**It’s a Small World**

you can [download the data for yourself](https://smallworldofwords.org/en/project/research).

library(tidyverse)

swow\_raw <- read\_csv("SWOW-EN.R100.csv") %>%

select(-education, -X1) %>%

mutate(gender = case\_when(gender == "Fe" ~ "Women",

gender == "Ma" ~ "Men",

gender == "X" ~ "Non-binary"),

gender = fct\_infreq(gender))

swow\_raw

## # A tibble: 1,228,200 x 11

## id participantID age gender nativeLanguage country created\_at cue R1 R2

##

## 1 29 3 33 Women United States Australia 2011-08-12 02:19:38 although nevertheless yet

## 2 30 3 33 Women United States Australia 2011-08-12 02:19:38 deal no cards

## 3 31 3 33 Women United States Australia 2011-08-12 02:19:38 music notes band

## 4 32 3 33 Women United States Australia 2011-08-12 02:19:38 inform tell rat on

## 5 33 3 33 Women United States Australia 2011-08-12 02:19:38 way path via

## 6 34 3 33 Women United States Australia 2011-08-12 02:19:38 none zilch nada

## 7 35 3 33 Women United States Australia 2011-08-12 02:19:38 extra plus special

## 8 36 3 33 Women United States Australia 2011-08-12 02:19:38 will free future

## 9 37 3 33 Women United States Australia 2011-08-12 02:19:38 paper book news

## 10 38 3 33 Women United States Australia 2011-08-12 02:19:38 w v x

## R3

##

## 1 but

## 2 shake

## 3 rhythm

## 4

## 5 method

## 6 zero

## 7 additional

## 8 death

## 9 writing

## 10 what

## # ... with 1,228,190 more rows

The available dataset as it exists when I downloaded it includes 1,228,200 word associations (each of which involve four words, i.e. three connections) by 83,864 unique participants. When a participant starts on a word association, this project has them move forward through three hops in a chain, from cue to R1 to R2 to R3, and then start over with a new cue. Participants can go through many cues in any given session.

Participants can also report other information about themselves. For example, what is the age distribution?

median\_age <- swow\_raw %>%

distinct(participantID, age) %>%

pull(age) %>%

median()

swow\_raw %>%

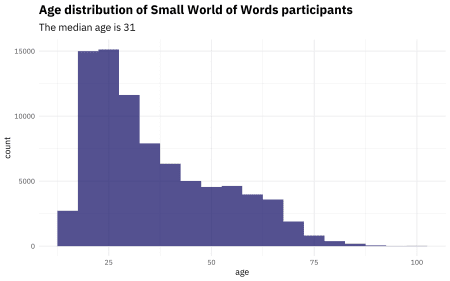
distinct(participantID, age) %>%

ggplot(aes(age)) +

geom\_histogram(alpha = 0.8, binwidth = 5, fill = "midnightblue") +

labs(title = "Age distribution of Small World of Words participants",

subtitle = paste("The median age is", median\_age))



There are lots of young folks represented in this project, as is typical for online surveys. What about gender?

swow\_raw %>%

distinct(participantID, gender) %>%

count(gender) %>%

mutate(Percent = n / sum(n)) %>%

ggplot(aes(fct\_reorder(gender, n), Percent)) +

geom\_col(alpha = 0.8, fill = "midnightblue") +

coord\_flip() +

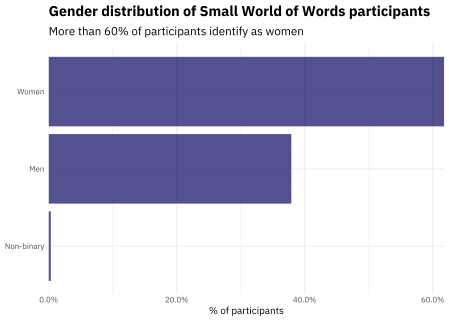
scale\_y\_continuous(labels = percent\_format(),

expand = c(0,0)) +

labs(x = NULL, y = "% of participants",

title = "Gender distribution of Small World of Words participants",

subtitle = "More than 60% of participants identify as women")



In this project, women were more likely to participate than other genders.

