STA 235 - Natural Language Processing & Twitter Data

Spring 2021

McCombs School of Business, UT Austin

Some reminders

Prediction Project due this Friday 12.00pm (noon)

Please make sure you upload the right files (.pdf and .R)

Late submissions (including wrong files) will be penalized

Some reminders (Cont.)

Materials for review will be uploaded this week

Office hours are Monday and Thursdays: If needed, I will add more slots

Some reminders (Cont.)

Any re-grading will only be accepted until Friday

Check your assignments, JITTs, etc.

JITT make-up will be included after today's class (based on in-class participation)

Some reminders (Cont.)

If you a conflict with the final exam, let me know today!

You will only be allowed to take the make-up exam if you have my authorization

Tips for de-bugging

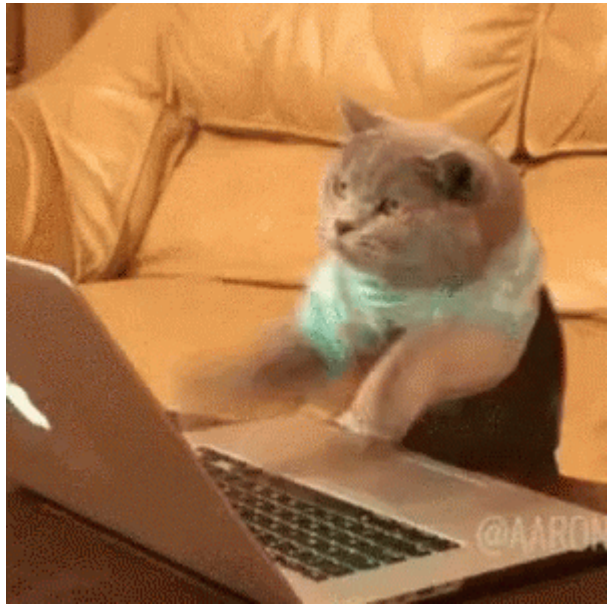
1) Check the **error message** (e.g. is it an error or warning?, is it giving you hints on what the problem is?)

2) Check your **code** (e.g. typos, arguments, type of data, wrong package?)

3) **Google** your error (e.g. name of package + error message)

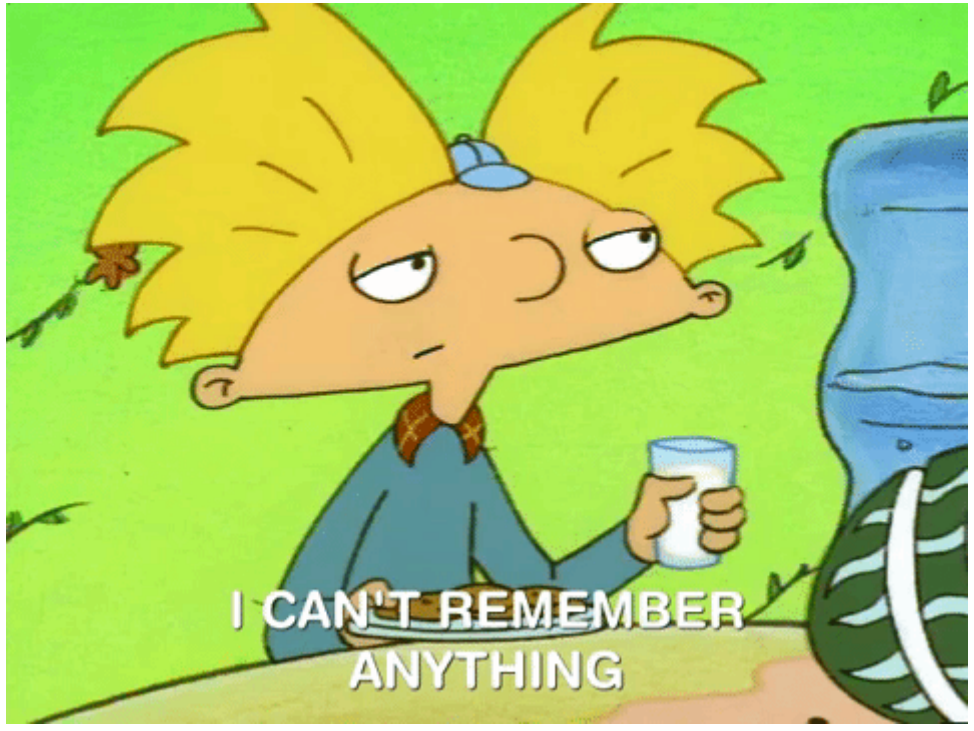
4) **Contact the instructor team** (i.e. Piazza or email)

- Please post code + error message (privately)



Questions?

