DSO 545: Statistical Computing and Data Visualization

Regular Expressions in R and Messy Data- Lab

Fall 2017- LAB

1. Create the following vector of strings in R:

```
fruit <- c("apple", "banana", "pear", "pinapple")
```

2. Run the following lines of code, and try to understand what's happening.

```
str_detect(fruit, "a")
str_detect(fruit, "^a")
str_detect(fruit, "a$")
str_detect(fruit, "[aeiou]")
str_detect(fruit, "[a-d]")
```

3. Using regular expressions, write down a line of R code to detect which of the fruits starts with an "a" and ends with an "e". The following table might help.

Character	Function
?	preceding pattern is optional (matched 0 or 1 time)
*****	preceding pattern is matched 0 or more times
+	preceding pattern is matched at least once (1 or more)
{n}	preceding pattern is matched exactly n times
{n,m}	preceding pattern is matched at least n times & up to m times
{n,}	preceding pattern is matched at least n times

4. Create a parser that detects phone numbers of this format 213 740 4826.
5. How are phone numbers formatted? Look at the body of messages 10 and 18 in the emails dataset. Create a parser that detects those formats of phone numbers.
6. Create a parser that detects zip codes. (e.g. 90028, 90028-0809)