This project is international, pulling participants from many native languages. It also allows folks to specify whether they are a US English speaker, a UK English speaker, etc.

native\_languages <- swow\_raw %>%

distinct(participantID, nativeLanguage) %>%

pull(nativeLanguage)

swow\_raw %>%

distinct(participantID, nativeLanguage) %>%

count(nativeLanguage) %>%

top\_n(15) %>%

ggplot(aes(fct\_reorder(nativeLanguage, n), n)) +

geom\_col(alpha = 0.8, fill = "midnightblue") +

coord\_flip() +

scale\_y\_continuous(labels = comma\_format(),

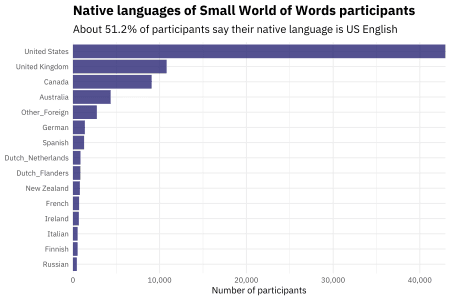
expand = c(0,0)) +

labs(x = NULL, y = "Number of participants",

title = "Native languages of Small World of Words participants",

subtitle = paste("About", percent(mean(native\_languages == "United States")),

"of participants say their native language is US English"))



So that’s a little bit of EDA to understand this project and its participants. Now let’s dig into the word associations!

**Building forward associations**

This is a rich, detailed dataset and there are so many directions we could go with it. In taking a first stab, let’s look at all the forward associations in the whole project. This means we will treat the “hop” from the cue to the first association the same as the “hop” from the first to second association, which certainly isn’t entirely correct. It’s a choice to start from, though.

swow\_forward <- swow\_raw %>%

select(from = cue, to = R1, gender, age, nativeLanguage) %>%

bind\_rows(

swow\_raw %>%

select(from = R1, to = R2, gender, age, nativeLanguage)

) %>%

bind\_rows(

swow\_raw %>%

select(from = R2, to = R3, gender, age, nativeLanguage)

) %>%

filter(!is.na(to))

swow\_forward

## # A tibble: 3,403,790 x 5

## from to gender age nativeLanguage

##

## 1 although nevertheless Women 33 United States

## 2 deal no Women 33 United States

## 3 music notes Women 33 United States

## 4 inform tell Women 33 United States

## 5 way path Women 33 United States

## 6 none zilch Women 33 United States

## 7 extra plus Women 33 United States

## 8 will free Women 33 United States

## 9 paper book Women 33 United States

## 10 w v Women 33 United States

## # ... with 3,403,780 more rows

Now that we have all the forward associations, we can find the most common associations for any individual word with some simple dplyr operations. What about… coffee? 

swow\_forward %>%

filter(from == "coffee") %>%

count(to, sort = TRUE)

## # A tibble: 711 x 2

## to n

##

## 1 tea 80

## 2 milk 33

## 3 drink 31

## 4 Starbucks 28

## 5 caffeine 15

## 6 cup 15

## 7 morning 15

## 8 hot 13

## 9 chocolate 12

## 10 island 12

## # ... with 701 more rows

Or… maybe you are in a holiday Christmas celebratory mood, and want to know what people associate with the word “Christmas”.

swow\_forward %>%

filter(from == "Christmas") %>%

count(to, sort = TRUE)

## # A tibble: 759 x 2

## to n

##

## 1 birthday 56

## 2 holiday 45

## 3 tree 43

## 4 Santa 34

## 5 snow 33

## 6 happy 31

## 7 red 29

## 8 presents 28

## 9 winter 24

## 10 Jesus 19

## # ... with 749 more rows

**Comparing groups**

This project recorded information about the participants themselves, so we can dig into how different kinds of people associate words. For example, let’s start with gender and comparing folks who identify as men and women. What differences do we see with the word “water”?