Last class



Finished with prediction methods:

- Bagging: Combine trees from bootstrap samples to reduce variance.
- Random Forests: Bagging + random covariate selection (de-correlate trees)
- Boosting: Sequential trees that learn from the previous one.

Today

1) **Natural Language Processing**:

- Amazon reviews
- Twitter data

2) **Course Instructor Survey**

3) **Wrapping things up**

- Example combining causal inference + prediction.

Let's get texting!

Not everything is numbers

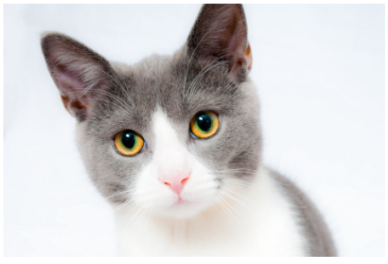
- So far, we have only seen **numeric** or **categorical** data both as predictors and outcomes.
- But what about other types of data?

Images

isthisacat.com

IS THIS A CAT?

Upload an image: Choose File No file chosen



YES

Not everything is numbers

- So far, we have only seen **numeric** or **categorical** data both as predictors and outcomes.
- But what about other types of data?

Images

Sounds

Song similarity using dynamic time warping

Posted on September 12, 2016

Here I show how to use the `dfDTW` function in warbleR to compare acoustics signals using dynamic time warping (DTW).

First load these packages (if not installed the code will install it):

```
x<-c("vegan", "warbleR")

A <- lapply(x, function(y) {
  if(!y %in% installed.packages()[,"Package"])  install.packages(y)
require(y, character.only = T)
  })
```

and load example data from `warbleR`

```
# optional, save it in a temporal folder
# setwd(tempdir())

data(list = c( "Phae.long1", "Phae.long2","Phae.long3", "Phae.long4","selec.table"))

writeWave(Phae.long1, "Phae.long1.wav")
writeWave(Phae.long2, "Phae.long2.wav")
writeWave(Phae.long3, "Phae.long3.wav")
writeWave(Phae.long4, "Phae.long4.wav")
```

These recordings all come from long-billed hermits with different song types.

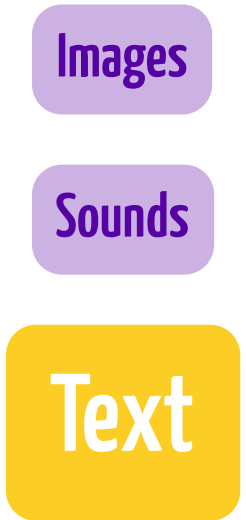
We can run the DTW analysis to compare these time series usin the warbleR function `dfDTW` which calculates the dominant frequency contours of each sgnals and compares using dynamic time warping. Internally it applies the `dtwDist` function from the dtw package.

```
dm <- dfDTW(selec.table, length.out = 30, flim = c(1, 12), bp = c(2, 9), wl = 300, ir
```

Let's see if the dissimilarity from dtw represents the acoutic differences. First we need a binary matrix representing same recording with 0s, and different recording with 1s. The following functions does exactly that:

