STAT 19000 Project 11

Topics: Pattern Matching

Motivation: Many computing platforms and languages have methods to enable easy pattern matching. A common method of identifying patterns is to use regular expressions.

Context: Regular expressions are usually the same (or mostly the same) from one language to another. So if we get familiar with regular expressions in R, we will be able to easily understand how to use regular expressions in other computing platforms too.

Scope: There are a handful of common types of regular expressions. The Basic Regular Expressions in R Cheat Sheet from RStudio is a great resource for remembering many of the common types of regular expressions:

https://rstudio.com/wp-content/uploads/2016/09/RegExCheatsheet.pdf

Please also see the examples contained in the file `project11examples.R` from the `/class/datamine/data/examples` folder.

How to format your solution: Use the `project11template.R` file from the `/class/datamine/data/examples` folder.

Question 1: City name endings

1a. In the 2020 election data (only considering the 9th field, not the 10th field) how many donations are from cities whose name ends in "burg"? How about "boro"? "shire"? "ton"? "town"? "ville"?

Question 2: Analyzing names of donors

2a. How many donations in the 2020 election data have a consecutive, repeated vowel in the (personal) name of the donor? In other words, how many donations are from a donor with AA or EE or II or OO or UU in the donor's name?

2b. Which donor(s) has/have the longest name(s) in the 2020 election data, in terms of character length?

Hint: You need to convert the names to characters, using the `as.character` function, and then you can check the length using the `nchar` function.

2c. How many donations in the 2020 election have donors with the same last name as yours? (For instance, Dr Ward would look for people whose name starts with Ward. You want to make sure to check the beginning of the name, since the last names come first.)

Question 3: Analysis across all election years

As you recall from the examples, there are 235188 town names in (only) the 2020 election data that end in "ton". If we make a table of these names, we see that 679 of those town names from 2020 (ending in "ton") are unique.

3a. Use the method you learned in Project 10, Question 1, to cut *only* the 9th field from the data for *all* election years, and save the result in a file in your home directory called: `myelectiontowns.txt`

Hint: After you cut this data down, using UNIX, and you are ready to read this data into R, you might want to do the following:

```
myDF <- read.delim("myelectiontowns.txt", quote="")
```

3b. How many donations come from cities whose names end in the phrase "ton", across all election years?

3c. How many unique city names are there in question 3b? (For this question, it is safe to only consider the city name and to ignore the State name.)

Project Submission:

Submit your solutions for the project at this URL: https://classroom.github.com/a/7MopMb51 using the instructions found in the GitHub Classroom instructions folder on Blackboard.