swow\_forward %>%

filter(from == "water",

gender %in% c("Men", "Women")) %>%

group\_by(to) %>%

filter(n() > 30) %>%

ungroup %>%

count(gender, to, sort = TRUE) %>%

spread(gender, n, fill = 0) %>%

mutate\_if(is.numeric, funs((. + 1) / (sum(.) + 1))) %>%

mutate(logratio = log2(Women / Men)) %>%

top\_n(15, abs(logratio)) %>%

ggplot(aes(fct\_reorder(to, logratio), logratio,

fill = logratio < 0)) +

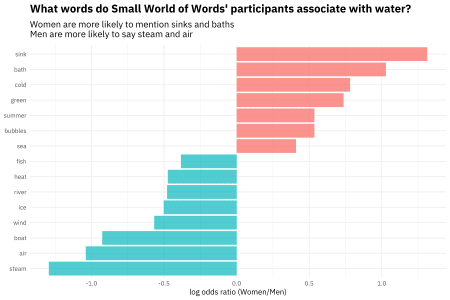
geom\_col(alpha = 0.8, show.legend = FALSE) +

coord\_flip() +

labs(x = NULL, y = "log odds ratio (Women/Men)",

title = "What words do Small World of Words' participants associate with water?",

subtitle = "Women are more likely to mention sinks and baths\nMen are more likely to say steam and air")



Notice the dramatic contrasts between domestic water uses like sinks and baths with more scientific word about water like steam. We see how socialized and differentiated women’s language is, even with something that seems neutral like water.

What about differences between US and UK English?

swow\_forward %>%

filter(from == "love",

nativeLanguage %in% c("United States", "United Kingdom")) %>%

group\_by(to) %>%

filter(n() > 30) %>%

ungroup %>%

count(nativeLanguage, to, sort = TRUE) %>%

spread(nativeLanguage, n, fill = 0) %>%

mutate\_if(is.numeric, funs((. + 1) / (sum(.) + 1))) %>%

mutate(logratio = log2(`United States` / `United Kingdom`)) %>%

ggplot(aes(fct\_reorder(to, logratio), logratio,

fill = logratio < 0)) +

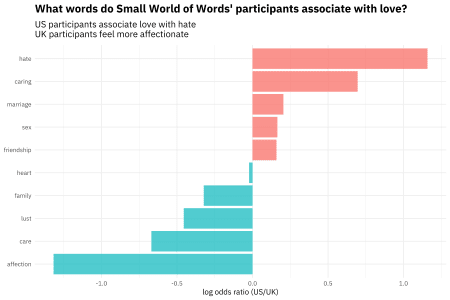
geom\_col(alpha = 0.8, show.legend = FALSE) +

coord\_flip() +

labs(x = NULL, y = "log odds ratio (US/UK)",

title = "What words do Small World of Words' participants associate with love?",

subtitle = "US participants associate love with hate\nUK participants feel more affectionate")



Well, alrighty then. 

**Changes with age**

We can apply some functional programming and modeling to look at how these word associations change with age. Let’s take the word “money”, and start by calculating, for 5-year bins, the number and proportion of words associated for each bin.

