Assignment1

Tian Walker

2024-10-25

EDLD 654, Fall 2022: Assignment 1

The purpose of this assignment is to get you working with the `recipes` package and preprocessing the variables in two different datasets. You will use the same datasets with processed variables to build models in the next assignments.

There are alternative ways to submit your assignment depending on your preference.

1. You can Copy/Edit this notebook and complete it with your responses. Then, you can save and run the completed Kaggle notebook and submit the link through Canvas. If you keep your notebook Private, do not forget to share it with 'UOCOEEDS' so I can access it.
2. You copy/paste the questions and download the datasets from the notebook to your computer. Then, complete the assignment as an R markdown document. Then, you can knit the R Markdown document to a PDF and submit both the .Rmd and PDF files by uploading them on Canvas.
3. You knit the R Markdown document to an HTML document and host it on your website/blog or any publicly available platform. Then, you can submit the .Rmd file by uploading it on Canvas and putting the link for the HTML document as a comment.
4. If you have a GitHub repo and store all your work for this class in a GitHub repo, you can create a folder for this assignment in that repo and put the.Rmd file and PDF document under a specific folder. Then, you can submit the link for the GitHub repo on Canvas.

To receive full credit, you must complete the following tasks. Please make sure that all the R code you wrote for completing these tasks and any associated output are explicitly printed in your submitted document. If the task asks you to submit the data files you created, please upload these datasets along with your submission.

If you have any questions, please do not hesitate to reach out to me.

Task 1: Preprocessing Text Data

Description

For this part of the assignment, you will work with a Twitter dataset which is randomly sampled from a larger dataset on the Kaggle platform (see this link for the original data). In this subset data, there are 1,500 tweets and three variables. A description of the three variables in the dataset follows:

- **sentiment**: a character string variable with two values (Positive and Negative) for the outcome variable to predict.
- **time**: a character string variable indicating time of a tweet (e.g.,Thu Jun 18 07:35:01 PDT 2009)
- **tweet**: a character string variable that provides the full text of a tweet.

This subset data is available as an input data in this R notebook (**'../input/tweets/tweet_sub.csv'**).

Our ultimate goal is to build a model to predict whether or not a tweet has a positive sentiment by using the information from time of the tweet and text of the tweet. We will do this in the following assignments. For

this assignment, we will only engineer features to use them later for building our models and prepare the dataset for model development.

Please complete the following tasks. Provide the R code you wrote and any associated output for each task.

Tasks

Task 1.1 Import the tweet data into the R environment. You can give any name to this data object. Print the structure of this data object using the `str` function.

```
library(rio)
```

```
## Warning: package 'rio' was built under R version 4.4.1
```

```
library(here)
```

```
## here() starts at /Users/tianwalker/Documents/Everything/PhD_harddrive/EDLD_654
```

```
library(recipes)
```

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
##
## Attaching package: 'recipes'
```

```
## The following object is masked from 'package:stats':
##
##     step
```

```
data <- import(here("data/tweet_sub.csv"))
```

```
str(data)
```

```
## 'data.frame':    1500 obs. of  3 variables:
##  $ sentiment: chr  "Negative" "Positive" "Positive" "Positive" ...
##  $ time     : chr  "Thu Jun 18 07:35:01 PDT 2009" "Sun May 10 00:31:52 PDT 2009" "Sun May 31 09:15:19
##  $ tweet    : chr  "I think my twitter is attackd by a kind of worm" "@ddlovato demi if you can i thi
```

Task 1.2 The `time` variable in this dataset is a character string such as *Thu Jun 18 07:35:01 PDT 2009*. Create four new columns in the dataset using this time variable to show the day, date, month, and hour of a tweet. The table below provides some examples of how these four new columns would look like given time as a character string.

time

day

month

date

hour

Thu Jun 18 07:35:01 PDT 2009

4

Jun

18

7

Sun May 10 00:31:52 PDT 2009

7

May

10

0

Sun May 31 09:15:19 PDT 2009

7

May

31

9

Fri May 22 07:25:52 PDT 2009

5

May

22

7

Sun May 31 02:09:52 PDT 2009

7

May

31

2

Sun Jun 07 09:13:08 PDT 2009

7

Jun

7

9

Make sure that `day` column is a numeric variable from 1 to 7 (Monday = 1, Sunday =7), `date` column is a numeric variable from 1 to 31, and `hour` column is a numeric variable from 0 to 23, and `month` column is a factor variable.

Calculate and print the frequencies for each new column (day, month, date, and hour) you created from the original `time`.