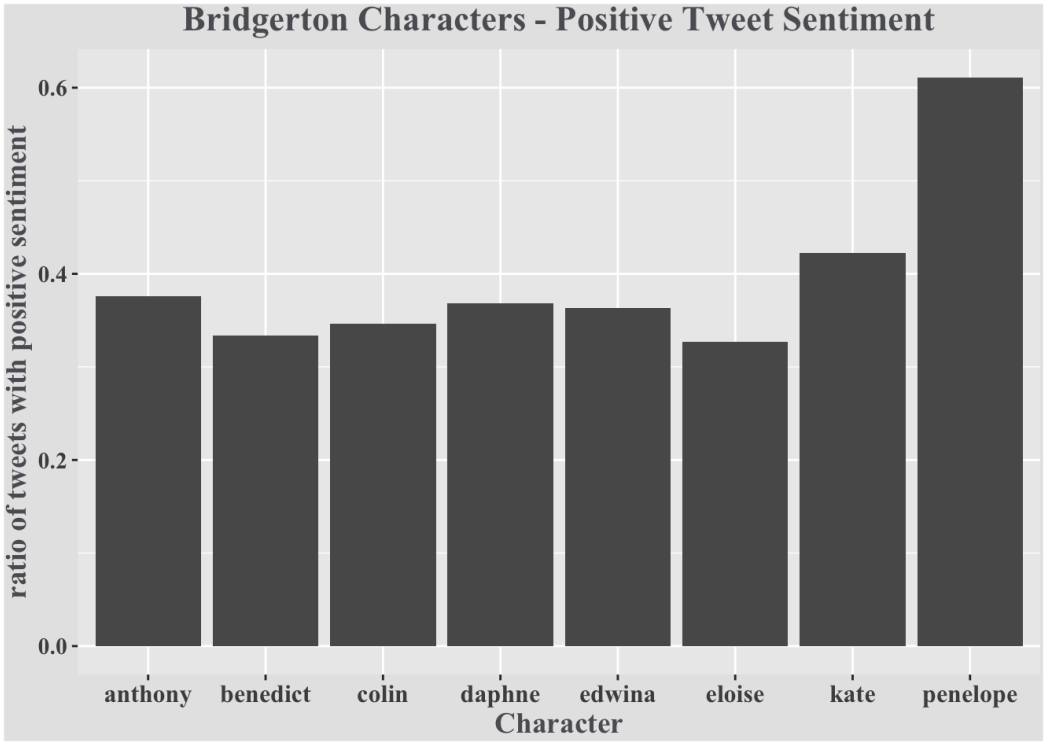
```
tweets <- read.csv("twitter#bridgerton.csv")
tweetsNoRT <- tweets[!substr(tweets$text,1,2)=="RT",] #removing retweets
write.table(tweetsNoRT$text,"tweetsNoRT.txt",row.names = FALSE, col.names=FALSE, quote=FALSE) #expo
rting text file for MEH
TopChar <- read.csv("MEH_topchar.csv") #reading edited MEH frequency list with top character names
TopChar
```

```
##   Character Frequency
## 1      kate       116
## 2   anthony       109
## 3    edwina        55
## 4    eloise        49
## 5     colin        26
## 6  benedict        21
## 7    daphne        19
## 8  penelope        18
```

```
tweetsNoRT$text <- tolower(tweetsNoRT$text) #converting to lower case
vadertweets <- vader_df(tweetsNoRT$text) #running sentiment analysis with vader
#preparing data frame to compile top character names with counts of tweets with positive/negative s
entiment
names <- TopChar$Character
neg <- rep(0,length(names))
pos <- rep(0,length(names))
negratio <- rep(0,length(names))
posratio <- rep(0,length(names))
freq <- TopChar$Frequency
char_sentiment <- data.frame(names,neg,pos,freq,negratio,posratio)

#loop to count tweets containing each character name that scored on negative/positive sentiment
for(i in 1:length(names))
{
  nametemp<- as.character(char_sentiment$names[i])
  vadertemp <- vadertweets[str_detect(vadertweets$text,nametemp),]
  char_sentiment$neg[i] <- nrow(vadertemp %>% filter(neg>0))
  char_sentiment$pos[i] <- nrow(vadertemp %>% filter(pos>0))
}
#calculating ratios for better comparison
char_sentiment$negratio <- char_sentiment$neg/char_sentiment$freq
char_sentiment$posratio <- char_sentiment$pos/char_sentiment$freq

ggplot(char_sentiment, aes(x = names, y =negratio)) +
  geom_bar(stat = "identity") +
  xlab("Character") +
  ylab("ratio of tweets with negative sentiment") +
  ggtitle("Bridgerton Characters - Negative Tweet Sentiment") +
  scale_fill_manual(values=c("#CA3145","#13CC66")) +
  theme(text = element_text(family = "Times New Roman", face = "bold", size =14, color = "#5B5E63"
)) +
  theme (plot.background = element_rect(
    fill = "grey90",
    size = 4), axis.text.x = element_text(size = 12), plot.title=element_text(hjust=0.5))
```



```
ggplot(char_sentiment, aes(x = names, y =posratio)) +
  geom_bar(stat = "identity") +
  xlab("Character") +
  ylab("ratio of tweets with positive sentiment") +
  ggtitle("Bridgerton Characters - Positive Tweet Sentiment") +
  theme(text = element_text(family = "Times New Roman", face = "bold", size =14, color = "#5B5E63"
)) +
  theme (plot.background = element_rect(
    fill = "grey90",
    size = 4), axis.text.x = element_text(size = 12), plot.title=element_text(hjust=0.5))
```



