Codebook for data sets used in "Women's Authority in Patriarchal Social Movements: The Case of Female Salafi Preachers," AJPS

Richard Nielsen
rnielsen@mit.edu
3/7/2019

NOTE: The analysis for this paper relies on primary source documents I collected from the website saaid.net, it's associated Twitter feed, and archived versions on the website archive.org. As such, the data are unstructured and don't start as "variables" in a rectangular data matrix. I analyzed these unstructured data in a variety of ways, some of which do involve rectangular data matrices. The codebook below explains the variables in these data sets. All of these variables are constructed in the replication code, so you can check the code to see precisely what each variable represents.

I may have forgotten something. Please contact me if the data remain unclear to you after consulting this codebook and the replication code.

Saaid.net text data sources ~/data/saaid.net_text_data/

clericTextHolder_4feb2019.rds
This is the raw text data I scraped from saaid.net. It is saved as a .rds file (readable in R with the readRDS() function). It is a list where each element is the identifier for a preacher on the website, and each sub-element is the full text of a document by that preacher.

dtm.csv
This is a document term matrix for the text analysis. The rows are individual documents (labeled by the filepath to the document). The columns are individual terms in the documents. Each cell indicates the number of times each term appears in each document. These terms are stemmed.

dtmPrefixes.csv
This is a document term matrix for the text analysis. The rows are individual documents (labeled by the filepath to the document). The columns are individual terms in the documents. Each cell indicates the number of times each term appears in each document. These terms are stemmed, but I append columns for the Arabic prefixes that were removed in the stemming process.

dtmUnstemmed.csv

This is a document term matrix for the text analysis. The rows are individual documents (labeled by the filepath to the document). The columns are individual terms in the documents. Each cell indicates the number of times each term appears in each document. These terms are unstemmed.

mainRegressionData.csv

This is the rectangular data matrix for the regressions reported in the Appendix, Table 1. The unit of observation is the document. The variables are:

genderF: A binary indicator coded as 1 if the author is female and 0 if male.

id: A string that serves as a unique document id. It is the filepath to where the original document is saved.

doclength: The number of words in each document.

Monotheism: The estimated proportion of this topic from the STM.

Doctrine: The estimated proportion of this topic from the STM.

Heterodoxy: The estimated proportion of this topic from the STM.

Teaching: The estimated proportion of this topic from the STM.

Jurisprudence: The estimated proportion of this topic from the STM.

Worship: The estimated proportion of this topic from the STM.

Hadith: The estimated proportion of this topic from the STM.

Women: The estimated proportion of this topic from the STM.

Salawat: The estimated proportion of this topic from the STM.

Westernization: The estimated proportion of this topic from the STM.

SlAallh.freq: The frequency of this phrase (in Arabic) in the document.

SlAallh.dummy: A binary indicator for the presence of this phrase in the document.

bKary.freq: The frequency of this hadith narrator's name (in Arabic) in the document.

bKary.dummy: A binary indicator for the presence of this hadith narrator's name in the document.

tymy0.freq: The frequency of this hadith narrator's name (in Arabic) in the document.

tymy0.dummy: A binary indicator for the presence of this hadith narrator's name in the document.

byhQy.freq: The frequency of this hadith narrator's name (in Arabic) in the document.

byhQy.dummy: A binary indicator for the presence of this hadith narrator's name in the document.

hryr0.freq: The frequency of this hadith narrator's name (in Arabic) in the document.

hryr0.dummy: A binary indicator for the presence of this hadith narrator's name in the document.

Tbrany.freq: The frequency of this hadith narrator's name (in Arabic) in the document.

Tbrany.dummy: A binary indicator for the presence of this hadith narrator's name in the document.

dawd.freq: The frequency of this hadith narrator's name (in Arabic) in the document.

dawd.dummy: A binary indicator for the presence of this hadith narrator's name in the document.

S7y7.mslm.freq: The frequency of this hadith narrator's name (in Arabic) in the document.

