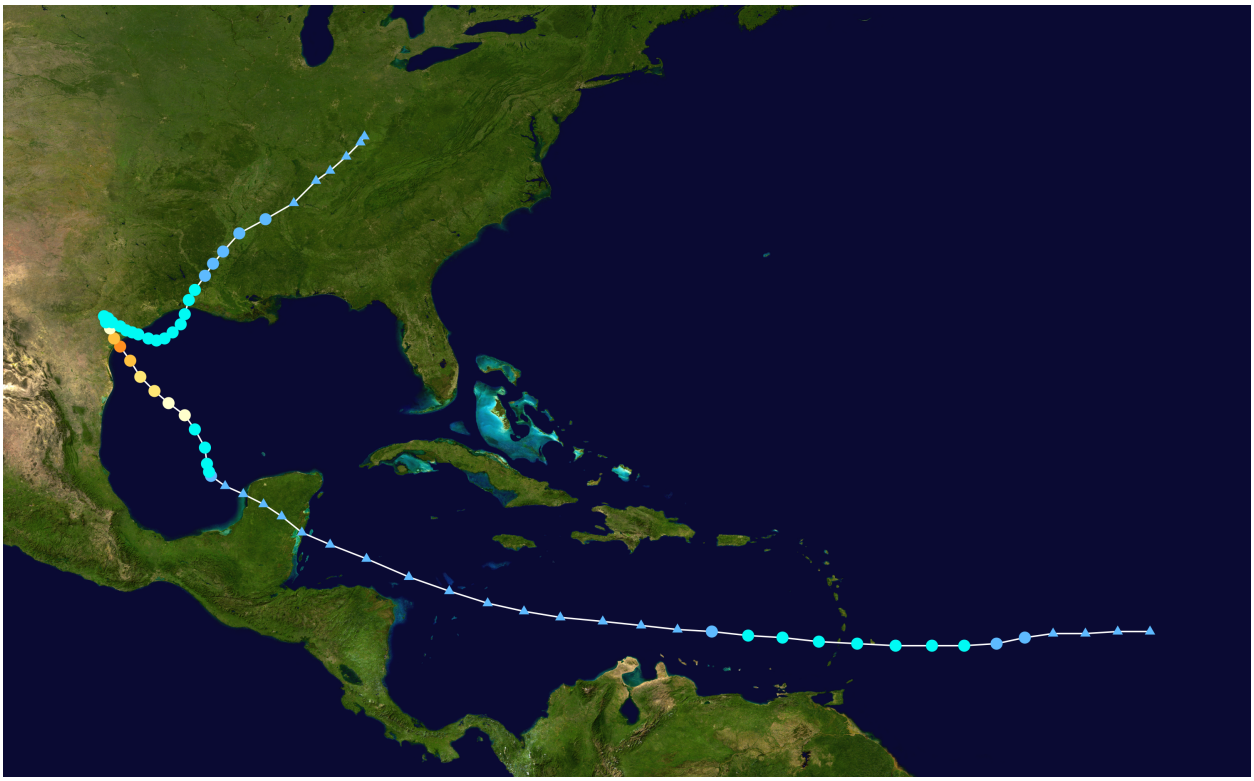
Module 12 Homework: Mining Hurricane Harvey tweets

Sheida Moin and Natalie Nelson



Source: SSHWS/NWS

Background

Hurricane Harvey made landfall as a category 4 storm in Texas on August 26, 2017. Classified as a major hurricane, Harvey brought massive amounts of rain to southeast Texas, resulting in widepsread and devastating flooding.

In this assignment, you will analyze Twitter data tagged "Hurricane Harvey" using text mining and sentiment analysis methods. Through this assignment, we will gain insight on the general tweeting behavior and content of tweets from August 17-29, 2017.

Grading

You will submit one R script for this assignment. To grade this assignment, I will be checking your script to make sure you prepared the correct code for each of the four questions outlined in the assignment. Each question is worth 2.5 points.