Date: Tuesday 3/22/22

- Grey's idea of developing semantic spaces for a specific topic (in reference to twitter climate change article), after my thought of developing a dictionary with boiled down tweets and labels.

Analysis

- RQ: how does the salience of religion vary depending on theme? when is religion more likely to be salient?
- Use dedoose to qualitatively code passages (utterances, or responses to questions, could it be decided based on context how long it is/)
- Each identified passage would be a unit, and the speaker's characteristics would be associated variables - including previous codes identified in whole interview?
- Possible dataframe structure (compiled from Dedoose export):
  - ID | parent ID | main theme | parent gender | relevant codes associated with interview
- Potential main themes (to be updated once interviews are initially scanned and coded):
  - child description
  - daily activities
  - parent education
  - perceived parental roles
  - moral tarbeya
  - discipline
  - school

Relevant feedback from lab meeting:

- Themes vs topics?
- weighting measures:
  - weighting codes in Dedoose
  - proportion within topic, calculate per person and take average?
  - multilevel model with passages nested within topic within the person, with religious salience as DV and topic as IV

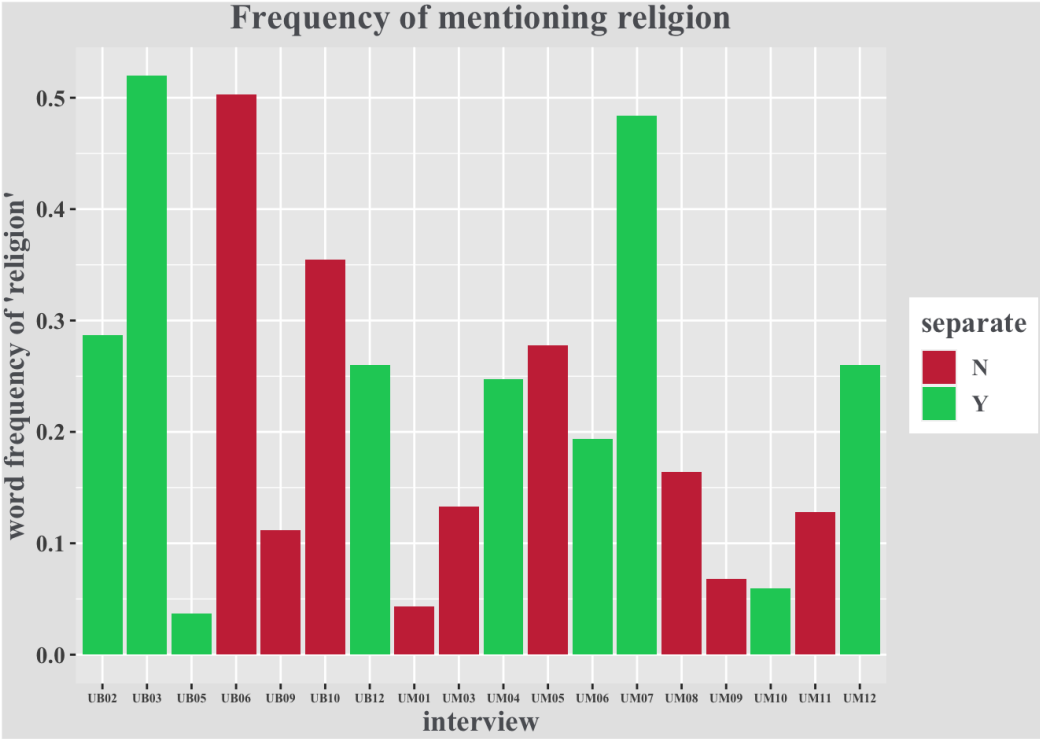
Testing R code to covert docx files to txt files

```
 library(textreadr)
#setwd("~/Documents/UConn/15- Spring 2022/PSYC 6783/Analysis Tools")
## .docx
#docx_t <- system.file("test.docx",
#package = "textreadr")  # not necessary ad did not work
test <- read_document("UB03.docx") # worked! seems to importan as character vector separated by emp
ty lines
test2 <- read_document("UB03.docx", combine=TRUE) #yay! worked and put it all in one string
#test2
write.table(test2, file="test2.txt", row.names = FALSE, col.names=FALSE, fileEncoding = "UTF-8", qu
ote=FALSE)
write.table(test, file="test.txt", row.names = FALSE, col.names=FALSE, fileEncoding = "UTF-8", quot
e=FALSE)
#while combining into a single string with combine=TRUE may be better for some things, it will not
 include separate lines when writing into a text file, with the noncombined dataframe was exported
 into a text file maintaining line structure but no empty lines
test3 <- read_document("UB03.docx",remove.empty=FALSE,skip=0) #reads empty lines as empty rows
write.table(test3, file="test3.txt", row.names = FALSE, col.names=FALSE, fileEncoding = "UTF-8", qu
ote=FALSE) #this works and includes empty lines
#other option to put in something instead of empty lines
test4 <- test3
test4[test4 == ""] <- "##########"
write.table(test4, file="test4.txt", row.names = FALSE, col.names=FALSE, fileEncoding = "UTF-8", qu
ote=FALSE) #perfect!

#now need loop to do this will all files
#the following reads all file names in a folder and changes extension.

docx_files <- list.files(path = "~/Documents/UConn/15- Spring 2022/PSYC 6783/Analysis Tools/Final C
omplete Transcripts",pattern="*.docx")

noext_files <- substr(docx_files,1,nchar(docx_files)-5)

txt_files <- paste(noext_files, ".txt",sep="")

#loop function to read all files and write them as txt works!!!

#setwd("~/Documents/UConn/15- Spring 2022/PSYC 6783/Analysis Tools/Final Complete Transcripts")
for(i in 1:length(docx_files))
{
  tempdata <- read_document(docx_files[i],remove.empty=FALSE,skip=0)
  tempdata[tempdata == ""] <- "######################"
  write.table(tempdata, file=txt_files[i], row.names = FALSE, col.names=FALSE, fileEncoding = "UTF-
8", quote=FALSE)
}
```