S7y7.mslm.dummy: A binary indicator for the presence of this hadith narrator's name in the document.

trmiy.freq: The frequency of this hadith narrator's name (in Arabic) in the document.

trmiy.dummy: A binary indicator for the presence of this hadith narrator's name in the document.

compiler.dummy: A binary indicator for whether a document includes any of the hadith narrator's names.

compiler.freq: The frequency of all of the hadith narrator's names in the document.

mainTopic: The main topic of the document, based on which STM topic is most prevalent.

author: A string unique to the name of each author.

dukturaOnSaaid: Whether the author is listed as having a Doctoral degree in saaid.net.

includeAsPhd: A variable that is 1 for all men (whether they have a PhD or not), 1 for women with a PhD, and 0 for women without a PhD.

hasPhd: A variable that is 1 for preachers who have PhDs and 0 for preachers who don't regardless of gender.

doclength2: The number of words in each document divided by 1000.

FPhdInt: An interaction term between **genderF** and **hasPhD**.

termRegNData_allh.csv

This data set is one of a series that I use to estimate the proportions of men's and women's documents using certain hadith-related words in Table 1. The data sets are produced in a function in the code and saved. They all have a similar format.

The first column (no variable name) is a unique id for each document.

term: the frequency of the term "allh" in each document.

termdummy: an indicator variable for the presence of "allh" in each document.

female: an indicator for whether the author is female.

termRegNData_bKar.csv

This data set is one of a series that I use to estimate the proportions of men's and women's documents using certain hadith-related words in Table 1. The data sets are produced in a function in the code and saved. They all have a similar format.

The first column (no variable name) is a unique id for each document.

term: the frequency of the term "bKar" in each document.

termdummy: an indicator variable for the presence of "bKar" in each document.

female: an indicator for whether the author is female.

termRegNData_byhQ.csv

This data set is one of a series that I use to estimate the proportions of men's and women's documents using certain hadith-related words in Table 1. The data sets are produced in a function in the code and saved. They all have a similar format.

The first column (no variable name) is a unique id for each document.

term: the frequency of the term "byhQ" in each document.

termdummy: an indicator variable for the presence of "byhQ" in each document.

female: an indicator for whether the author is female.

termRegNData_dawd.csv

This data set is one of a series that I use to estimate the proportions of men's and women's documents using certain hadith-related words in Table 1. The data sets are produced in a function in the code and saved. They all have a similar format.

The first column (no variable name) is a unique id for each document.

term: the frequency of the term "dawd" in each document.

termdummy: an indicator variable for the presence of "dawd" in each document.

female: an indicator for whether the author is female.

termRegNData_hadith.csv

This data set is one of a series that I use to estimate the proportions of men's and women's documents using certain hadith-related words in Table 1. The data sets are produced in a function in the code and saved. They all have a similar format.

The first column (no variable name) is a unique id for each document.

term: the frequency of the term "hadith" in each document.

termdummy: an indicator variable for the presence of "hadith" in each document.

female: an indicator for whether the author is female.

termRegNData_hryr.csv

This data set is one of a series that I use to estimate the proportions of men's and women's documents using certain hadith-related words in Table 1. The data sets are produced in a function in the code and saved. They all have a similar format.

The first column (no variable name) is a unique id for each document.

term: the frequency of the term "hryr" in each document.

termdummy: an indicator variable for the presence of "hryr" in each document.

female: an indicator for whether the author is female.

termRegNData_Qal-Qwl (al)rswl.csv

This data set is one of a series that I use to estimate the proportions of men's and women's documents using certain hadith-related words in Table 1. The data sets are produced in a function in the code and saved. They all have a similar format.

The first column (no variable name) is a unique id for each document.

term: the frequency of the term "Qal-Qwl (al)rswl" in each document.

termdummy: an indicator variable for the presence of "Qal-Qwl (al)rswl" in each document.

female: an indicator for whether the author is female.

termRegNData_rDy allh.csv

This data set is one of a series that I use to estimate the proportions of men's and women's documents using certain hadith-related words in Table 1. The data sets are produced in a function in the code and saved. They all have a similar format.

