Assignment 14

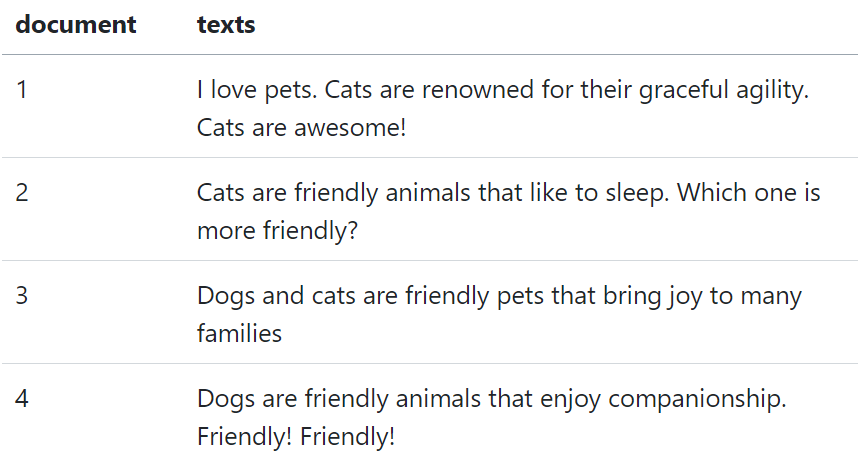
Allison Tessman

2024-04-11

Question 1. Given the following text dataset. Calculate by hand for this question.

* Calculate term frequency of term “animals” for each document

knitr::include\_graphics("~/Applied Analystics SAS Prog/mymath475/assignment14Table.png")



#TF(term freq) = # times term t appears in a doc / total number of terms in doc  
TF1 = 0/13  
TF1

## [1] 0

TF2 = 1/13  
TF2

## [1] 0.07692308

TF3 = 0/12  
TF3

## [1] 0

TF4 = 1/9  
TF4

## [1] 0.1111111

* Calculate the idf of the term animals

#idf(t) = ln (total num of doc / num of docs with term t in it)  
  
idf = log(4/2)  
idf

## [1] 0.6931472

* Calculate the tf-idk of the term animals for each document

#tf-idf(t) = tf(t) x idf(t)  
  
tfidf1 = 0\*log(4/2)  
tfidf1

## [1] 0

tfidf2 = (1/13)\*log(4/2)  
tfidf2

## [1] 0.05331901

tfidf3 = 0\*log(4/2)  
tfidf3

## [1] 0

tfidf4 = (1/9)\*log(4/2)  
tfidf4

## [1] 0.07701635

Question 2. Reproduce the results the codes in the sample codes

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.3 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.5.0 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ 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(ggplot2)  
library(tidytext)  
df <- read\_csv("https://bryantstats.github.io/math475/assignments/amazon\_reviews.csv")

## Rows: 34660 Columns: 21  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (12): id, name, asins, brand, categories, keys, manufacturer, reviews.d...  
## dbl (3): reviews.id, reviews.numHelpful, reviews.rating  
## lgl (4): reviews.didPurchase, reviews.doRecommend, reviews.userCity, revie...  
## dttm (2): reviews.date, reviews.dateAdded  
##   
## 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.

head(df)

## # A tibble: 6 × 21  
## id name asins brand categories keys manufacturer reviews.date   
## <chr> <chr> <chr> <chr> <chr> <chr> <chr> <dttm>   
## 1 AVqkIhwDv… All-… B01A… Amaz… Electroni… 8416… Amazon 2017-01-13 00:00:00  
## 2 AVqkIhwDv… All-… B01A… Amaz… Electroni… 8416… Amazon 2017-01-13 00:00:00  
## 3 AVqkIhwDv… All-… B01A… Amaz… Electroni… 8416… Amazon 2017-01-13 00:00:00  
## 4 AVqkIhwDv… All-… B01A… Amaz… Electroni… 8416… Amazon 2017-01-13 00:00:00  
## 5 AVqkIhwDv… All-… B01A… Amaz… Electroni… 8416… Amazon 2017-01-12 00:00:00  
## 6 AVqkIhwDv… All-… B01A… Amaz… Electroni… 8416… Amazon 2017-01-12 00:00:00  
## # ℹ 13 more variables: reviews.dateAdded <dttm>, reviews.dateSeen <chr>,  
## # reviews.didPurchase <lgl>, reviews.doRecommend <lgl>, reviews.id <dbl>,  
## # reviews.numHelpful <dbl>, reviews.rating <dbl>, reviews.sourceURLs <chr>,  
## # reviews.text <chr>, reviews.title <chr>, reviews.userCity <lgl>,  
## # reviews.userProvince <lgl>, reviews.username <chr>