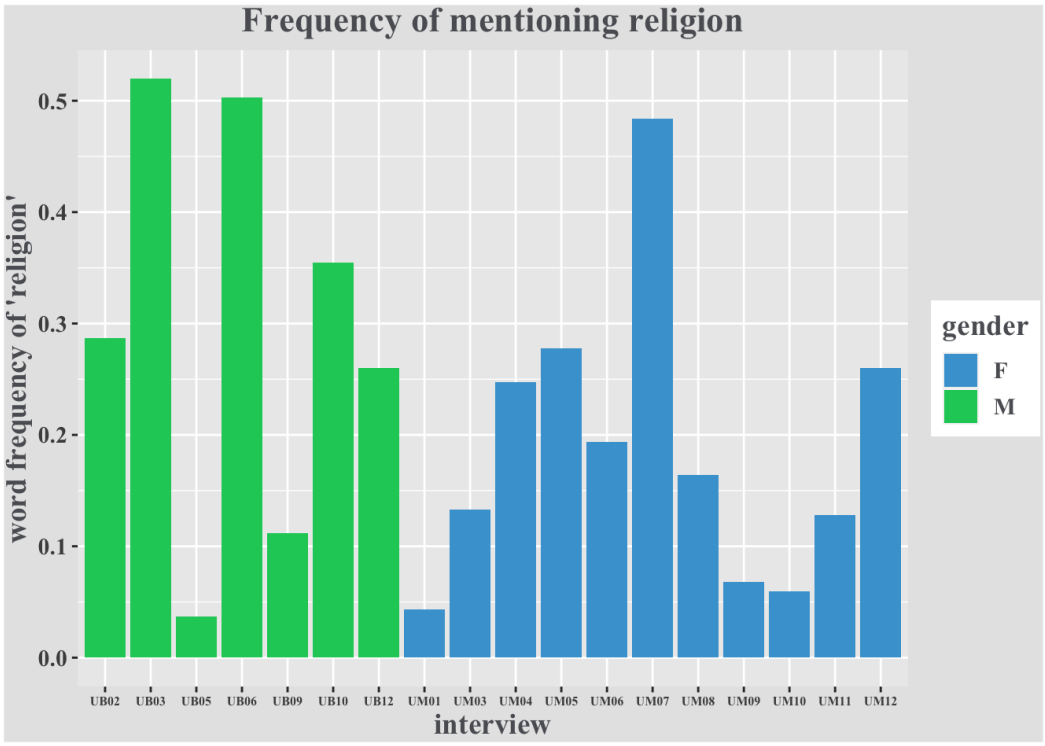
Date: Friday 3/11/22

- Ran MEH on updated transcript files, excluding noted ones. Still not complete finalized list, few more transcripts left.
    - Added compiled coversion list, including some English to Arabic ones
    - Need to compile list of stop words, start with top function words in MEH output
- Columns added for gender and whether theme of separating religion & morality was identified in interview (qualitatively coded)
- The following charts are an attempt to show frequency (verbose, frequency/word count) of the word “religion” in Arabic for each interview color coded for the separation theme (separate) and for parent gender. They will likely need to be redone once interview transcripts are finalized and cleaned up. I might then reconsider whether it’s meaningful.