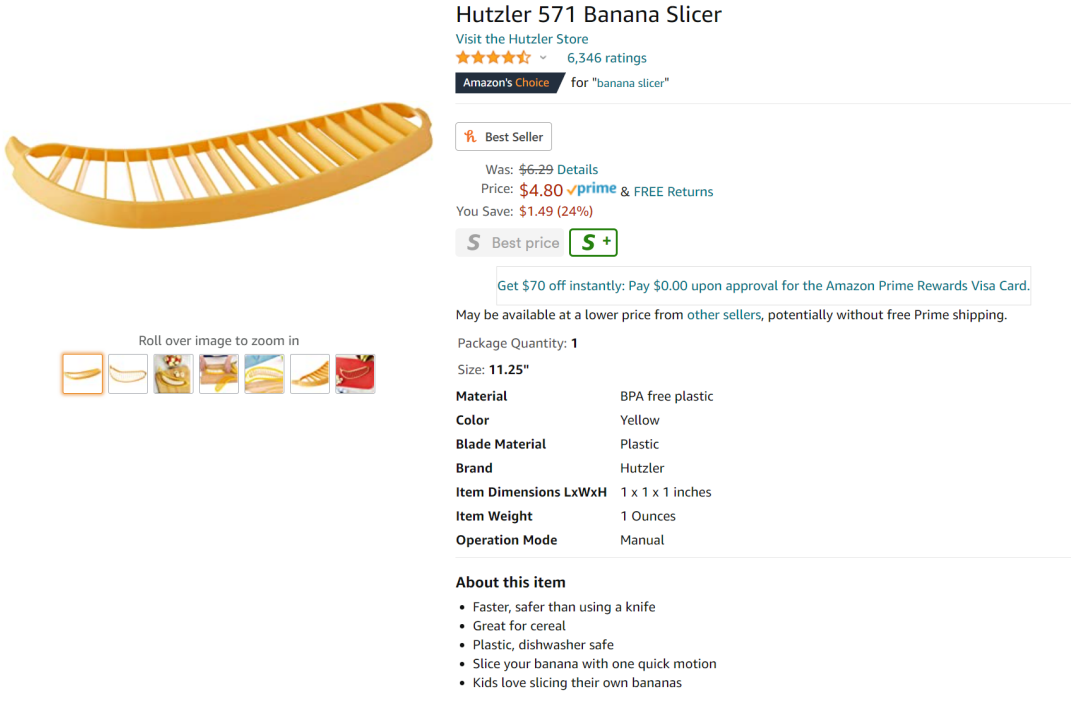
Not everything is numbers

- So far, we have only seen **numeric** or **categorical** data both as predictors and outcomes.
- But what about other types of data?



What is Natural Language Processing?

- NLP focuses on analyzing and understanding text systematically **through code**.
- E.g.: If you were a seller on Amazon who wanted to capture feedback through reviews, how would you do it?



Roll over image to zoom in

Hutzler 571 Banana Slicer

Visit the Hutzler Store

4.5 out of 5 stars 6,346 ratings

Amazon's Choice for "banana slicer"

Best Seller

Was: $6.29 Details

Price: $4.80 prime & FREE Returns

You Save: $1.49 (24%)

Best price S+

Get $70 off instantly: Pay $0.00 upon approval for the Amazon Prime Rewards Visa Card.

May be available at a lower price from other sellers, potentially without free Prime shipping.

Package Quantity: 1

Size: 11.25"

Material	BPA free plastic
Color	Yellow
Blade Material	Plastic
Brand	Hutzler
Item Dimensions LxWxH	1 x 1 x 1 inches
Item Weight	1 Ounces
Operation Mode	Manual

About this item

- Faster, safer than using a knife
- Great for cereal
- Plastic, dishwasher safe
- Slice your banana with one quick motion
- Kids love slicing their own bananas

What is Natural Language Processing?

Customer reviews

4.5 out of 5

6,346 global ratings

5 star 73%
4 star 13%
3 star 7%
2 star 4%
1 star 3%

How are ratings calculated?

By feature

Battery life 5.0
Easy to use 4.6
Easy to clean 4.5

See more

Review this product

Share your thoughts with other customers

Write a customer review

Reviews with images




See all customer images

Read reviews that mention

slice bananas | banana slices | banana slicers | changed my life
peanut butter | perfectly sliced | grocery store | needless to say
cutting board | left handed | every time | every day | ice cream

Top reviews

Top reviews from the United States

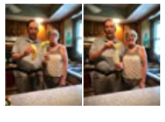
Emily S

Be careful of wrong-way bananas!!

Reviewed in the United States on June 1, 2018

Package Quantity: 1 | Size: 11.25" | Verified Purchase

We were so excited to get our Hutzler 571...until we realized that our bananas curved the wrong way. Gonna have to go to the store for new bananas...🍌🍌🍌



1,458 people found this helpful

Human vs. Machine

Most of the times, **machines are great at things that humans are not (and vice-versa)**

Human vs. Machine

Most of the times, **machines are great at things that humans are not (and vice-versa)**

- Humans are great at understanding **context**

Human vs. Machine

Most of the times, **machines are great at things that humans are not (and vice-versa)**

- Humans are great at understanding **context**
- Machines are great at **detecting patterns**

Let's put this to use!

```
library(tidyverse)
library(rvest)

scrape_amazon <- function(ASIN, page_num){ #Write our own function (source: https://martinctc.github.io/blog/vig

  url_reviews <- paste0("https://www.amazon.com/product-reviews/",ASIN,"/?pageNumber=",page_num)

  doc <- read_html(url_reviews) # Assign results to `doc`

  # Review Title
  doc %>%
    html_nodes("[class='a-size-base a-link-normal review-title a-color-base review-title-content a-text-bold']")
    html_text() -> review_title

  # Review Text
  doc %>%
    html_nodes("[class='a-size-base review-text review-text-content']") %>%
    html_text() -> review_text

  # Number of stars in review
  doc %>%
    html_nodes("[data-hook='review-star-rating']") %>%
    html_text() -> review_star

  # Return a tibble
  tibble(review_title, review_text, review_star, page = page_num) %>% return()
}
```