#define documents  
df %>%   
 group\_by(name) %>%   
 count(sort = TRUE)

## # A tibble: 49 × 2  
## # Groups: name [49]  
## name n  
## <chr> <int>  
## 1 "Fire Tablet, 7 Display, Wi-Fi, 8 GB - Includes Special Offers, Magent… 10966  
## 2 <NA> 6760  
## 3 "Echo (White),,,\nEcho (White),,," 3309  
## 4 "Amazon Kindle Paperwhite - eBook reader - 4 GB - 6 monochrome Paperwh… 3176  
## 5 "All-New Fire HD 8 Tablet, 8 HD Display, Wi-Fi, 16 GB - Includes Speci… 2814  
## 6 "Amazon Fire Tv,,,\nAmazon Fire Tv,,," 2527  
## 7 "Fire Kids Edition Tablet, 7 Display, Wi-Fi, 16 GB, Green Kid-Proof Ca… 1685  
## 8 "Brand New Amazon Kindle Fire 16gb 7 Ips Display Tablet Wifi 16 Gb Blu… 1038  
## 9 "Kindle Voyage E-reader, 6 High-Resolution Display (300 ppi) with Adap… 580  
## 10 "Fire Tablet, 7 Display, Wi-Fi, 8 GB - Includes Special Offers, Black" 372  
## # ℹ 39 more rows

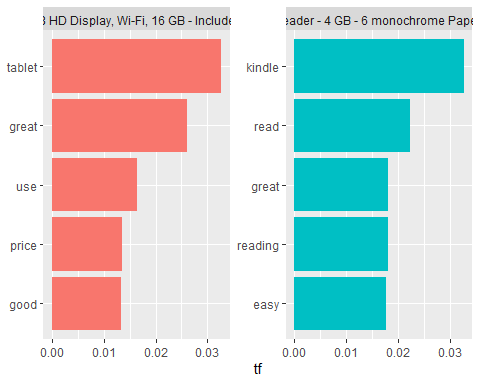
#filter to 2 items for analysis  
df = df %>%   
 select(name, reviews.text) %>%   
 filter(name %in% c('Amazon Kindle Paperwhite - eBook reader - 4 GB - 6 monochrome Paperwhite - touchscreen - Wi-Fi - black,,,',   
'All-New Fire HD 8 Tablet, 8 HD Display, Wi-Fi, 16 GB - Includes Special Offers, Magenta')) %>%   
 rename(texts = reviews.text,  
 document = name)  
  
#plot term frequency for each doc  
stop\_word2 = tibble(word = c(letters, LETTERS, "oh", 'just'))  
  
df\_words = df %>%   
 unnest\_tokens(input = texts, output = word) %>%   
 count(document, word, sort = TRUE) %>%   
 anti\_join(get\_stopwords()) %>%   
 anti\_join(stop\_word2)

## Joining with `by = join\_by(word)`  
## Joining with `by = join\_by(word)`

total\_words <- df\_words %>%   
 group\_by(document) %>%   
 summarize(total = sum(n))  
  
df\_words <- left\_join(df\_words, total\_words)