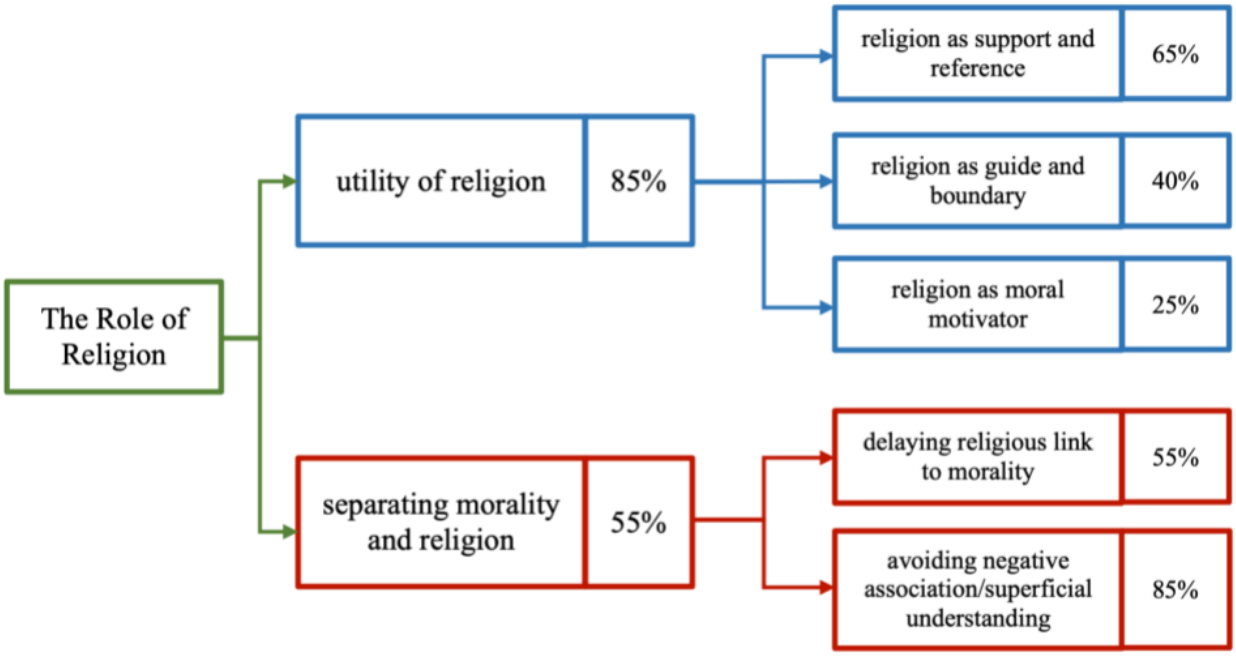
```
MEHoutput <- read.csv("MEHoutput0311.csv")

ggplot(MEHoutput, aes(x = Filename, y = دین۲, fill=separate)) +
  geom_bar(stat = "identity") +
  xlab("interview") +
  ylab("word frequency of 'religion'") +
  ggtitle("Frequency of mentioning religion") +
  scale_fill_manual(values=c("#CA3145","#13CC66")) +
  theme(text = element_text(family = "Times New Roman", face = "bold", size =14, color = "#5B5E63"
)) +
  theme (plot.background = element_rect(
    fill = "grey90",
    size = 4), axis.text.x = element_text(size = 6), plot.title=element_text(hjust=0.5))
```




```
ggplot(MEHoutput, aes(x = Filename, y = دین۲, fill=gender)) +
  geom_bar(stat = "identity") +
  xlab("interview") +
  ylab("word frequency of 'religion'") +
  ggtitle("Frequency of mentioning religion") +
  scale_fill_manual(values=c("#45A2D5","#13CC66")) +
  theme(text = element_text(family = "Times New Roman", face = "bold", size = 14, color = "#5B5E63"
)) +
  theme (plot.background = element_rect(
    fill = "grey90",
    size = 4),axis.text.x = element_text(size = 6), plot.title=element_text(hjust=0.5))
```



- I also included the following figure, which identifies the qualiatively coded themes related to how parents related morality and religion.



