Lab 4/CIS*2250

Pair Programming 3: Visualization

courselink.uoguelph.ca

A. Hamilton-Wright &
K. Raymond

Overview

Learning objectives: ○ Strategies for processing many files ◑ Generating output for use in another program ● Programatically processing general input data ◐ Introduction to visualization using `ggplot2`

Skills

coordination + communication (1/6)

organization + planning (3/6)

teamwork (3/6)

programming + tools (5/6)

strategy (4/6)

visualization (6/6)

(*)[The skill scale is from 0 (Fundamental Awareness) to 6 (Main Focus).]

Image description

A pair of glasses. Image source Wikipedia CC BY-SA 4.0

Overview

In this lab we will again be working in new pairs. With your partner, choose whose lab from last week will be the starting place. Again, a solution to last week's lab is available, but using it will result in a 10% reduction to your grade for this week's lab.

Task 1 Description

Retrieve a copy of your `findFirstNames.pl` Perl script from *last week's lab*. Decide with your new partner which of your implementations to use. If you did not get the previous assignment working, you can download working code from the CourseLink site for but you will lose 10% of your grade for this lab. You will see this available as "Emergency Kit: Solution for Lab 3."

We will now change gears slightly to create code that will track a name's popularity across a number of years. This will require changes to the number of files that will be read in so let us start there.

Copy `findFirstName.pl` to another file named `firstNamesByTime.pl`.

Change your code so that instead of reading in one SS name year file you will read in any number of input files. Your command line will look like the following:

```
$ perl firstNamesByTime.pl 1900 2000 20 querynames.txt
```

Here, the first two parameters (1900, 2000) denote the start and end years that you want to cover in the SS name files, *i.e.*; `yob1900.txt` and `yob2000.txt`.

The next parameter is increment in years for the files in between the start and end years. So in our example we would be considering all of these files: `yob1900.txt`, `yob1920.txt`, `yob1940.txt`, `yob1960.txt`, `yob1980.txt`, and `yob2000.txt`.

The last parameter is the file containing the names that you want to examine for their popularity in the years described in the first three parameters.

The `querynames.txt` file will have the following format:
name,sex
where sex = `F` or `M` (female or male)
For example, see the file listing shown to the right:

```
Andrew,M
Kassandra,F
Davis,M
Julia,F
```

For each name and sex in the `querynames.txt` file, we want to print out the ranking from each of the indicated SS files in the format of a new `.csv` file. This file should have a header line consisting of the field names "`Name,Year,Ranking`" and the remaining lines should consist of the data values for one of the names for a given year. All the data for a given name should appear together, and the years should be in ascending order.

For example, if we run the following command using the above `querynames.txt` file, we should see this output:

```
$ perl firstNamesByTime.pl 1990 2000 3 queryNames.txt
Name,Year,Ranking
Andrew,1990,7
Andrew,1993,10
Andrew,1996,10
Andrew,1999,7
Kassandra,1990,312
Kassandra,1993,120
Kassandra,1996,203
Kassandra,1999,264
Davis,1990,598
Davis,1993,461
Davis,1996,409
Davis,1999,371
Julia,1990,83
Julia,1993,72
Julia,1996,48
Julia,1999,30
```

Lab 4/CIS*2250

Pair Programming 3: Visualization

courselink.uoguelph.ca

A. Hamilton-Wright &
K. Raymond

Overview

Learning objectives: ○ Strategies for processing many files ◑ Generating output for use in another program ● Programatically processing general input data ◐ Introduction to visualization using `ggplot2`

Skills

coordination + communication (1/6)

organization + planning (3/6)

teamwork (3/6)

programming + tools (5/6)

strategy (4/6)

visualization (6/6)

(*)[The skill scale is from 0 (Fundamental Awareness) to 6 (Main Focus).]

Image description

A pair of glasses. Image source Wikipedia CC BY-SA 4.0

Task 2 Description

We will now explore using the `Statistics::R` package to use the powerful `ggplot2` library to produce plots of our data.

You will need to download all of the YoB (Year of Birth) files from CourseLink. You will find the file, `names.zip` in the *Labs* section on *CourseLink*. This contains all the YoB files.