## Joining with `by = join\_by(document)`

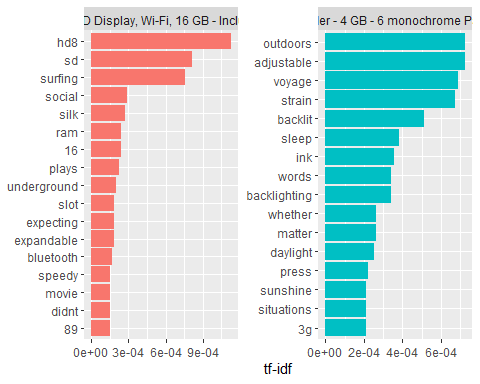
df\_tf = df\_words %>%   
 group\_by(document) %>%   
 mutate(tf = n/total)  
  
df\_tf %>%   
 group\_by(document) %>%   
 slice\_max(tf, n = 5) %>%   
 ungroup() %>%  
 ggplot(aes(tf, fct\_reorder(word, tf), fill = document)) +  
 geom\_col(show.legend = FALSE) +  
 facet\_wrap(~document, ncol = 2, scales = "free") +  
 labs(x = "tf", y = NULL)



#plot the tf-idf for each doc  
df\_tf\_idf <- df\_words %>%  
 bind\_tf\_idf(word, document, n)  
  
df\_tf\_idf

## # A tibble: 8,609 × 7  
## document word n total tf idf tf\_idf  
## <chr> <chr> <int> <int> <dbl> <dbl> <dbl>  
## 1 Amazon Kindle Paperwhite - eBook reade… kind… 1717 52532 0.0327 0 0  
## 2 All-New Fire HD 8 Tablet, 8 HD Display… tabl… 1342 41027 0.0327 0 0  
## 3 Amazon Kindle Paperwhite - eBook reade… read 1169 52532 0.0223 0 0  
## 4 All-New Fire HD 8 Tablet, 8 HD Display… great 1066 41027 0.0260 0 0  
## 5 Amazon Kindle Paperwhite - eBook reade… read… 950 52532 0.0181 0 0  
## 6 Amazon Kindle Paperwhite - eBook reade… great 947 52532 0.0180 0 0  
## 7 Amazon Kindle Paperwhite - eBook reade… easy 925 52532 0.0176 0 0  
## 8 Amazon Kindle Paperwhite - eBook reade… books 778 52532 0.0148 0 0  
## 9 Amazon Kindle Paperwhite - eBook reade… love 772 52532 0.0147 0 0  
## 10 Amazon Kindle Paperwhite - eBook reade… light 686 52532 0.0131 0 0  
## # ℹ 8,599 more rows

library(forcats)  
  
df\_tf\_idf %>%  
 group\_by(document) %>%  
 slice\_max(tf\_idf, n = 15) %>%  
 ungroup() %>%  
 ggplot(aes(tf\_idf, fct\_reorder(word, tf\_idf), fill = document)) +  
 geom\_col(show.legend = FALSE) +  
 facet\_wrap(~document, ncol = 2, scales = "free") +  
 labs(x = "tf-idf", y = NULL)



Question 3. Redo the sample codes with different selections of documents

library(tidyverse)  
library(ggplot2)  
library(tidytext)  
df <- read\_csv("https://bryantstats.github.io/math475/assignments/amazon\_reviews.csv")

## Rows: 34660 Columns: 21  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (12): id, name, asins, brand, categories, keys, manufacturer, reviews.d...  
## dbl (3): reviews.id, reviews.numHelpful, reviews.rating  
## lgl (4): reviews.didPurchase, reviews.doRecommend, reviews.userCity, revie...  
## dttm (2): reviews.date, reviews.dateAdded  
##   
## 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.

head(df)