Role of Religion in Moral Tarbeya (Upbringing)

Date: Tuesday 3/8/22

Could there be a benefit from using multilevel/mixed effects modeling? Perhaps nest by interview? The idea would be that if we're looking at word frequencies, that may reasonably be weighted by their "talkativeness"? In that case, we could add random intercepts and random slopes. If we're looking at how theme may predict religious reference (as represented by frequency of religion related words), then we would perhaps need …

relevant questions:

- what is the main outcome?
  - religious salience?
- what is the unit of analysis?
  - themed paragraph, where there are several per interview
- what are the possible predictors?

    - gender
    - theme
    - qualitatively coded themes per interview
- is this sample too small for this analysis?

Note: create themed dictionaries? like for religious salience?

if i only have if transcript is diced down into multiple pieces, then I would want to account for the person with randome affects.

determine unit of analaysis, based on what level for you think the outcome variable will vary –> determine this based on theoretical expertise of the subject rather than for statistical reasons. At which level do I want to look at how religious salience varies? Is it at the person level? the sentence level? the answer level? the diced down interview into themes level/

Date: Friday 3/4/22

Based on feedback class activity:

- On the issue of what to do with English words within the Arabic transcripts:
    - It's probably worth it to keep them in the original language unless I'm only looking an frequency of concepts without regard to language
    - Could look into creating dataset where unique words with the same meaning but different language have the same "code"
    - Are there any potential inferences to be made from which English words are most frequent? Are there particular issues/ceoncepts that parents tend to talk about in English rather than Arabic? **potential RQ**
- RQ related to single question where parents describe their children - probably not the best way to go due to small sample size.

The following table was an attempt to identify most appropriate analytical approach. I believe my priority would be to explore the feasibility of semantic analysis with Arabic text, and if not then meaning extraction method would likely be the way to go.

```
analysis_app <- read.csv("approaches.csv")

kable(analysis_app, caption="Analytical Approaches & RQs")
```

Analytical Approaches & RQs

Approach	Brief.Description	Notes	Applicable.RQs
Bag of Words	count of words and their types, no attention to syntax or order	works with Arabic, doesn't require dictionary	
Dictionary	categorizes words according to pre-existing "dictionary", no attention to word order	no dictionary available for Arabic, would have to create targeted dictionary	probably too much work to build relevant dictionary
Vector Based Semantic	extracting meaning based on context and co-occurrence patterns	does not need dictionary or human coding, needs corpora for training?, lack of available tools for Arabic, most papers found too technical — can something be implemented in R? can I find appropriat corpora for training?	In which contexts is religion more salient? Explore co-occurrences
Meaning & Sentiment Analysis	word frequency analysis with list of stop words and conversion list	useful for including stop words and conversion list, verified for Arabic	- Explore differences in top frequent words between parents who "separate" religion & morality and those who don't. (based on previous qualitative coding). - Compare frequency of religion-related words within different interview themes (based on divided questions or qualitative coding)

Date: Friday 2/25/22

Thinking about how to structure and clean my transcripts into a usable dataframe in R:

- word files converted into text files manually, for new transcripts coming in - do I need to do that? if changing file extension only, encoding is messed up, could change encoding with R code?
- think about how to structure different parts of a transcript into data frame:
    - need to remove my voice - should it be completely removed or put into it's own dataframe?
    - should have one dataframe with entire transcript in a column, should I also attempt to divide transcript portions based on questions?
    - should I attempt to write code that adds new column for transcript after removing stop words, or after making conversions? is that necessary if I will use MEH or LIWC?
- looking into possibilities in tidyverse and regular expressions - found book online for latter and reading

Brainstorming for which portion of transcript to use for analysis and RQs

- parents asked to describe their children to someone who doesn't know them
    - with only 19 interviews, would this be any more valuable than regular qualitative coding?
    - possible RQ would be to see how gender of parent and gender of child affect types of descriptors - but N probably too small
- exploratory approach to identify the top contexts in which religion is mentioned;
    - identify all religion related words, and then find the highest co-ocurrences.
    - identify general themes in interviews, based on qualitative coding and sentence structure, accordingly divide up transcripts into columns (e.g. child outcome goals, school, morality, discipline…), and then explore how and when religion is most likely mentioned?

Possible categories of religion related words/phrases

- referring to religion using that word specifically "دين" with all possible forms (added pronouns, articles, prepositions), review and keep adding to conversion list
- referring to God explicitly not within phrase or idiom
- everyday colloquial phrases that include the word Allah — can an R program identify them and convert them, potentially make them dash separated to be identified as a single word?

