36-315 Homework 9, Fall 2019

Eu Jing Chua

Due Nov 13, 2019 (11pm) on Canvas

Homework 9

Problem 1

(20 points)

Get the datasets wineQualityReds.csv from Canvas.

(a) (10 points) Plot a dependence graph for the red wines. You need to choose a significance level alpha. This acts as a tuning parameter: small alpha gives a sparse graph. Choose alpha so that the graph is fairly sparse. (No right answer here; just use your judgement.) What alpha did you choose?

(b) (5 points) What variables are related to quality?

(c) (5 points) Repeat (a) and (b) using a conditional independence graph.

Problem 2

(20 points)

Word Clouds and Tidy Text Mining

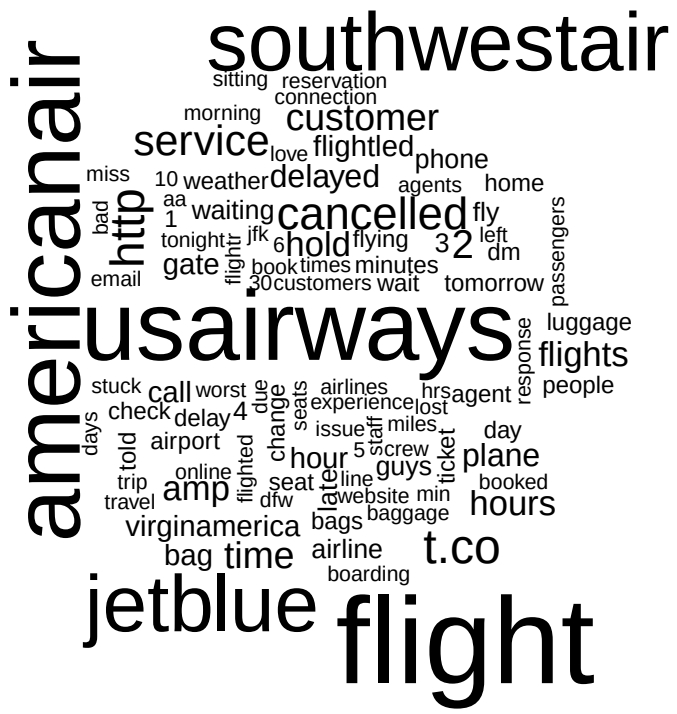
a. (5 points) Read Sections 2, 2.1, and 2.5 of the Tidy Text Mining book (free online). What does the `unnest_tokens()` function do?

b. (5 points) Load the Airline Tweets dataset from one of the first assignments. What column contains the text of the tweets? Run the following code and give an interpretation of the resulting word cloud (you may need to install the `tidytext` and `wordcloud` packages first):

```
#install.packages("tidytext")
#install.packages("wordcloud")
library(tidyverse)
library(tidytext)
library(wordcloud)
data(stop_words)

airline_tweets <- read_csv("https://raw.githubusercontent.com/mateyneykov/315_code_data/master/data/Twe

my_tweets <- dplyr::select(airline_tweets, tweet_id, text) %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words) %>%
  count(word) %>%
  with(wordcloud(word, n, max.words = 100))
```



c. (10 points) Create a separate wordcloud for each airline. Arrange the results into a 2x3 grid. Interpret the results: Are there any words that are more/less common for certain airlines?

Problem 3

(20 points; 4 points for each part)

Sentiment Analysis and Word Clouds with `ggplot()`

Load the airline tweets dataset from here.

Following the example here, create three graphs using the airline tweet text:

a. A word cloud with the words colored by airline.

b. A facetted word cloud (facetting by airline), colored by `user_timezone`.

c. Interpret the plot in (b). Are there any interesting features across the airlines?

d. Follow the example in Section 2.5 of the Tidy Text Mining book to join the **sentiment** of each word to the word counts. Then create a facetted word cloud (facetting by airline), colored by the `sentiment` of the word.

e. Interpret the plot in (d). Are there any interesting features across the airlines?

Problem 4

(20 points)

Topic Modeling

a. (0 points) Read Chapter 6 of the Tidy Text Mining book on Topic Modeling.

b. (0 points) Download the News Articles dataset from Kaggle.

c. (20 points) Recreate the analysis in Chapter 6.1.1 using this dataset (as we did in class).

Problem 5

(20 points)

Briefly describe the dataset that your team will analyze for the final project.