Let's put this to use!

```
library(tidyverse)
library(rvest)

scrape_amazon <- function(ASIN, page_num){ #Write our own function (source: https://martinctc.github.io/blog/vig

  url_reviews <- paste0("https://www.amazon.com/product-reviews/",ASIN,"/?pageNumber=",page_num)

  doc <- read_html(url_reviews) # Assign results to `doc`

  # Review Title
  doc %>%
    html_nodes("[class='a-size-base a-link-normal review-title a-color-base review-title-content a-text-bold']")
    html_text() -> review_title

  # Review Text
  doc %>%
    html_nodes("[class='a-size-base review-text review-text-content']") %>%
    html_text() -> review_text

  # Number of stars in review
  doc %>%
    html_nodes("[data-hook='review-star-rating']") %>%
    html_text() -> review_star

  # Return a tibble
  tibble(review_title, review_text, review_star, page = page_num) %>% return()
}
```

Amazon.com: Customer reviews:

amazon.com/product-reviews/B0047E0EII/?pageNumber=1

Customer reviews

4.5 out of 5

6,346 global ratings

Hutzler 571 Banana Slicer

by Hutzler

Package Quantity: 1 | Size: 11.25" | Change

5 star 73%
4 star 13%
3 star 7%
2 star 4%
1 star 3%

Write a review

How are ratings calculated?

See All Buying Options

Add to Wish List

Top positive review

All positive reviews ›

Emily S

Be careful of wrong-way bananas!!

Reviewed in the United States on June 1, 2018

We were so excited to get our Hutzler 571...until we realized that our bananas curved the wrong way. Gonna have to go to the store for new bananas...🍌🍌🍌

1,458 people found this helpful

Top critical review

All critical reviews ›

L. Wurts

Perfect, If You Want To Get Fired

Reviewed in the United States on November 5, 2017

I was sitting on the couch and my doorbell rang. I leapt off the sofa and ran to the door screaming, “My banana slicer!” I opened the package and immediately snatched a banana to slice. Without instructions included I did not realize I had to peel the banana first. It was a gooey mess and I had to grab another. This time I peeled it, but my banana was too small. It didn’t fill the whole slicer.
I went to Walmart Customer Service, “Do you have any giant

Read more

1,409 people found this helpful

Questions? Get fast answers from reviewers

Let's put this to use!

```
library(tidyverse)
library(rvest)

scrape_amazon <- function(ASIN, page_num){ #Write our own function (source: https://martinctc.github.io/blog/vig

  url_reviews <- paste0("https://www.amazon.com/product-reviews/",ASIN,"/?pageNumber=",page_num)

  doc <- read_html(url_reviews) # Assign results to `doc`

  # Review Title
  doc %>%
    html_nodes("[class='a-size-base a-link-normal review-title a-color-base review-title-content a-text-bold']")
    html_text() -> review_title

  # Review Text
  doc %>%
    html_nodes("[class='a-size-base review-text review-text-content']") %>%
    html_text() -> review_text

  # Number of stars in review
  doc %>%
    html_nodes("[data-hook='review-star-rating']") %>%
    html_text() -> review_star

  # Return a tibble
  tibble(review_title, review_text, review_star, page = page_num) %>% return()
}
```

Let's put this to use!

```
library(tidyverse)
library(rvest)

scrape_amazon <- function(ASIN, page_num){ #Write our own function (source: https://martinctc.github.io/blog/vig

  url_reviews <- paste0("https://www.amazon.com/product-reviews/",ASIN,"/?pageNumber=",page_num)

  doc <- read_html(url_reviews) # Assign results to `doc`

  # Review Title
  doc %>%
    html_nodes("[class='a-size-base a-link-normal review-title a-color-base review-title-content a-text-bold']")
    html_text() -> review_title

  # Review Text
  doc %>%
    html_nodes("[class='a-size-base review-text review-text-content']") %>%
    html_text() -> review_text

  # Number of stars in review
  doc %>%
    html_nodes("[data-hook='review-star-rating']") %>%
    html_text() -> review_star

  # Return a tibble
  tibble(review_title, review_text, review_star, page = page_num) %>% return()
}
```

Amazon.com: Customer reviews:

amazon.com/product-reviews/B0047E0EII/?pageNumber=1

Search customer reviews | Search

SORT BY: Top reviews | FILTER BY: All reviewers | All stars | All formats | Text, image, video

6,346 global ratings | 5,780 global reviews

From the United States

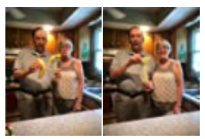
Emily S

★★★★★ **Be careful of wrong-way bananas!!**

Reviewed in the United States on June 1, 2018

Package Quantity: 1 | Size: 11.25" | Verified Purchase

We were so excited to get our Hutzler 571...until we realized that our banan store for new bananas... 🍌🍌🍌



1,458 people found this helpful

Helpful | Report abuse

L. Wurts

★★☆☆☆ **Perfect, If You Want To Get Fired**

Reviewed in the United States on November 5, 2017

Package Quantity: 1 | Size: 11.25" | Verified Purchase

I was sitting on the couch and my doorbell rang. I leapt off the sofa and ran to the door screaming, “My banana slicer!” I opened the package and immediately snatched a banana to slice. Without instructions included I did not realize I had to peel the banana first. It was a gooey mess and I had to grab another. This time I peeled it, but my banana was too small. It didn’t fill the whole slicer.