NEW RESOURCES DISCOVERED FOR ARABIC

- **Quranic Arabic Corpus** (corpus.quran.com), website developed in university of leeds, still exploring access. Can be very useful to identify conversion words, has dictionary list of all words in the Quran categorized by root word.
- Open Source Arabic Corpora (see reference below)
- Found other collection of Arabic corpora (Dr. Mourad Abbas, Algeria), more here: http://aracorpus.e3rab.com/index.php?content=english (http://aracorpus.e3rab.com/index.php?content=english)

New Literature

- Sawalha, M., Atwell, E., & Abushariah, M. A. (2013, February). SALMA: standard Arabic language morphological analysis. In 2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA) (pp. 1-6). IEEE
- Alayba, A. M., Palade, V., England, M., & Iqbal, R. (2017, April). Arabic language sentiment analysis on health services. In 2017 1st international workshop on arabic script analysis and recognition (asar) (pp. 114-118). IEEE.
- Alyafeai, Z., Al-shaibani, M. S., Ghaleb, M., & Ahmad, I. (2021). Evaluating various tokenizers for arabic text classification. arXiv preprint arXiv:2106.07540.
- Saad, Motaz & Ashour, Wesam. (2010). OSAC: Open Source Arabic Corpora. 10.13140/2.1.4664.9288.
    - authors discuss creating open source accessible arabic corpora collected from different contexts (news, religion, family, education, sports…), still trying to figure out how to access it, only found java file

---

Date: Friday 2/18/22

MEH Analysis Test

I ran the Meaning Extraction Helper on 19 transcript files (parent interviews). The files include some full transcripts, partial transripts, and some notes from interviews. This is definitely not the dataset that I will continue working with, I am yet to finish cleaning the data and getting the rest of the full transcripts (hired transcriber working on the rest). Before further analysis, I will also need to remove my voice from the interview (the interview questions), so that the analysis runs only on what the parents are saying. The files include both Arabic and English text. I still need to decide what to do with that later (some content words appear in both English and Arabic and are counted separately - should they be added to a conversion list?). Although this is not currently a great or ready dataset, I thought it would be most beneficial as a learning experience. There are more details about running the data on MEH in the notes beneath the table.
The following table shows the top 30 frequent words after running MEH. I included translations for the words in Arabic. You'll find a notes column indicating which words I need to include in conversion lists in the future. Looking at the most frequent words you can clearly tell this was text related to families and parenting. There are typical words like home, children, nursery, school…etc. However, the significant thing here is that there are multiple words related to religion (translated: religion, prayer,

Allah, Our Lord) that are very high up on the list, with very high frequencies. Although the first word on the list which translates to “we” would usually be counted as a function word, I included it here because I wanted to see how often it’s used compared to the singular “I”. Another interesting thing I noticed is words that indicate a binary right/wrong (wrong, the right thing to do, how things should be). This may not be useful, but I’m personally interested in the idea of always preaching how things “should be done” - although that may be a common thing in parenting conversations. Finally, the word mom appears a lot more than the word dad, and I’d be interested to map that to specific interviews of mothers and fathers. It’s worth noting that a lot of conversions related to the words mom and dad in Arabic need to be added to get a more accurate count. I think an important lesson from this excercise is how important conversions are to Arabic text - things like pronouns, prepositions and articles are typically part of a word (e.g. for mom, my mom, so mom would all be a single word).

```
top78 <- read.csv("Top78FrequentWords_0218.csv")

kable(top78[1:30,], caption="Top 30 Frequent Content Words")
```

Top 30 Frequent Content Words

Token	Translation.if.needed	Frequency	Docs_With_Token	ObservationPct	IDF	type	notes
احنا	we	538	16	84.21053	0.17185	content	
مدرسة	school	389	17	89.47368	0.11123	content	
دين	religion	287	19	100.00000	0.00000	content	
البيت	home	158	9	47.36842	0.74721	content	
ماما	mom	157	11	57.89474	0.54654	content	
صلاة	prayer	149	15	78.94737	0.23639	content	
الله	Allah	114	11	57.89474	0.54654	content	
بابا	dad	110	8	42.10526	0.86500	content	
ربنا	Our Lord	100	14	73.68421	0.30538	content	
غلط	wrong	99	12	63.15789	0.45953	content	
الولاد	children	90	11	57.89474	0.54654	content	conversion
الحضانة	nursery/preschool	83	7	36.84211	0.99853	content	
التربية	upbringing/raising kids	82	13	68.42105	0.37949	content	
مشكلة	problem	80	8	42.10526	0.86500	content	
الأخلاق	morals	74	17	89.47368	0.11123	content	
parent		62	14	73.68421	0.30538	content	
الشغل	work	60	9	47.36842	0.74721	content	
المفروض	how it’s supposed to be/ how it should be	56	10	52.63158	0.64185	content	
family		53	15	78.94737	0.23639	content	
والله	swear to God	50	9	47.36842	0.74721	content	
الطفل	child	45	7	36.84211	0.99853	content	conversion
الصح	the right thing (to do)/what’s right	45	10	52.63158	0.64185	content	
control		42	6	31.57895	1.15268	content	
قصة	story	39	5	26.31579	1.33500	content	
community		37	10	52.63158	0.64185	content	
strict		35	9	47.36842	0.74721	content	
التمرين	practice (sports)	33	4	21.05263	1.55814	content	
طفل	child	32	10	52.63158	0.64185	content	conversion
work		32	5	26.31579	1.33500	content	