## # A tibble: 6 × 21  
## id name asins brand categories keys manufacturer reviews.date   
## <chr> <chr> <chr> <chr> <chr> <chr> <chr> <dttm>   
## 1 AVqkIhwDv… All-… B01A… Amaz… Electroni… 8416… Amazon 2017-01-13 00:00:00  
## 2 AVqkIhwDv… All-… B01A… Amaz… Electroni… 8416… Amazon 2017-01-13 00:00:00  
## 3 AVqkIhwDv… All-… B01A… Amaz… Electroni… 8416… Amazon 2017-01-13 00:00:00  
## 4 AVqkIhwDv… All-… B01A… Amaz… Electroni… 8416… Amazon 2017-01-13 00:00:00  
## 5 AVqkIhwDv… All-… B01A… Amaz… Electroni… 8416… Amazon 2017-01-12 00:00:00  
## 6 AVqkIhwDv… All-… B01A… Amaz… Electroni… 8416… Amazon 2017-01-12 00:00:00  
## # ℹ 13 more variables: reviews.dateAdded <dttm>, reviews.dateSeen <chr>,  
## # reviews.didPurchase <lgl>, reviews.doRecommend <lgl>, reviews.id <dbl>,  
## # reviews.numHelpful <dbl>, reviews.rating <dbl>, reviews.sourceURLs <chr>,  
## # reviews.text <chr>, reviews.title <chr>, reviews.userCity <lgl>,  
## # reviews.userProvince <lgl>, reviews.username <chr>

#define documents  
df %>%   
 group\_by(name) %>%   
 count(sort = TRUE)

## # A tibble: 49 × 2  
## # Groups: name [49]  
## name n  
## <chr> <int>  
## 1 "Fire Tablet, 7 Display, Wi-Fi, 8 GB - Includes Special Offers, Magent… 10966  
## 2 <NA> 6760  
## 3 "Echo (White),,,\nEcho (White),,," 3309  
## 4 "Amazon Kindle Paperwhite - eBook reader - 4 GB - 6 monochrome Paperwh… 3176  
## 5 "All-New Fire HD 8 Tablet, 8 HD Display, Wi-Fi, 16 GB - Includes Speci… 2814  
## 6 "Amazon Fire Tv,,,\nAmazon Fire Tv,,," 2527  
## 7 "Fire Kids Edition Tablet, 7 Display, Wi-Fi, 16 GB, Green Kid-Proof Ca… 1685  
## 8 "Brand New Amazon Kindle Fire 16gb 7 Ips Display Tablet Wifi 16 Gb Blu… 1038  
## 9 "Kindle Voyage E-reader, 6 High-Resolution Display (300 ppi) with Adap… 580  
## 10 "Fire Tablet, 7 Display, Wi-Fi, 8 GB - Includes Special Offers, Black" 372  
## # ℹ 39 more rows

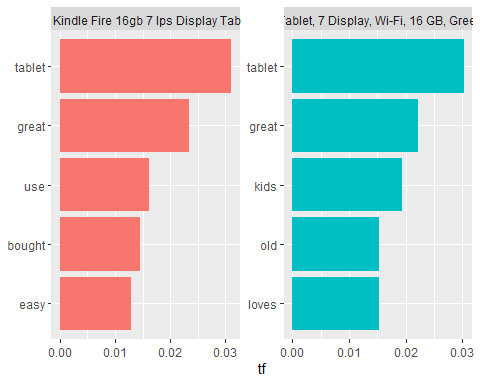
#filter to 2 items for analysis  
df = df %>%   
 select(name, reviews.text) %>%   
 filter(name %in% c('Fire Kids Edition Tablet, 7 Display, Wi-Fi, 16 GB, Green Kid-Proof Case',   
'Brand New Amazon Kindle Fire 16gb 7 Ips Display Tablet Wifi 16 Gb Blue,,,')) %>%   
 rename(texts = reviews.text,  
 document = name)  
  
#plot term frequency for each doc  
stop\_word2 = tibble(word = c(letters, LETTERS, "oh", 'just'))  
  
df\_words = df %>%   
 unnest\_tokens(input = texts, output = word) %>%   
 count(document, word, sort = TRUE) %>%   
 anti\_join(get\_stopwords()) %>%   
 anti\_join(stop\_word2)

## Joining with `by = join\_by(word)`  
## Joining with `by = join\_by(word)`

total\_words <- df\_words %>%   
 group\_by(document) %>%   
 summarize(total = sum(n))  
  
df\_words <- left\_join(df\_words, total\_words)