Now you can download the Perl script named `createNameRankPlot.pl` and test it out with your new output.

Run your code and redirect the output into a file (note that here we are using every year's data, not skipping 3 as in the last run):

```
$ perl firstNamesByTime.pl 1990 2000 1 queryNames.txt > plot1.txt
```

To run the plotting programs (even in the THRN labs) you must do the following before running `createNameRankPlot.pl`:

1. Go to the Applications Folder on the machine and double click on R (this is a statistics program)
2. In R type the following command to load the plotting library (the `>` is the R prompt):

   ```
   > install.packages("ggplot2")
   ```

3. Now you can continue on with your perl programming and `createNameRankPlot.pl` will produce lovely PDFs of your plots.

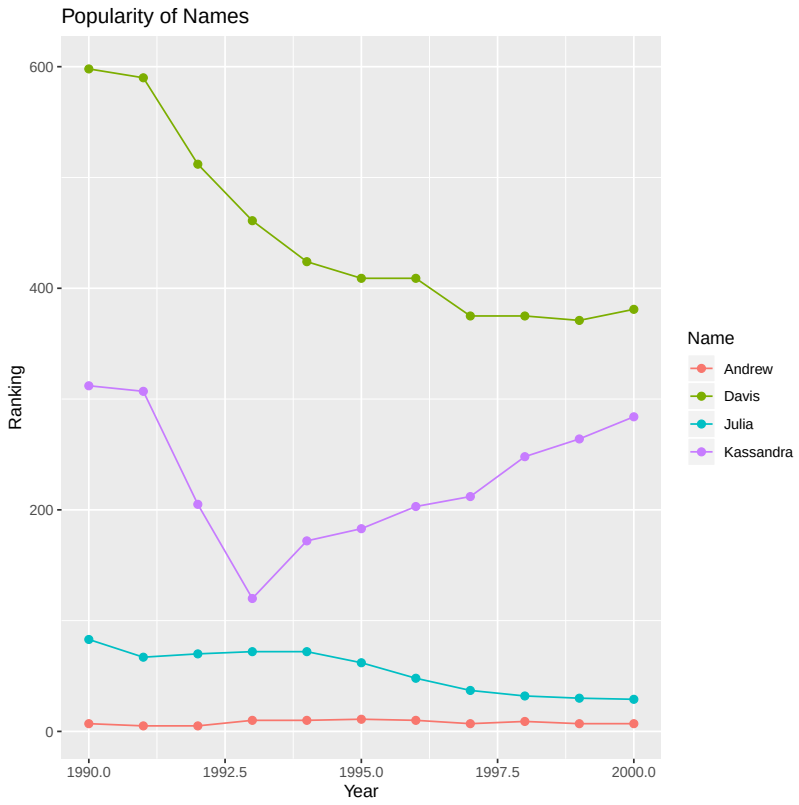
Then run the plotting script:

```
$ perl createNameRankPlot.pl plot1.txt plot1.pdf
```

Then open the PDF file to see what you have created:

```
$ open plot1.pdf
```

You should see a plot likt this:



Lab 4/CIS*2250

Pair Programming 3: Visualization

courselink.uoguelph.ca

A. Hamilton-Wright &
K. Raymond

Overview

Learning objectives: ○ Strategies for processing many files ◑ Generating output for use in another program ● Programatically processing general input data ◐ Introduction to visualization using `ggplot2`

Skills

coordination + communication (1/6)

organization + planning (3/6)

teamwork (3/6)

programming + tools (5/6)

strategy (4/6)

visualization (6/6)

(*)[The skill scale is from 0 (Fundamental Awareness) to 6 (Main Focus).]

Image description

A pair of glasses. Image source Wikipedia CC BY-SA 4.0

Task 3 Description

We can calculate different presentations of data. Examine the script `convertRankingToRankCategory.pl`. This script reads a `.csv` file and will convert the values in a "`Ranking`" column according to the conversion shown to the right.