Packages

Load the following packages to your workspace:

- `tidyverse`
- `tidytext`

- `wordcloud`
- `textdata`

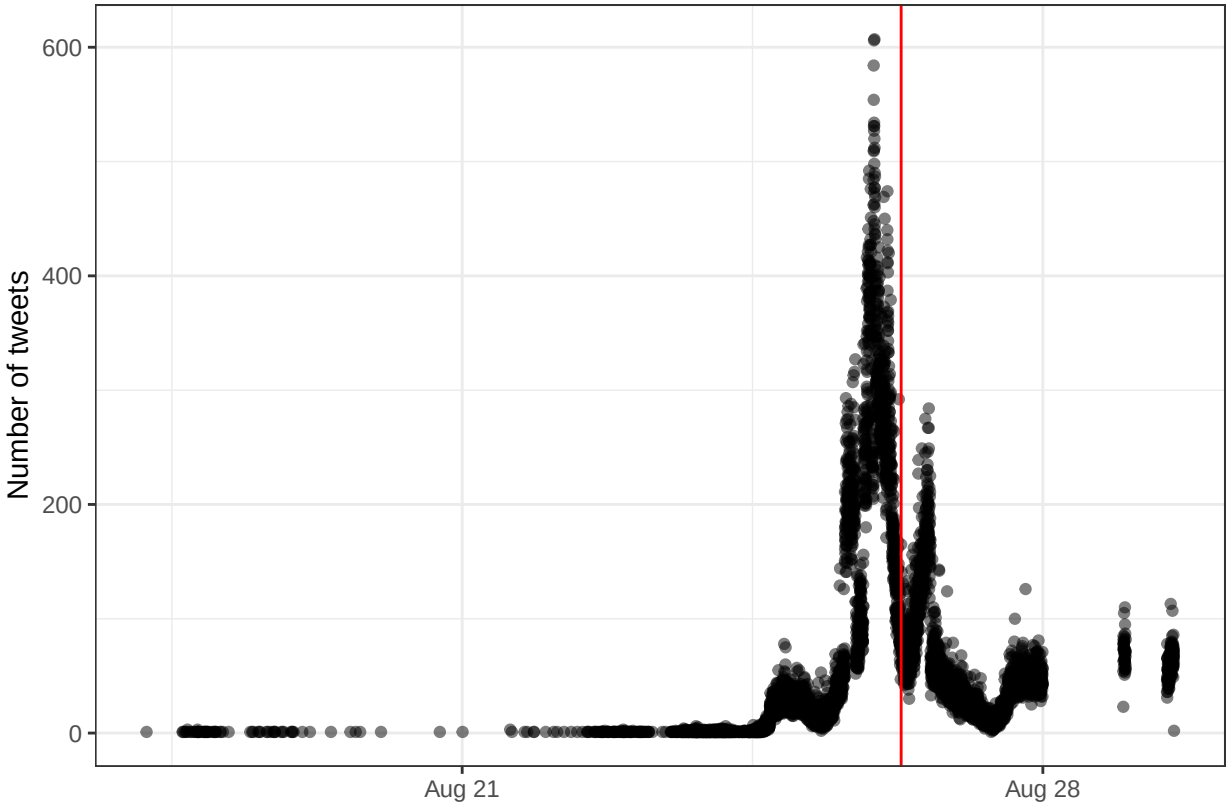
Load and organize data

Tweets that referenced Hurricane Harvey were tabulated and made available on Kaggle under the CC0 Public Domain license. These data are in the `hurricane-harvey-tweets.csv` file included with this assignment. The head of the data are shown below for reference.

```
## # A tibble: 6 x 7
##       id datetime            date       likes replies retweets tweet
##    <dbl> <dttm>              <date>     <dbl>   <dbl>    <dbl> <chr>
## 1     0 2017-08-25 14:44:00 2017-08-25     3       0        0 If you do de~
## 2     1 2017-08-25 14:44:00 2017-08-25     0       0        0 "As Hurrican~
## 3     2 2017-08-25 14:44:00 2017-08-25     6       0        1 "Is @JerryJo~
## 4     3 2017-08-25 14:44:00 2017-08-25     0       0        0 I'm waiting ~
## 5     4 2017-08-25 14:44:00 2017-08-25     0       0        0 The name of ~
## 6     5 2017-08-25 14:44:00 2017-08-25     0       0        0 "@realDonald~
```

Question 1: When did Harvey-related tweets peak in relation to when the hurricane made landfall?