## Joining with `by = join\_by(document)`

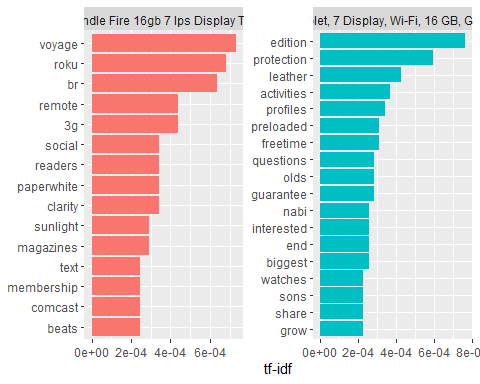
df\_tf = df\_words %>%   
 group\_by(document) %>%   
 mutate(tf = n/total)  
  
df\_tf %>%   
 group\_by(document) %>%   
 slice\_max(tf, n = 5) %>%   
 ungroup() %>%  
 ggplot(aes(tf, fct\_reorder(word, tf), fill = document)) +  
 geom\_col(show.legend = FALSE) +  
 facet\_wrap(~document, ncol = 2, scales = "free") +  
 labs(x = "tf", y = NULL)



#plot the tf-idf for each doc  
df\_tf\_idf <- df\_words %>%  
 bind\_tf\_idf(word, document, n)  
  
df\_tf\_idf

## # A tibble: 5,363 × 7  
## document word n total tf idf tf\_idf  
## <chr> <chr> <int> <int> <dbl> <dbl> <dbl>  
## 1 Fire Kids Edition Tablet, 7 Display, W… tabl… 742 24553 0.0302 0 0  
## 2 Fire Kids Edition Tablet, 7 Display, W… great 543 24553 0.0221 0 0  
## 3 Fire Kids Edition Tablet, 7 Display, W… kids 477 24553 0.0194 0 0  
## 4 Brand New Amazon Kindle Fire 16gb 7 Ip… tabl… 441 14181 0.0311 0 0  
## 5 Fire Kids Edition Tablet, 7 Display, W… old 377 24553 0.0154 0 0  
## 6 Fire Kids Edition Tablet, 7 Display, W… loves 376 24553 0.0153 0 0  
## 7 Fire Kids Edition Tablet, 7 Display, W… year 372 24553 0.0152 0 0  
## 8 Fire Kids Edition Tablet, 7 Display, W… love 341 24553 0.0139 0 0  
## 9 Brand New Amazon Kindle Fire 16gb 7 Ip… great 331 14181 0.0233 0 0  
## 10 Fire Kids Edition Tablet, 7 Display, W… boug… 306 24553 0.0125 0 0  
## # ℹ 5,353 more rows

library(forcats)  
  
df\_tf\_idf %>%  
 group\_by(document) %>%  
 slice\_max(tf\_idf, n = 15) %>%  
 ungroup() %>%  
 ggplot(aes(tf\_idf, fct\_reorder(word, tf\_idf), fill = document)) +  
 geom\_col(show.legend = FALSE) +  
 facet\_wrap(~document, ncol = 2, scales = "free") +  
 labs(x = "tf-idf", y = NULL)



Question 4. Do the follows:

* Define your own documents for the analysis.
* Plot the term frequency in the documents
* Plot the tf-idf in the documents.

#dataset  
library(tidyverse)  
library(ggplot2)  
library(tidytext)  
df <- read\_csv("https://bryantstats.github.io/math475/assignments/netflix\_titles.csv")

## Rows: 7787 Columns: 12  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (11): show\_id, type, title, director, cast, country, date\_added, rating,...  
## dbl (1): release\_year  
##   
## 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.

head(df)

