Exercise 23

Marc Dotson

1. Run a topic model with three topics in roomba\_reviews.
2. Visualize the topics and name them. Justify your names.
3. Render the Quarto document into Word and upload to Canvas.

**Five points total, one point each for:**

* **Tokenizing Review into words.**
* **Using cast\_dtm() to create a DTM.**
* **Running LDA with k = 3.**
* **Visualizing and naming the topics, with justification for each name.**
* **Submitting a rendered Word document.**

## Run a Topic Model

# Load packages.  
library(tidyverse)

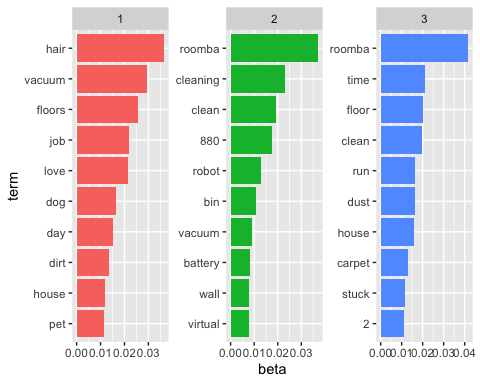
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ dplyr 1.1.4 ✔ readr 2.1.4  
✔ forcats 1.0.0 ✔ stringr 1.5.1  
✔ ggplot2 3.4.4 ✔ tibble 3.2.1  
✔ lubridate 1.9.3 ✔ tidyr 1.3.0  
✔ purrr 1.0.2   
── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
✖ dplyr::filter() masks stats::filter()  
✖ dplyr::lag() masks stats::lag()  
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(tidytext)  
library(topicmodels)  
  
# Import data, tokenize, remove stop words, cast a DTM, run a topic model.  
set.seed(42)  
fit\_lda3 <- read\_csv(here::here("Data", "roomba\_reviews.csv")) |>  
 mutate(review\_id = row\_number()) |>   
 unnest\_tokens(word, Review) |>   
 anti\_join(stop\_words, join\_by(word)) |>   
 count(word, review\_id) |>  
 cast\_dtm(review\_id, word, n) |>   
 LDA(k = 3, method = "Gibbs")

Rows: 1833 Columns: 5  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (4): Date, Product, Title, Review  
dbl (1): Stars  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

## Name the Topics

# Visualize the topics.  
fit\_lda3 |>   
 tidy(matrix = "beta") |>   
 group\_by(topic) |>   
 top\_n(10, beta) |>  
 ungroup() |>  
 mutate(term = reorder\_within(term, beta, topic)) |>  
 ggplot(aes(x = beta, y = term, fill = as.factor(topic))) +  
 geom\_col(show.legend = FALSE) +  
 facet\_wrap(~ topic, scales = "free") +  
 scale\_y\_reordered()



Let’s name the topics.

* **Features**: This topic is highest on roomba and 880 and the features associated with the features of a robotic vacuum: cleaning, clean, vacuum, battery. This seems positive overall.
* **Flaws**: This topic is highest overall on the time, floor, house, clean, and stuck – all possible flaws or pain points with robotic vacuums.
* **Functions**: This topic is highest overall on hair, floors, love, dog, and dirt all part of the function of robotic vacuums and likely positive overall.