swow\_freq <- swow\_forward %>%

filter(from == "money",

age < 80) %>%

mutate(age = age %/% 5 \* 5) %>%

count(age, to) %>%

complete(age, to, fill = list(n = 0)) %>%

group\_by(age) %>%

mutate(age\_total = sum(n),

percent = n / age\_total) %>%

ungroup %>%

group\_by(to) %>%

filter(sum(n) > 25) %>%

ungroup

swow\_freq

## # A tibble: 741 x 5

## age to n age\_total percent

##

## 1 15 bank 47 1118 0.0420

## 2 15 banks 7 1118 0.00626

## 3 15 bills 2 1118 0.00179

## 4 15 business 11 1118 0.00984

## 5 15 buy 7 1118 0.00626

## 6 15 car 4 1118 0.00358

## 7 15 cash 30 1118 0.0268

## 8 15 change 5 1118 0.00447

## 9 15 charity 5 1118 0.00447

## 10 15 cheap 5 1118 0.00447

## # ... with 731 more rows

Now let’s fit some models using glm() since this is count data to predict the counts out of the total for each age bin from the age. We can then tidy() the output of the modeling, adjust the p-values for multiple comparisons since we looked at a bunch of words at one time, and make a volcano-style plot to compare the effect size with the p-value.

library(broom)

library(ggrepel)

slopes <- swow\_freq %>%

nest(-to) %>%

mutate(models = map(data, ~ glm(cbind(n, age\_total) ~ age, .,

family = "binomial"))) %>%

unnest(map(models, tidy)) %>%

filter(term == "age") %>%

arrange(estimate) %>%

mutate(p.value = p.adjust(p.value))

slopes %>%

ggplot(aes(estimate, p.value)) +

geom\_vline(xintercept = 0, lty = 2, alpha = 0.7, color = "gray50") +

geom\_point(color = "midnightblue", alpha = 0.8, size = 2.5) +

scale\_y\_log10() +

geom\_text(data = filter(slopes,

p.value >= 0.5),

aes(label = to),

family = "IBMPlexSans",

vjust = 0, nudge\_y = 0.02,

check\_overlap = TRUE) +

geom\_text\_repel(data = filter(slopes,

p.value < 0.5),

aes(label = to),

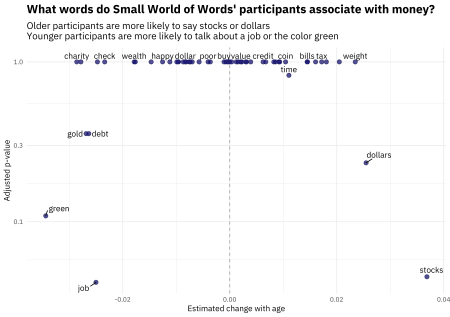
family = "IBMPlexSans") +

labs(x = "Estimated change with age",

y = "Adjusted p-value",

title = "What words do Small World of Words' participants associate with money?",

subtitle = "Older participants are more likely to say stocks or dollars\nYounger participants are more likely to talk about a job or the color green")



The younger someone is, the more likely they are to associate money with a job, or the color green or gold. The older someone is, the more likely they are to associate money with dollars and stocks. Let’s look at the top terms associated with money that exhibit change with age in terms of a small p-value.

slopes %>%

top\_n(6, -p.value) %>%

inner\_join(swow\_freq) %>%

ggplot(aes(age, percent)) +

geom\_point() +

geom\_smooth() +

facet\_wrap(~ to) +

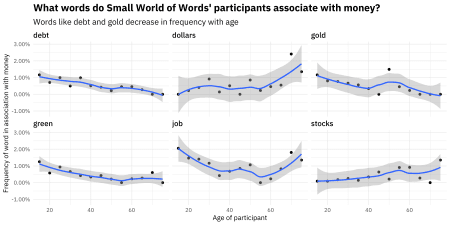
scale\_y\_continuous(labels = percent\_format()) +

labs(x = "Age of participant",

y = "Frequency of word in association with money",

title = "What words do Small World of Words' participants associate with money?",

subtitle = "Words like debt and gold decrease in frequency with age")



Younger respondents were more likely to associate words like DEBT with money. 

**The End**

There is so much more that could be done with this dataset. You could build a network data structure between the words and do various kinds of network analysis, and I didn’t touch any of the differences in the intial cue vs. the later hops. Notice that this dataset is all about words, but I didn’t ever load the [tidytext](https://cran.r-project.org/package=tidytext) package for this analysis.