The first column (no variable name) is a unique id for each document.

term: the frequency of the term "rDy allh" in each document.

termdummy: an indicator variable for the presence of "rDy allh" in each document.

female: an indicator for whether the author is female.

termRegNData_SlA allh.csv

This data set is one of a series that I use to estimate the proportions of men's and women's documents using certain hadith-related words in Table 1. The data sets are produced in a function in the code and saved. They all have a similar format.

The first column (no variable name) is a unique id for each document.

term: the frequency of the term "SlA allh" in each document.

termdummy: an indicator variable for the presence of "SlA allh" in each document.

female: an indicator for whether the author is female.

termRegNData_Tbran.csv

This data set is one of a series that I use to estimate the proportions of men's and women's documents using certain hadith-related words in Table 1. The data sets are produced in a function in the code and saved. They all have a similar format.

The first column (no variable name) is a unique id for each document.

term: the frequency of the term "Tbran" in each document.

termdummy: an indicator variable for the presence of "Tbran" in each document.

female: an indicator for whether the author is female.

termRegNData_tymy0.csv

This data set is one of a series that I use to estimate the proportions of men's and women's documents using certain hadith-related words in Table 1. The data sets are produced in a function in the code and saved. They all have a similar format.

The first column (no variable name) is a unique id for each document.
term: the frequency of the term "tymy0" in each document.
termdummy: an indicator variable for the presence of "tymy0" in each document.
female: an indicator for whether the author is female.

Twitter data ~/data/twitter_data

tweetdat_9aug2018.rds

This contains some information about the tweets I scraped. I am providing it in .rds format to try to make it clear that I'm intending this only for researchers seeking to replicate my results, rather than as an attempt to distribute information about tweets (which are proprietary to Twitter).

id: Unique identifier for each tweet.
genderF: The gender of the saaid.net author whose work is the subject of the tweet.
retweets: The number of retweets for the tweet.
favorites: The number of times the tweet was "favorited".
totalRetweetAudience: The sum of the Twitter users who were friends with Twitter users who retweeted each Tweet. For example, if 3 people retweeted the original tweet and those people had 10 friends each, then this would 30.
meanRetweetAudience: The average number of Twitter users who were friends with Twitter users who retweeted each Tweet. For example, if 3 people retweeted the original tweet and those people had 10 friends each, then this would 10.
pctRetweetsMissingGender: This is the proportion of retweets from users where I was not able to confidently guess the user gender.
pctRetweetsFemale: The proportion of retweets by twitter users I code as women.
nTweetersForGenderPct: The total number of retweeters I used to calculate the previous two variables.
pctRetweetsFemaleUpper: The proportion of retweets by female twitter users *if I assume that all of the users with unknown gender are female.*
pctRetweetsFemaleLower: The proportion of retweets by female twitter users *if I assume that all of the users with unknown gender are male.*
pctRetweetsMissingCountry: The percent of retweets for which I can't identify the country.
pctRetweetsSaudi: The percent of retweets that are from Saudi Arabia.
expandedURLs: The URLs in the tweets (links to the writing on saaid.net).

Archive.org data ~/data/archive.org_data

~/data/archive.org_data/archive.org scrape

This subdirectory contains 1,104 files of raw data I scraped from archive.org. These data are not structured. The file names correspond to the timestamp of the saved snapshot on archive.org. The contents are in html format.

yearAdded_13nov2018.rds

This is an .rds file (open in R using the readRDS() function). This is not a rectangular data matrix. It is a vector where the element names uniquely identify the female preachers and the elements are the year that the preacher was added to saaid.net for the first time. This .rds file is called in the main analysis script.

preacher info ~/data/preacher info/

daeyat_bio_data.csv

This is a hand-coded data set I made by looking at the subpage for each preacher on saaid.net and recording what credential they were listed as having

foldername: This is a unique identifier for each preacher.

genderF: This is an indicator variable for whether the preacher is female.

credentialOnSaaid: This is the credential listed on the website (coded by hand).