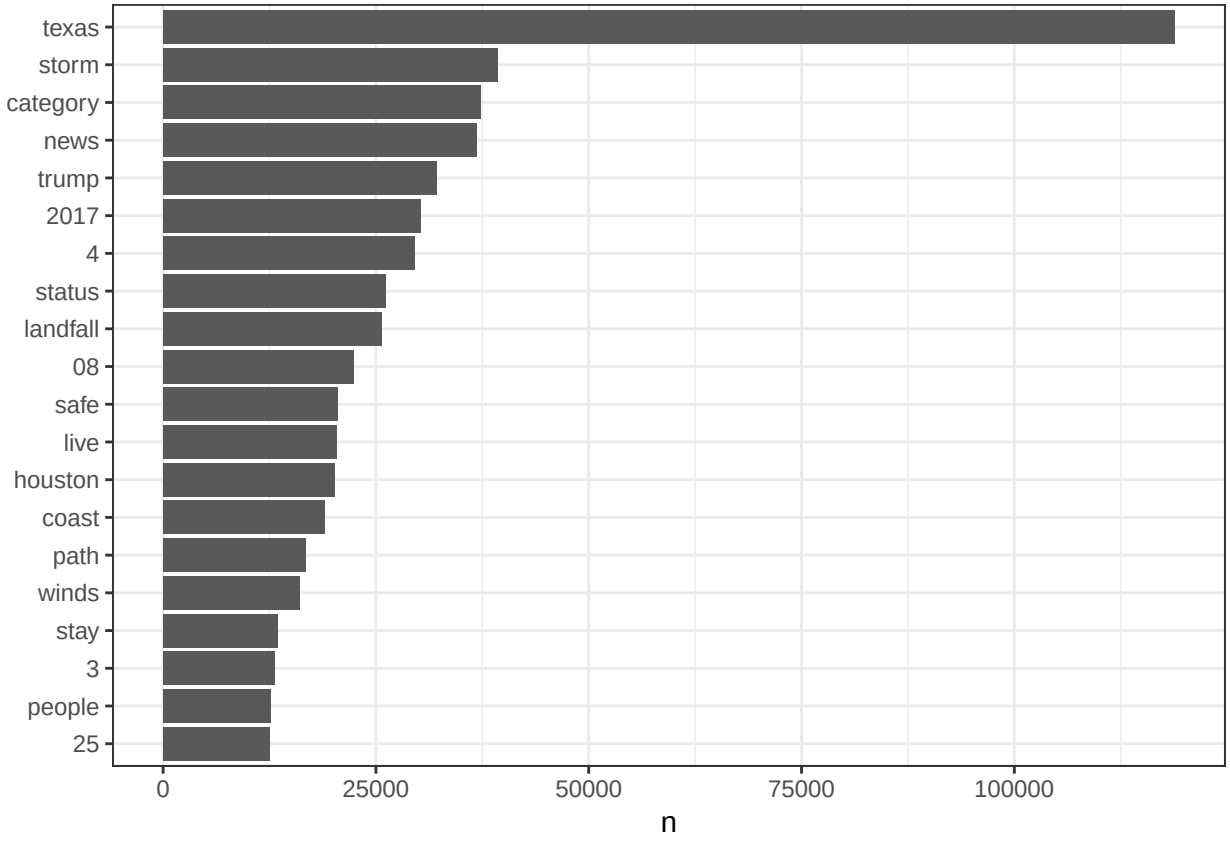
Prepare the figure shown below, which illustrates the number of tweets over time. Include a vertical red line to mark the time at which Harvey made landfall in Texas (`2017-08-26 03:00:00`). To create this plot, you will first need to compute the number (`count()`) of tweets per each `datetime` value (`group_by()`).



Question 2: What are the 20 most commonly used words in the Hurricane Harvey tweets?

Perform a word count analysis of all tweets in the dataset to determine the top 20 most commonly used words. (hint: use `top_n()`)

As part of your word count analysis, apply custom stop words in addition to those in `data(stop_words)`. Refer to chapter 1 in the "Text Mining with R" book for an example on how to include custom stop words. Include the following in your custom stop words: `c("hurricane", "harvey", "hurricaneharvey", "http", "https", "html", "ift.tt", "pic.twitter.com", "twitter.com", "fb.me", "bit.ly", "dlvr.it", "youtube", "youtu.be")`.



Question 3: What are common words used in tweets that reference refineries?

A large number of oil refineries are located in the area around Houston, which was directly impacted by Harvey. Were tweets that referenced refineries primarily focused on potential economic or environmental consequences?

Subset the tweets containing the word "refinery" in its different forms (refinery, refineries) using a function from 14.4 of R4DS. Present the results in a word cloud and include a 2-3 sentence comment on whether tweets that referenced refineries seemed to emphasize the potential economic or environmental impacts. When creating the word cloud, specify that `max.words = 100`. Use the same custom stop words from Question 2.

Question 4: How did the average sentiment of tweets change from August 17-29, 2019?

Using the `afinn` sentiment lexicon, determine the average sentiment of tweets on each `date` in the dataset and create the figure shown below. To create this figure you will:

- Perform an `inner_join` with your tokenized dataset and the `afinn` sentiment lexicon
- Calculate the `mean()` sentiment value per `date`
- Create a column chart. To adjust the x-axis as shown in the figure, use the following scale function: `scale_x_date(date_breaks = "day", date_labels = "%d")`