Ranking	*Category*
0	0
> 2000	1
1000–2000	2
500–999	3
200–499	4
100–199	5
50–99	6
10–49	7
1–10	8

Run this script to convert the `plot1.txt` file to a new `plot2.txt` file:

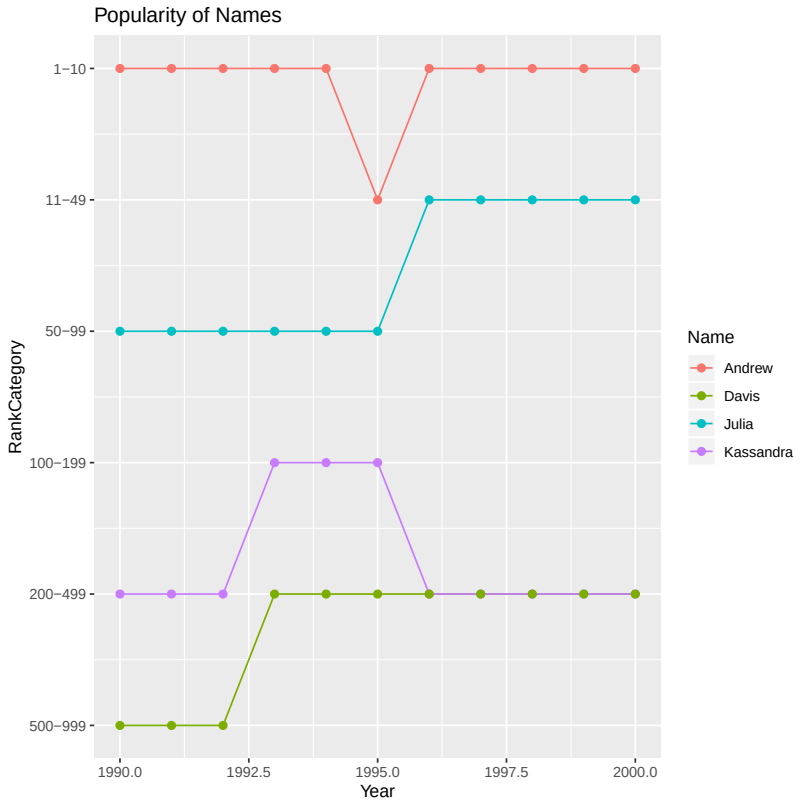
```
$ perl convertRankingToRankCategory.pl plot1.txt > plot2.txt
```

The file `open plot2.txt` will contain the following:

```
Name,Year,RankCategory
Andrew,1990,8
Andrew,1993,8
Andrew,1996,8
Andrew,1999,8
Kassandra,1990,4
Kassandra,1993,5
Kassandra,1996,4
Kassandra,1999,4
Davis,1990,3
Davis,1993,4
Davis,1996,4
Davis,1999,4
Julia,1990,6
Julia,1993,6
Julia,1996,7
Julia,1999,7
```

Creating a plot and then viewing it with produce the plot below:

```
$ perl createNameRankCategoryPlot.pl plot2.txt plot2.pdf
```



Lab 4/CIS*2250

Pair Programming 3: Visualization

courselink.uoguelph.ca

A. Hamilton-Wright &
K. Raymond

Overview

Learning objectives: ○ Strategies for processing many files ◑ Generating output for use in another program ● Programatically processing general input data ◐ Introduction to visualization using `ggplot2`

Skills

coordination + communication (1/6)

organization + planning (3/6)

teamwork (3/6)

programming + tools (5/6)

strategy (4/6)

visualization (6/6)

(*)[The skill scale is from 0 (Fundamental Awareness) to 6 (Main Focus).]

Image description

A pair of glasses. Image source Wikipedia CC BY-SA 4.0

Task 4 Description

As a final lab task, consider what the differences are between `createNameRankPlot.pl` and `createNameRankCategoryPlot.pl`.

Why does `createNameRankPlot.pl` not work (it gives an error) when it is run on the data in `plot2.txt`?

Complete the Quiz in Courselink that is part of Lab 4 to provide your answer, and be sure to upload your `firstNameByTime.pl` file, along with the `plot1.pdf` and `plot2.pdf` visualization files that you created.