LLCU255 Assignment2

Shenshun Yao 260709204

15/09/2020

Abstract

In this assignment, our goal is to get comfortable working with tables in R based on Novel150 data set.

First, we set working directory, load table and print the name of each column.

```
setwd("~/Desktop/Fall_2020/LLCU255_IntrotoTextMining/Individual Assignment 2")
a<-read.csv("txtlab_Novel150_English.csv")
colnames(a)
```

```
## [1] "filename" "id"       "language" "date"     "author"   "title"    "gender"
## [8] "person"   "length"
```

Problem 1

How many documents are there? How many variables?

Solution:

```
# To get the number of documents, count the number of rows using nrow()
nrow(a)
```

```
## [1] 150
```

There are 150 documents(rows) in total, each document has 9 features(columns), so we have 1350 variables.

Problem 2

How many arthurs are there?

Solution:

```
# unique() returns a vector, data frame or array with duplicate elements/rows remove
length(unique(a$author) )
```

```
## [1] 98
```

There are 98 authors.

Problem 3

Name a novel other than one by Jane Austen that you've read in this data.

Solution:

```r
# Create a subset of Novel150 data set without the novels written by Jane Austen
a_1<-subset(a, a$author!="Austen,Jane")
# Select a random element in the column"title"
sample(a_1$title, 1)
```

```
## [1] "Basil"
```

A random selected answer was printed.

Problem 4

What is the ratio of first-person to third person novels in our sample?

Solution

```r
# Count the number of first-person and third-person novels respectively.
numberofFirstperson<-length(a$person[a$person == "first"])
numberofThirdperson<-length(a$person[a$person == "third"])
# Then divide them to get the ratio
ratio<-numberofFirstperson/numberofThirdperson
ratio
```

```
## [1] 0.4018692
```

The ratio is 0.4018692.

Problem 5

What is the avg length of first and third person novels? Does the difference seem large to you?

```r
firstperson<-subset(a, a$person == "first")
firstperson_avg<-mean(firstperson$length)
firstperson_avg
```

```
## [1] 119562.1
```

```r
thirdperson<-subset(a, a$person == "third")
thirdperson_avg<-mean(thirdperson$length)
thirdperson_avg
```

```
## [1] 124718.6
```

I think the difference does not seam large to me.