I went to Walmart Customer Service, “Do you have any giant bananas?” I questioned. The attendant turned away. I think he was laughing. He called for another attendant, they went to the back, and brought out the biggest bananas I’ve ever seen.

I went home with the bananas. I peeled the bananas and used the banana slicer. It was so satisfying to cut the bananas. I did it all day and forgot to go

See all 816 answered questions | Ask

Need customer service? Click here

Amazon.com: Customer reviews:

amazon.com/product-reviews/B0047E0EII/?pageNumber=1

Search customer reviews | Search

SORT BY | FILTER BY

Top rev… | All reviewers | All stars | All formats | Text, image, …

6,346 global ratings | 5,780 global reviews

From the United States

a.a-size-base.a-link-normal.review-title.a-color-base.review-title-content.a-tex... 231.35 × 17.78

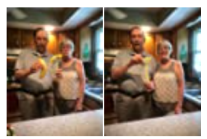
Emily S

★★★★★ **Be careful of wrong-way bananas!!**

Reviewed in the United States on June 1, 2018

Package Quantity: 1 | Size: 11.25" | Verified Purchase

We were so excited to get our Hutzler 571...until we realized that our bananas curved the wrong wa
Gonna have to go to the store for new bananas...💪💪💪



1,458 people found this helpful

Helpful | Report abuse

L. Wurts

★★☆☆☆ **Perfect, If You Want To Get Fired**

Reviewed in the United States on November 5, 2017

Package Quantity: 1 | Size: 11.25" | Verified Purchase

I was sitting on the couch and my doorbell rang. I leapt off the sofa and ran to the door screaming, "My banana slicer!" I opened the package and immediately snatched a banana to slice. Without instructions included I did not realize I had to peel the banana first. It was a gooey mess and I had t grab another. This time I peeled it, but my banana was too small. It didn't fill the whole slicer.
I went to Walmart Customer Service, "Do you have any giant bananas?" I questioned. The attendan

Elements | Console | Sources | Network | 14

```
before
iv id="customer_review-R1YUGEIEULVYES" class="a-section celwidget" data-csa-c-id
data-cel-widget="customer_review-R1YUGEIEULVYES">
<div data-hook="genome-widget" class="a-row a-spacing-mini">…</div>
<div class="a-row"> == $0
  ::before
▶<a class="a-link-normal" title="5.0 out of 5 stars" href="/gp/customer-reviews
p_d_rvw_ttl?ie=UTF8&ASIN=B0047E0EII">…</a>
  <span class="a-letter-space"></span>
▼<a data-hook="review-title" class="a-size-base a-link-normal review-title a-co
t a-text-bold" href="/gp/customer-reviews/R1YUGEIEULVYES/ref=cm_cr_arp_d_rvw_tt
    <span>Be careful of wrong-way bananas!!</span>
  </a>
  ::after
</div>
<span data-hook="review-date" class="a-size-base a-color-secondary review-date">
States on June 1, 2018</span>
<div class="a-row a-spacing-mini review-data review-format-strip">…</div>
<div class="a-row a-spacing-small review-data">…</div>
<div class="a-popover-preload" id="a-popover-R1YUGEIEULVYES_gallerySection_main"
```

… .a-spacing-none | div#customer_review-R1YUGEIEULVYES.a-section.celwidget | div.a-row …

Styles | Computed | Layout | Event Listeners | DOM Breakpoints | Properties

Filter :hov .cls

Console | What's New

Highlights from the Chrome 90 update

New CSS Flexbox debugging tools
Debug and inspect CSS Flexbox with the new CSS Flexbox debugging tools.