## # A tibble: 6 × 12  
## show\_id type title director cast country date\_added release\_year rating  
## <chr> <chr> <chr> <chr> <chr> <chr> <chr> <dbl> <chr>   
## 1 s1 TV Show 3% <NA> João… Brazil August 14… 2020 TV-MA   
## 2 s2 Movie 7:19 Jorge Mich… Demi… Mexico December … 2016 TV-MA   
## 3 s3 Movie 23:59 Gilbert Ch… Tedd… Singap… December … 2011 R   
## 4 s4 Movie 9 Shane Acker Elij… United… November … 2009 PG-13   
## 5 s5 Movie 21 Robert Luk… Jim … United… January 1… 2008 PG-13   
## 6 s6 TV Show 46 Serdar Akar Erda… Turkey July 1, 2… 2016 TV-MA   
## # ℹ 3 more variables: duration <chr>, listed\_in <chr>, description <chr>

#define documents  
df %>%   
 group\_by(title) %>%   
 count(sort = TRUE)

## # A tibble: 7,787 × 2  
## # Groups: title [7,787]  
## title n  
## <chr> <int>  
## 1 #Alive 1  
## 2 #AnneFrank - Parallel Stories 1  
## 3 #FriendButMarried 1  
## 4 #FriendButMarried 2 1  
## 5 #Roxy 1  
## 6 #Rucker50 1  
## 7 #Selfie 1  
## 8 #Selfie 69 1  
## 9 #blackAF 1  
## 10 #cats\_the\_mewvie 1  
## # ℹ 7,777 more rows

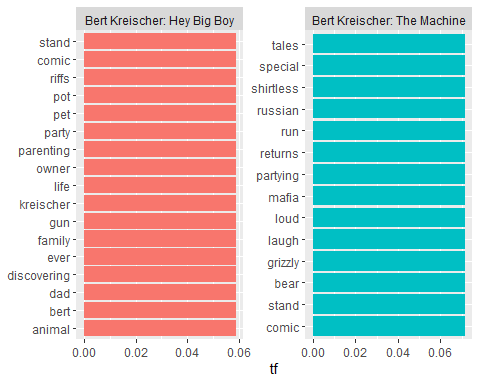
df = df %>%   
 select(title, description) %>%   
 filter(title %in% c('Bert Kreischer: Hey Big Boy',   
'Bert Kreischer: The Machine')) %>%   
 rename(texts = description,  
 document = title)  
  
#plot term frequency  
stop\_word2 = tibble(word = c(letters, LETTERS, "oh", 'just'))  
  
df\_words = df %>%   
 unnest\_tokens(input = texts, output = word) %>%   
 count(document, word, sort = TRUE) %>%   
 anti\_join(get\_stopwords()) %>%   
 anti\_join(stop\_word2)

## Joining with `by = join\_by(word)`  
## Joining with `by = join\_by(word)`

total\_words <- df\_words %>%   
 group\_by(document) %>%   
 summarize(total = sum(n))  
  
df\_words <- left\_join(df\_words, total\_words)

## Joining with `by = join\_by(document)`

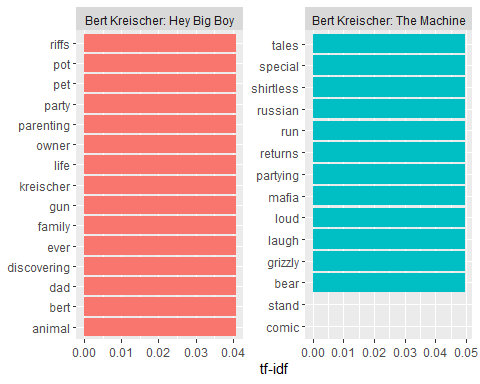
df\_tf = df\_words %>%   
 group\_by(document) %>%   
 mutate(tf = n/total)  
  
df\_tf %>%   
 group\_by(document) %>%   
 slice\_max(tf, n = 5) %>%   
 ungroup() %>%  
 ggplot(aes(tf, fct\_reorder(word, tf), fill = document)) +  
 geom\_col(show.legend = FALSE) +  
 facet\_wrap(~document, ncol = 2, scales = "free") +  
 labs(x = "tf", y = NULL)



