DQLab

MINI PROJECT

Working with Text Data using R Programming

Sahda Huwaidah Estiningtyas

Memformat Dataset yang Bersumber dari Data Teks

Instruksi:

1. Lakukan import data dan pemisahan setiap data sesuai dengan separatornya.
2. Tambahkan kolom baru yang berisi gabungan dari **Tempat Lahir** dan **Provinsi**.
3. Carilah data yang mengandung angka pada kolom **Nama** dan hapuslah angka yang terdapat pada data tersebut.

Memformat Dataset yang Bersumber dari Data Teks

- Melakukan import file dan menampilkan 5 data teratas

Input

```
#Melakukan import file
data <- read.table(
  file = "https://storage.googleapis.com/dqlab-dataset/datalahir_teks_dqlab.txt",
  header = FALSE,
  sep = "\n",
  na.strings=c("NA","N/A",""),
  col.names = 'data_list',
  skip = 1)

#Menampilkan 5 data teratas
head(data,5)
```

Output

```
> #Menampilkan 5 data teratas
> head(data,5)
                                                    data_list
1                     Agung|||Semarang|||03 Juni 1992|||Jawa Tengah
2                    Agus|||Probolinggo|||19 Juni 1989|||Jawa Timur
3            Andi|||Pangkalpinang|||11 April 1994|||Bangka Belitung
4                      Andika|||Surabaya|||24 Mei 1999|||Jawa Timur
5 Anggun|||Kota Administrasi Jakarta Pusat|||13 Juni 1981|||Jakarta
```

Memformat Dataset yang Bersumber dari Data Teks

- Menambahkan kolom baru yang berisi gabungan dari **Tempat Lahir** dan **Provinsi**

Input

```r
#Memisahkan data menggunakan strsplit
data <- strsplit(data$data_list,split = "|||", fixed = TRUE)

#Merubah data menjadi dataframe
df <- data.frame(matrix(unlist(data), nrow=length(data), byrow=TRUE))

#Memberikan nama pada setiap kolom
colnames(df) <- c('Nama','Tempat_Lahir', 'Tanggal_Lahir', 'Provinsi')

#Tampilkan 5 baris pertama dari df
head(df,5)

#Tambahkan kolom baru yang berisi tempat lahir dan provinsi
df$kota_provinsi <- paste(df$Tempat_Lahir,",",df$Provinsi)

#Tampilkan 5 data teratas dari df
head(df,5)
```

Output

```
> #Memisahkan data menggunakan strsplit
> data <- strsplit(data$data_list,split = "|||", fixed = TRUE)

> #Merubah data menjadi dataframe
> df <- data.frame(matrix(unlist(data), nrow=length(data), byrow=TRUE))

> #Memberikan nama pada setiap kolom
> colnames(df) <- c('Nama','Tempat_Lahir', 'Tanggal_Lahir', 'Provinsi')

> #Tampilkan 5 baris pertama dari df
> head(df,5)
    Nama                    Tempat_Lahir Tanggal_Lahir        Provinsi
1  Agung                        Semarang  03 Juni 1992     Jawa Tengah
2   Agus                     Probolinggo  19 Juni 1989      Jawa Timur
3   Andi                   Pangkalpinang 11 April 1994 Bangka Belitung
4 Andika                        Surabaya   24 Mei 1999      Jawa Timur
5 Anggun Kota Administrasi Jakarta Pusat  13 Juni 1981         Jakarta

> #Tambahkan kolom baru yang berisi tempat lahir dan provinsi
> df$kota_provinsi <- paste(df$Tempat_Lahir,",",df$Provinsi)

> #Tampilkan 5 data teratas dari df
> head(df,5)
    Nama                    Tempat_Lahir Tanggal_Lahir        Provinsi
1  Agung                        Semarang  03 Juni 1992     Jawa Tengah
2   Agus                     Probolinggo  19 Juni 1989      Jawa Timur
3   Andi                   Pangkalpinang 11 April 1994 Bangka Belitung
4 Andika                        Surabaya   24 Mei 1999      Jawa Timur
5 Anggun Kota Administrasi Jakarta Pusat  13 Juni 1981         Jakarta
                             kota_provinsi
1                  Semarang , Jawa Tengah
2                Probolinggo , Jawa Timur
3          Pangkalpinang , Bangka Belitung
4                   Surabaya , Jawa Timur
5 Kota Administrasi Jakarta Pusat , Jakarta
```

Memformat Dataset yang Bersumber dari Data Teks

- Mencari data yang mengandung angka pada kolom **Nama** dan menghapus angka yang terdapat pada data tersebut
- Menampilkan isi dari data

Input

```r
#Melakukan import file
data <- read.table(
  file = "https://storage.googleapis.com/dqlab-dataset/datalahir_teks_dqlab.txt",
  header = FALSE,
  sep = "\n",
  na.strings=c("NA","N/A",""),
  col.names = 'data_list',
  skip = 1)

#Menampilkan 5 data teratas
head(data,5)

#Memisahkan data menggunakan strsplit
data <- strsplit(data$data_list,split = "|||", fixed = TRUE)

#Merubah data menjadi dataframe
df <- data.frame(matrix(unlist(data), nrow=length(data), byrow=TRUE))

#Memberikan nama pada setiap kolom
colnames(df) <- c('Nama','Tempat_Lahir', 'Tanggal_Lahir', 'Provinsi')

#Tampilkan 5 baris pertama dari df
head(df,5)

#Tambahkan kolom baru yang berisi tempat lahir dan provinsi
df$kota_provinsi <- paste(df$Tempat_Lahir,",",df$Provinsi)

#Tampilkan 5 data teratas dari df
head(df,5)

#Menghapus karakter yang bukan termasuk alphabet pada kolom Nama
df$Nama <- gsub("[^A-Za-z]","",df$Nama)

#Tampilkan isi dari df
df
```

Memformat Dataset yang Bersumber dari Data Teks

- Mencari data yang mengandung angka pada kolom **Nama** dan menghapus angka yang terdapat pada data tersebut
- Menampilkan isi dari data

Output

```
> #Menghapus karakter yang bukan termasuk alphabet pada kolom Nama
> df$Nama <- gsub("[^A-Za-z]","",df$Nama)

> #Tampilkan isi dari df
> df
      Nama                      Tempat_Lahir     Tanggal_Lahir
1    Agung                          Semarang      03 Juni 1992
2     Agus                       Probolinggo      19 Juni 1989
3     Andi                      Pangkalpinang     11 April 1994
4   Andika                          Surabaya       24 Mei 1999
5   Anggun   Kota Administrasi Jakarta Pusat      13 Juni 1981
...
67    Yuli                               Jambi   11 Januari 1984
68   Zahra                             Langsa      25 Juli 2002
69    Zaki                            Cirebon     31 Maret 1999
                  Provinsi                                  kota_provinsi
1              Jawa Tengah                          Semarang , Jawa Tengah
2               Jawa Timur                       Probolinggo , Jawa Timur
3           Bangka Belitung                Pangkalpinang , Bangka Belitung
4               Jawa Timur                          Surabaya , Jawa Timur
5                  Jakarta   Kota Administrasi Jakarta Pusat , Jakarta
...
67                   Jambi                                  Jambi , Jambi
68                    Aceh                                  Langsa , Aceh
69              Jawa Barat                           Cirebon , Jawa Barat
```

certificate