New Core Web Vitals overlay

Let's put this to use!

```
library(tidyverse)
library(rvest)

scrape_amazon <- function(ASIN, page_num){ #Write our own function (source: https://martinctc.github.io/blog/vig

  url_reviews <- paste0("https://www.amazon.com/product-reviews/",ASIN,"/?pageNumber=",page_num)

  doc <- read_html(url_reviews) # Assign results to `doc`

  # Review Title
  doc %>%
    html_nodes("[class='a-size-base a-link-normal review-title a-color-base review-title-content a-text-bold']")
    html_text() -> review_title

  # Review Text
  doc %>%
    html_nodes("[class='a-size-base review-text review-text-content']") %>%
    html_text() -> review_text

  # Number of stars in review
  doc %>%
    html_nodes("[data-hook='review-star-rating']") %>%
    html_text() -> review_star

  # Return a tibble
  tibble(review_title, review_text, review_star, page = page_num) %>% return()
}
```

Let's put this to use!

```
library(tidyverse)
library(rvest)

scrape_amazon <- function(ASIN, page_num){ #Write our own function (source: https://martinctc.github.io/blog/vig

  url_reviews <- paste0("https://www.amazon.com/product-reviews/",ASIN,"/?pageNumber=",page_num)

  doc <- read_html(url_reviews) # Assign results to `doc`

  # Review Title
  doc %>%
    html_nodes("[class='a-size-base a-link-normal review-title a-color-base review-title-content a-text-bold']")
    html_text() -> review_title

  # Review Text
  doc %>%
    html_nodes("[class='a-size-base review-text review-text-content']") %>%
    html_text() -> review_text

  # Number of stars in review
  doc %>%
    html_nodes("[data-hook='review-star-rating']") %>%
    html_text() -> review_star

  # Return a tibble
  tibble(review_title, review_text, review_star, page = page_num) %>% return()
}
```

What about our banana slicer?

```
ASIN_banana <- "B0047E0EII"

banana <- scrape_amazon(ASIN = ASIN_banana, page_num = 1)

banana %>% head()
```

```
## # A tibble: 6 x 4
##   review_title                  review_text                   review_star   page
##   <chr>                         <chr>                         <chr>        <dbl>
## 1 Be careful of wrong-way bana~ "We were so excited to get o~ 5.0 out of ~     1
## 2 Perfect, If You Want To Get ~ "I was sitting on the couch ~ 2.0 out of ~     1
## 3 The greatest invention of al~ "Couldn’t live without this ~ 5.0 out of ~     1
## 4 Where are the instructions?   "I don't know how they expec~ 1.0 out of ~     1
## 5 This will end all your trepi~ "This slicer is the best!  F~ 5.0 out of ~     1
## 6 3 bananas at most!  No way t~ "My expectations may have be~ 2.0 out of ~     1
```

Word clouds

```
library(wordcloud)
githubURL <- "https://github.com/maibennett/sta235/blob/main/exampleSite/content/Classes/Week14/1_Tw
load(url(githubURL))

# Now let's play with the data
word_tb <- output_list %>% # This is the list of data we got from Amazon
              bind_rows() %>% #bind the rows to build a data frame
              unnest_tokens(output = "word", input = "review_text", token = "words") %>% #we separa
              count(word) %>% #and we count those words
              filter(!(str_detect(word,"banana") | str_detect(word,"slic") | str_detect(word,"57")
                       str_detect(word,"hutzler") | str_detect(word,"product") | str_detect(word,"de
              anti_join(tidytext::stop_words, by = "word") #And finally, we exclude (anti_join) the
```

Word clouds

```
wordcloud::wordcloud(words = word_tb$word, freq = word_tb$n, min.freq = 5,
                     max.words=200, random.order=FALSE, rot.per=0.35,
                     colors=brewer.pal(10, "Dark2"), family = "Fira Sans Condensed SemiBold")
```

Let's look at some reviews that include the word "time"

```
time <- output_list %>% bind_rows() %>%
  filter(str_detect(review_text,"time"))

head(time)
```

```
## # A tibble: 6 x 4
##   review_title                review_text                       review_star    page
##   <chr>                       <chr>                             <chr>         <int>
## 1 No more butter knives!      Incredible. The time I have sa~ 5.0 out of 5~   463
## 2 Unforunately, it's Too La~ This product is amazing! I wis~ 5.0 out of 5~   463
## 3 better than  chain saw      This product really works well~ 5.0 out of 5~   463
## 4 Fix for wrong bending ban~ Like many other reviewers, I w~ 4.0 out of 5~   463
## 5 Not sure what product eve~ After reading the other review~ 3.0 out of 5~   463
## 6 I used to carry a Glock -~ I had a concealed weapon permi~ 5.0 out of 5~   463
```