#plot tf-idf for each doc  
df\_tf\_idf <- df\_words %>%  
 bind\_tf\_idf(word, document, n)  
  
df\_tf\_idf

## # A tibble: 31 × 7  
## document word n total tf idf tf\_idf  
## <chr> <chr> <int> <int> <dbl> <dbl> <dbl>  
## 1 Bert Kreischer: Hey Big Boy animal 1 17 0.0588 0.693 0.0408  
## 2 Bert Kreischer: Hey Big Boy bert 1 17 0.0588 0.693 0.0408  
## 3 Bert Kreischer: Hey Big Boy comic 1 17 0.0588 0 0   
## 4 Bert Kreischer: Hey Big Boy dad 1 17 0.0588 0.693 0.0408  
## 5 Bert Kreischer: Hey Big Boy discovering 1 17 0.0588 0.693 0.0408  
## 6 Bert Kreischer: Hey Big Boy ever 1 17 0.0588 0.693 0.0408  
## 7 Bert Kreischer: Hey Big Boy family 1 17 0.0588 0.693 0.0408  
## 8 Bert Kreischer: Hey Big Boy gun 1 17 0.0588 0.693 0.0408  
## 9 Bert Kreischer: Hey Big Boy kreischer 1 17 0.0588 0.693 0.0408  
## 10 Bert Kreischer: Hey Big Boy life 1 17 0.0588 0.693 0.0408  
## # ℹ 21 more rows

library(forcats)  
  
df\_tf\_idf %>%  
 group\_by(document) %>%  
 slice\_max(tf\_idf, n = 15) %>%  
 ungroup() %>%  
 ggplot(aes(tf\_idf, fct\_reorder(word, tf\_idf), fill = document)) +  
 geom\_col(show.legend = FALSE) +  
 facet\_wrap(~document, ncol = 2, scales = "free") +  
 labs(x = "tf-idf", y = NULL)



Question 5. Redo Question 4 on your own dataset.

#dataset  
library(tidyverse)  
library(ggplot2)  
library(tidytext)  
df <- read\_csv("~/Applied Analystics SAS Prog/mymath475/CNNtext.csv")

## Rows: 11490 Columns: 3  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (3): id, article, highlights  
##   
## 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.

head(df)

## # A tibble: 6 × 3  
## id article highlights  
## <chr> <chr> <chr>   
## 1 92c514c913c0bdfe25341af9fd72b29db544099b Ever noticed how plane se… "Experts …  
## 2 2003841c7dc0e7c5b1a248f9cd536d727f27a45a A drunk teenage boy had t… "Drunk te…  
## 3 91b7d2311527f5c2b63a65ca98d21d9c92485149 Dougie Freedman is on the… "Nottingh…  
## 4 caabf9cbdf96eb1410295a673e953d304391bfbb Liverpool target Neto is … "Fiorenti…  
## 5 3da746a7d9afcaa659088c8366ef6347fe6b53ea Bruce Jenner will break h… "Tell-all…  
## 6 5ed5e3fbd235a8046cd3b87f4a1aa51b856c8ec3 This is the moment that a… "Giant pi…

#define documents  
df %>%   
 group\_by(id) %>%   
 count(sort = TRUE)

## # A tibble: 11,490 × 2  
## # Groups: id [11,490]  
## id n  
## <chr> <int>  
## 1 000571afe702684d90c1d222ce70b1e1375c1016 1  
## 2 000642916e3a6c33411c617cf2f3c134a206fba8 1  
## 3 00110802bc6eae0e8e4d3d22e27f458f41be2b22 1  
## 4 00119229166ae09a6ef25c0e10b101ef9eb9cca3 1  
## 5 0013aa16650fbcfbe6edb16ac614ad174cb5d1cf 1  
## 6 00180b7ce54794a52766d795506a94071f7c055b 1  
## 7 001ebaa80dca4a65adf2178b132113cb9e3d5431 1  
## 8 00200e794fa41d3f7ce92cbf43e9fd4cd652bb09 1  
## 9 00217448b38d81a23db66ac362bee25056f58fab 1  
## 10 0021fe8d65bd0d6d76d5fefba2ac02f0c48a43f4 1  
## # ℹ 11,480 more rows