Token	Translation.if.needed	Frequency	Docs_With_Token	ObservationPct	IDF	type	notes
god		32	8	42.10526	0.86500	content	conversion

Notes

- Found list of 750 Arabic stop words on github: https://github.com/mohataher/arabic-stop-words.git (https://github.com/mohataher/arabic-stop-words.git). These will be useful, but look like they're mostly fusha (formal Arabic).
- Ran MEH on transcript text files (without removing anything, and including some files with only noted transcripts), will identify more stop words based on most frequent words in file to add to stopwords file. It's really hard to pick out stop words from frequency list :(
- reviewed all words/tokens with frequency> 50 and identified stop words – added around 116 stop words to previous list (after removing duplicates)
- conversion list (*specifically searched for some root words of interest in initial frequency list*):
    - identified 15 words in frequency list with the root word of religion and added to conversion list, everything converted to religion or "دين", words that imply religiosity specifically were converted to "تدين", only 3 of those.
    - identified 24 words from root word for prayer in various noun and verb forms, all converted to prayer "صلاة"
    - 10 words related to supplication, all converted to "دعاء"
    - 3 words related to Quran
    - many more added, refer to conversion list file, many noted in output to be added to conversion list

Brainstorming

- one interview question that can work for easy analysis approaches is one asking parents to describe their children - easier to have an RQ here than with religion/morality question.
- another thought is to include the entire interview with an exploratory approach that aims to identify the top contexts in which religion is mentioned - in other words identify all religion related words, and then find the highest co-ocurrences.

Date: Tuesday 2/15/22

- MEH worked with Arabic and English text together
- would need to create Arabic stop words list
- also need to create conversion list for Arabic: for different verb applicaitons, perhaps bringing words back to their roots, also separating words from pronouns — this will be complicated…
- installed BUTTER, looks like it has more versatile functionality, but need to know exactly analysis pipeline and set it up before running - doesn't look like there are presets

---

Date: Monday 2/14/22

- created developer account for twitter API (keys in locked note on Notes) for Arabic corpora.
- try read_delim rather than read_Lines for loading text files

---

Date: Sunday 2/13/22

update on LIWC: using "Categorize Words" and file with mostly Arabic and some English text, it ran fine, and Arabic words were identified but not categorized - so at least I know the files can be read and parsed.

Watched video about creating dictionary file: https://www.youtube.com/watch?v=CXPfrkfs7eo (https://www.youtube.com/watch?v=CXPfrkfs7eo) Testing out simple Arabic dictionary – question: can files be analyzed in comparison to several dictionaries (English + Arabic), or would we need to combine dictionaries in one?

- with sample dictionary file, LIWC actually gave some output and identified the pronouns (yay!), verified through categorizing words:

- only worked when .txt file saved in Archive (note software) not TextEdit or Excel.
- actual text file looks weird when listing pronouns in Arabic, looks like numbers come first, probably just because of text direction - looks like it worked fine on LIWC

learned that with Arabic text in RMarkdown file, pdf output will not work! something to do with unicode text and LateX :/

Date: Saturday 2/12/22

Played around with some Arabic text:

- was able to import file with both Arabic and English text, looks like it delimits by new line into a vector
- flattening worked to put it all in one string
- some simple string operations seem to work as well, counting specific Arabic words in the file works well
- next step, figure out how to create data frame with all transcripts, each file being a row of data with different identifying columnw (file name, gender, length…?)

```
arabicTest <- readLines("UB_RTest.txt",encoding='UTF-8')
arabicTest # aaaand this worked!! yippee!!
arabicTest2 <- readLines("UB06.txt",encoding='UTF-8') #bigger text file include some English words
 within Arabic text
arabicTest2 # this also worked well, loks like delimiter is new line
# now trying some string operations
#arabicTest2 %>% str_count(pattern="حاجة")
#does not work, weird output
#head(arabicTest2)
UB06flat <- arabicTest2 %>% str_flatten() #flattening is the key!
UB06flat
UB06flat %>% str_count(pattern = "حاجة") #works and results verified, there are indeed 33 instances
of حاجةin the file
```

Date: Friday 2/11/22

Testing out how LIWC and R work with Arabic characters:
* Have not yet been able to find available Arabic dictionaries to use, might need to reach out to authors (though Nairan mentioned she was told it didn't work very well) * R can take Arabic text in code and store it in strings, but need to figure out how to import text file with Arabic text

```
#arabicTest <- readLines("/Users/huda/Documents/UConn/15- Spring 2022/PSYC 6783/Testing Arabic on
 R/UB.rtf",encoding='UTF-8') #runs but problem with rtf and seems to have only read formatting stuf
f
#arabicTest
arabchar<-'هدى'
arabchar  # works!
```

```
## [1] "هدى"
```