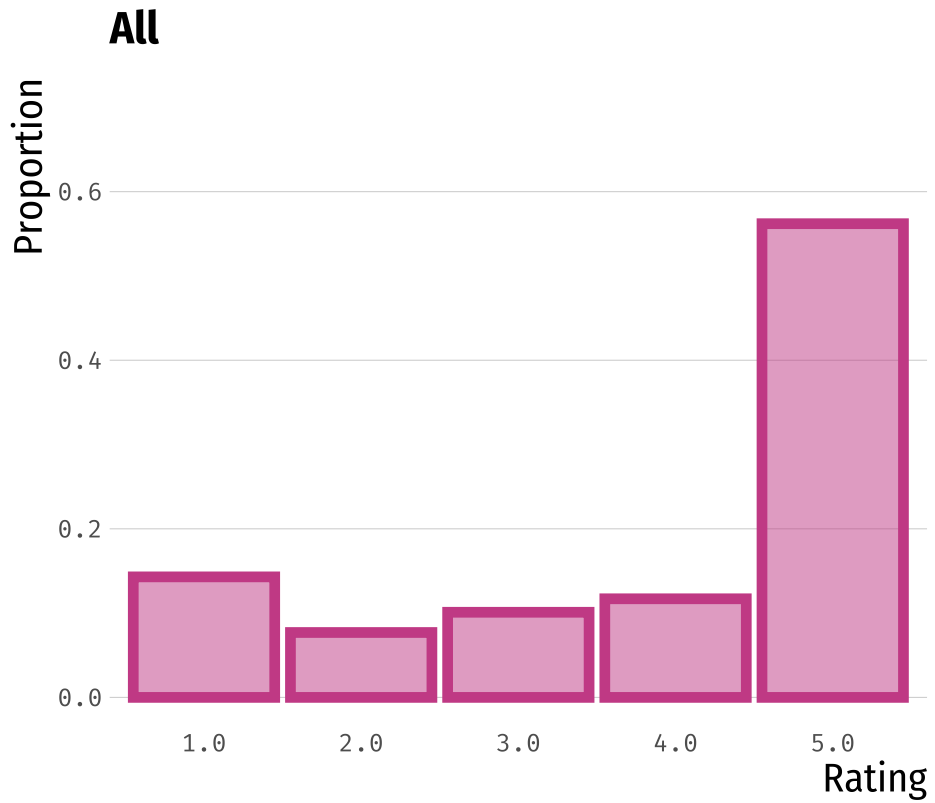
Let's look at some reviews that include the word "time"

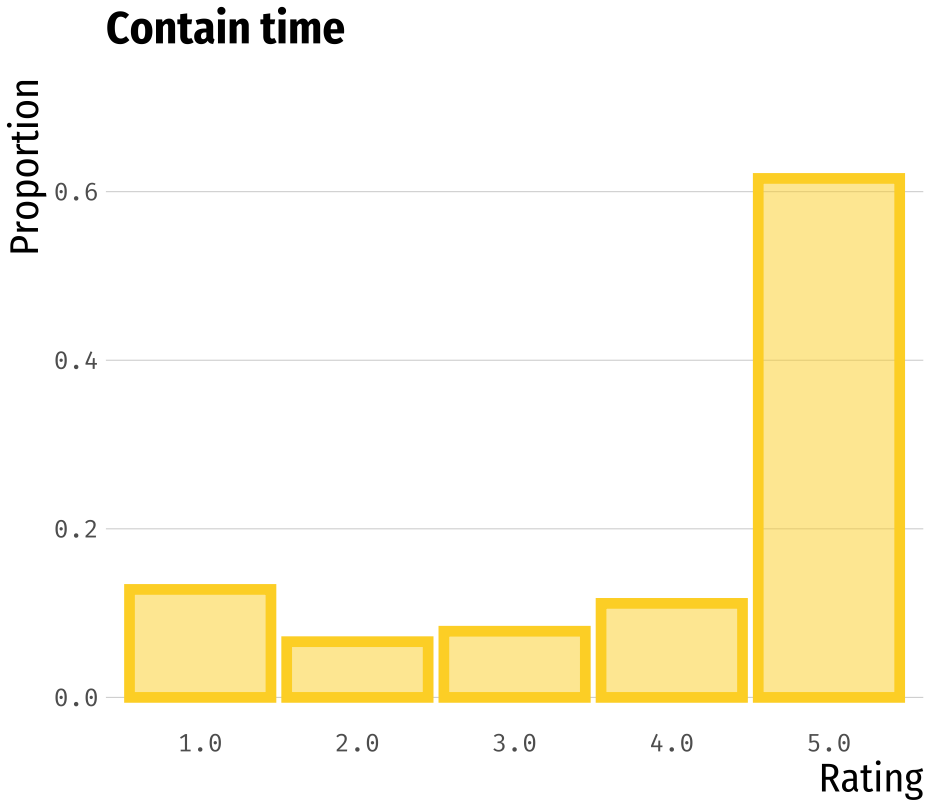
```
df <- output_list %>% bind_rows() %>%
  mutate(time = str_detect(review_text,"time"))

df %>%
  count(review_star) %>% mutate(perc = n / nrow
  ggplot(., aes(x = factor(review_star), y = pe
  geom_bar(stat = "identity") + xlab("Rating")
  ylab("Percentage (%)") + theme_bw() + ggtitle
```