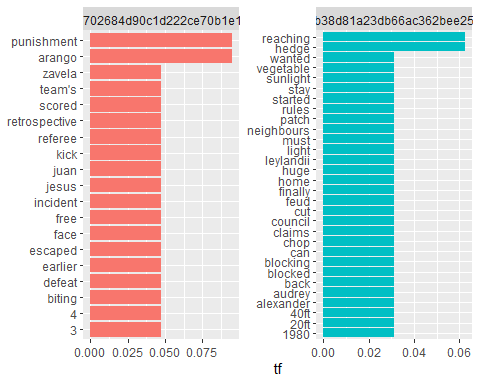
df = df %>%   
 select(id, highlights) %>%   
 filter(id %in% c('00217448b38d81a23db66ac362bee25056f58fab',   
'000571afe702684d90c1d222ce70b1e1375c1016')) %>%   
 rename(texts = highlights,  
 document = id)  
  
#plot term frequency  
stop\_word2 = tibble(word = c(letters, LETTERS, "oh", 'just'))  
  
df\_words = df %>%   
 unnest\_tokens(input = texts, output = word) %>%   
 count(document, word, sort = TRUE) %>%   
 anti\_join(get\_stopwords()) %>%   
 anti\_join(stop\_word2)

## Joining with `by = join\_by(word)`  
## Joining with `by = join\_by(word)`

total\_words <- df\_words %>%   
 group\_by(document) %>%   
 summarize(total = sum(n))  
  
df\_words <- left\_join(df\_words, total\_words)

## Joining with `by = join\_by(document)`

df\_tf = df\_words %>%   
 group\_by(document) %>%   
 mutate(tf = n/total)  
  
df\_tf %>%   
 group\_by(document) %>%   
 slice\_max(tf, n = 5) %>%   
 ungroup() %>%  
 ggplot(aes(tf, fct\_reorder(word, tf), fill = document)) +  
 geom\_col(show.legend = FALSE) +  
 facet\_wrap(~document, ncol = 2, scales = "free") +  
 labs(x = "tf", y = NULL)



#plot tf-idf for each doc  
df\_tf\_idf <- df\_words %>%  
 bind\_tf\_idf(word, document, n)  
  
df\_tf\_idf

## # A tibble: 49 × 7  
## document word n total tf idf tf\_idf  
## <chr> <chr> <int> <int> <dbl> <dbl> <dbl>  
## 1 000571afe702684d90c1d222ce70b1e1375c10… aran… 2 21 0.0952 0.693 0.0660  
## 2 000571afe702684d90c1d222ce70b1e1375c10… puni… 2 21 0.0952 0.693 0.0660  
## 3 00217448b38d81a23db66ac362bee25056f58f… hedge 2 32 0.0625 0.693 0.0433  
## 4 00217448b38d81a23db66ac362bee25056f58f… reac… 2 32 0.0625 0.693 0.0433  
## 5 000571afe702684d90c1d222ce70b1e1375c10… 3 1 21 0.0476 0.693 0.0330  
## 6 000571afe702684d90c1d222ce70b1e1375c10… 4 1 21 0.0476 0.693 0.0330  
## 7 000571afe702684d90c1d222ce70b1e1375c10… biti… 1 21 0.0476 0.693 0.0330  
## 8 000571afe702684d90c1d222ce70b1e1375c10… defe… 1 21 0.0476 0.693 0.0330  
## 9 000571afe702684d90c1d222ce70b1e1375c10… earl… 1 21 0.0476 0.693 0.0330  
## 10 000571afe702684d90c1d222ce70b1e1375c10… esca… 1 21 0.0476 0.693 0.0330  
## # ℹ 39 more rows

library(forcats)  
  
df\_tf\_idf %>%  
 group\_by(document) %>%  
 slice\_max(tf\_idf, n = 15) %>%  
 ungroup() %>%  
 ggplot(aes(tf\_idf, fct\_reorder(word, tf\_idf), fill = document)) +  
 geom\_col(show.legend = FALSE) +  
 facet\_wrap(~document, ncol = 2, scales = "free") +  
 labs(x = "tf-idf", y = NULL)