Date: Thursday 2/7/22 - Class

- Look into whether i'll need a non-human subjects approval if I'll use any online data

Date: Thursday 2/3/22 - Class

Used the LSA web tool to compare articles written about motherhood, fatherhood, and parenthood from the NY Times. T1=motherhood, T2=fatherhood, T3=parenthood
Doc T1 T2 T3
Text 1 1 0.48 0.64
Text 2 0.48 1 0.49
Text 3 0.64 0.49 1

Date: Friday 1/28/22

Additional prompts for this week

For this week, your lab notebook submission must include some output from one dictionary using LIWC OR bag-of-words text analysis tool (i.e., any approach discussed in class) and a brief description (3-5 sentences) of the interpretation of that result. You may use any dataset that you would like, but you might consider using a dataset that is somewhat related to your final project or research area.

Date: Tuesday 1/25/22 (from class)

Brainstorming:

Ideas for questions to approach analyzing my data and some issues to pay attention to

- how gender affects pronoun use, and sense of accountability and responsibility
- use of religious phrases overall and in relation to certain context
- differentiating religious terms used within everyday phrases, and intentionally to imply religious emphasis or framing
- identifying any mention or reference to sociopolitical events/changes
- be clear about what variables are controlled for and which ones can be used as points of comparison

Tools: Praat, audacity to identify speech and silence segments and timing/rhythm of conversations in general

---

Date: Friday 1/21/22

Brainstorming:

This week, I think I just barely scratched the surface. I know that I want to use my interview transcript data and these are the preliminary ideas I had in mind for how to approach analysis: (I will use English terms for some ideas but these will actually be in Arabic)

- iterative runs to identify recurrent words and build an Egyptian dialect dictionary
- identifying sets of words that tend to be present in the same sentence together
- analysis of content word frequency in comparison to coded themes (this isn't very clear in my head, but an example would be if there are specific types of recurrent words present when a certain theme is identified in a transcript)

Literaure Review:

So far, most papers I find looking at language analysis of Arabic text are doing so with social media content. I'm also not yet familiar with all the terminology used in this area, so it'll be a learning curve.

- Hayeri, N. (2014). Does gender affect translation?: analysis of English talks translated to Arabic (Doctoral dissertation).
    - This paper discusses the development of Arabic LIWC, and Pennebaker is on the committee. This is a dissertation examining how gender impacts translation from English to Arabic. They posit that the gender of the translator may have an effect on the style of translation which may in turn lead to certain meaning being lost.
- Oueslati, O., Cambria, E., HajHmida, M. B., & Ounelli, H. (2020). A review of sentiment analysis research in Arabic language. Future Generation Computer Systems, 112, 408-430.
    - reviews sentiment analysis in Arabic, specifically discusses dialect issues
- Essam, B.A. & Abdo, M.S. (2020). How Do Arab Tweeters Perceive the COVID-19 Pandemic? Journal of Psycholinguistic Research., 50(3), 507–521. https://doi-org.ezproxy.lib.uconn.edu/10.1007/s10936-020-09715-6 (https://doi-org.ezproxy.lib.uconn.edu/10.1007/s10936-020-09715-6)
    - This study analyzed Arabic tweets to understand Arab perceptions of COVID. They conducted a lexicon-based thematic analysis using corpus tools (I don't yet entirely understand what this means), LIWC, and applied R stylo. Linguists manually annotated the top 1000 keywords and came up with categories, which they then compared to the psychological attitudes of users through LIWC's algorithm. I like that they combined a somewhat qualitative technique and then utilized LIWC to use generated categories in another analysis. Among the findings, the one I found interesting is that the psychological categories of religion and health were dominant in COVID related Arabic tweets. This makes a lot of sense to me, but makes me wonder ho this category of religion is identified. The authors mention using an existing Arabic dictionary compitable with LIWC, but also mention translating something related to LIWC- it was a bit hard to understand.
- Al-Ali, M. N., & Shatat, H. A. (2021). Discoursal representation of masculine parenting in Arabic and English websites. Pragmatics.
    - I couldn't yet get access to this paper, but it's super interesting. It's looking at parenting articles in English and Arabic, and analyzing the style of discourse. English articles revealed more "shared parenting" approaches where as Arabic ones revealed more traditional approaches.

Note: This was a really hectic first week for me, as I tested positive for COVID a few days before the semester started and had to delay my return till I tested negative. I'm leaving tonight (1/21) so this was really rushed, I'm completing and submitting this from the airport at the gate - I apologize. Have a great weekend!