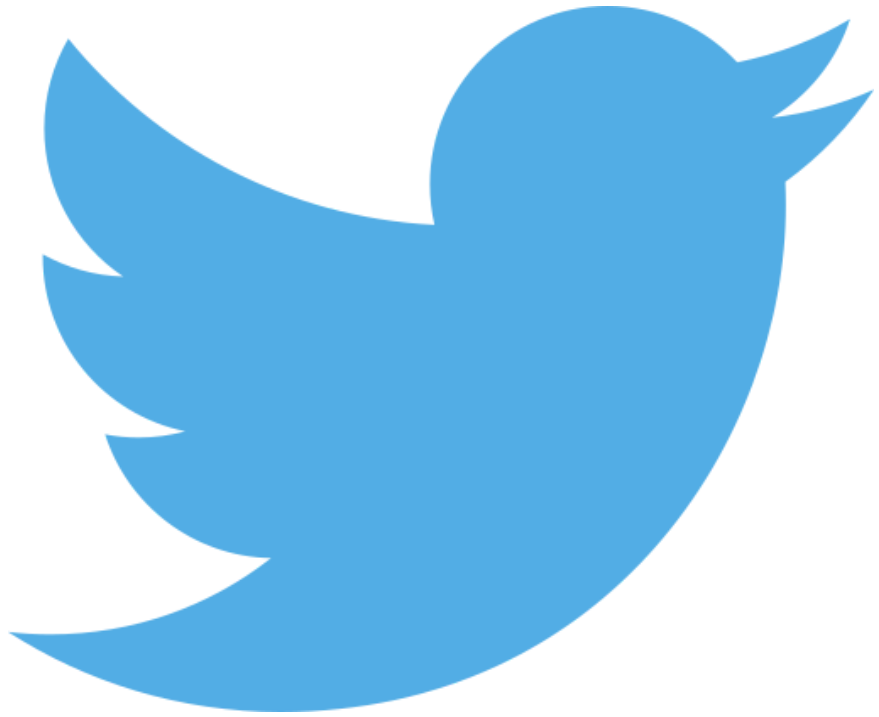
```
df %>% filter(time==TRUE) %>%
  count(review_star) %>% mutate(perc = n / nrow
  ggplot(., aes(x = factor(review_star), y = pe
  geom_bar(stat = "identity") + xlab("Rating")
  ylab("Percentage (%)") + theme_bw() + ggtitle
```

Let's look at some reviews that include the word "time"





Now let's look at Twitter



Let's go to R

Course Instructor Survey

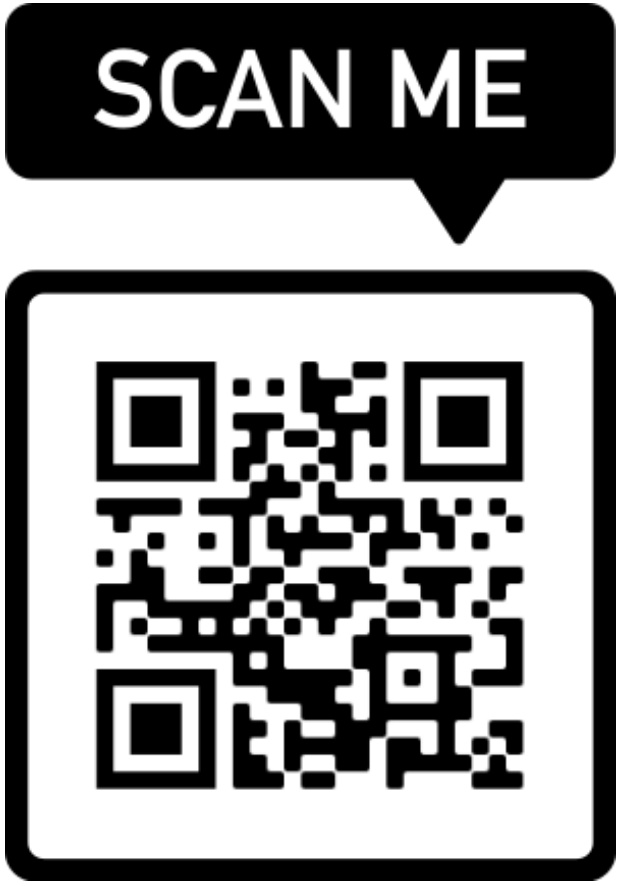
Course Instructor Survey

- To complete the survey either point with your phone to the QR code or go to:

bit.ly/longhorn-cis

The instructor will be back in:

10:00



References

- Chan, M. (2019). "Vignette: Scraping Amazon Reviews in R"
- Udacity (2020). "Natural Language Processing With R"