CS209 Assignment2

DUE: 21 April

Data visualisation

This assignment continues on the same topic as the first assignment.

In Assignment 1 we preprocessed some data files from your friend, and collected them into a systemically classified, and consistently encoded state. Now, you are able to find some general information from these files. You also want to present the information graphically in several charts that your friend and others can see. The visualization will help your friends to understand the training data used in the machine learning algorithms they develop...

In summary, to complete this assignment, you need to write a JavaFX program for data visualisation and complete the following tasks.

summary :

1. Load a default config file to get the 'root' and the 'intervals'
2. Draw the bar chart for number of files per genre

   > This chart gives you a brief idea of the distributions of files (text files in Assignment 1) with respect to genres (fields in Assignment 1, e.g. Technology, Economy, Society, Sports).

3. Draw the charts for different file size intervals of each genre
4. User can change the 'root' and the 'intervals' by button or other interaction
5. (Optional) User can change the chart type for 'intervals' charts by choice box or other interaction

1. Load a default config file

The default config file is named config.yml and put in the root of the project, which means the path is "./config.yml".

A possible configuration file:

```
Root: ./test_datas
Intervals:
  - 100
  - 2000
  - 4000
IntervalNames:
  - small
  - mid
  - big
  - giant
```

The basic rules for configuration are as follows:

- Root means the directory root for Counter
- Intervals mean the size interval for classification , in this example, it means [0,100), [100,2000),[2000,4000),[4000,+) bytes.
- IntervalNames means the name for each interval.
- The length of IntervalNames will be always one more than the intervals'

The encoding of this file will be always UTF-8.

2.Draw the first chart (Genre Statistics)

For this example, the file tree is shown as following:

```
test_datas
    |  newout.csv
    |  output_for_test_case_1
    |
    ├─专题报道
    |  ├─南科大
    |  |  └─半年以上
    |  |          sustech_0_cs.txt
    |  |          sustech_13_cs.txt
    |  |          sustech_1_cs.txt
    |  |          sustech_4_cs.txt
    |  |          sustech_7_cs.txt
    |  |
    |  ├─文匯報
    |  |  └─一月内
    |  |          sustech_16_cs.txt
    |  |          sustech_17_cs.txt
    |  |
    |  └─深圳本地宝
    |      └─半年以上
    |              sustech_15_cs.txt
    |
    ├─书院新闻
    |  └─南科大
```

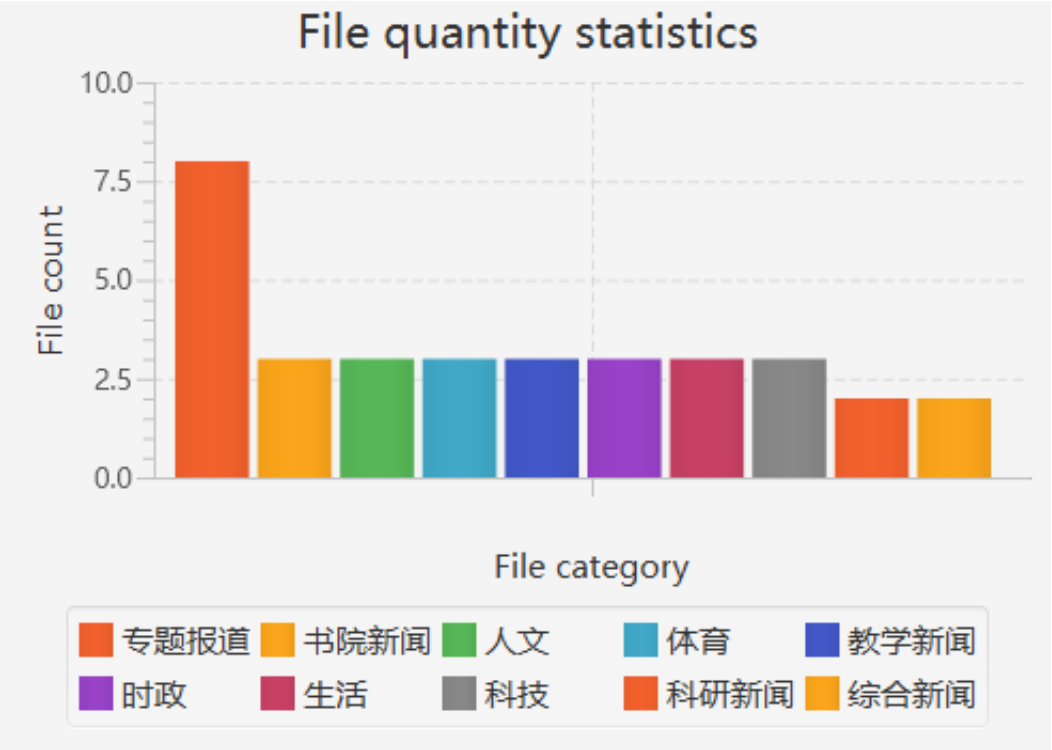
```
|       └半年以上
|               sustech_11_cs.txt
|               sustech_6_cs.txt
|               sustech_9_cs.txt
|
├人文
|   ├人日
|   |   └一月内
|   |           hi_5_there.txt
|   |
|   ├天涯
|   |   └半年以上
|   |           hi_10_there.txt
|   |
|   └腾讯
|       └半年以上
|               hi_15_there.txt
|
├体育
|   ├人日
|   |   └半年内
|   |           hi_13_there.txt
|   |
|   ├新浪
|   |   └一周内
|   |           hi_8_there.txt
|   |
|   └腾讯
|       └半年内
|               hi_3_there.txt
|
├教学新闻
|   └南科大
|       ├半年以上
|       |       sustech_14_cs.txt
|       |       sustech_3_cs.txt
|       |
|       └半年内
|               sustech_12_cs.txt
|
├时政
|   ├人日
|   |   └一周内
|   |           hi_1_there.txt
|   |
|   ├天涯
|   |   └一月内
|   |           hi_6_there.txt
|   |
```

```
|   └─腾讯
|       └─半年内
|               hi_11_there.txt
|
├─生活
|   ├─天涯
|   |   └─一周内
|   |           hi_2_there.txt
|   |
|   ├─新浪
|   |   └─半年内
|   |           hi_12_there.txt
|   |
|   └─腾讯
|       └─半年内
|               hi_7_there.txt
|
├─科技
|   ├─人日
|   |   └─半年内
|   |           hi_9_there.txt
|   |
|   ├─天涯
|   |   └─半年以上
|   |           hi_14_there.txt
|   |
|   └─新浪
|       └─一月内
|               hi_4_there.txt
|
├─科研新闻
|   └─南科大
|       └─半年以上
|               sustech_10_cs.txt
|               sustech_5_cs.txt
|
└─综合新闻
    └─南科大
        └─一月内
                sustech_2_cs.txt
                sustech_8_cs.txt
```

- you can ignore the files in root.
- you should categorize according to the first level catalog.
- this chart should be a bar chart.

The statistical logic has been provided in Counter.java. And you should read and use it.

One possible result is shown following:

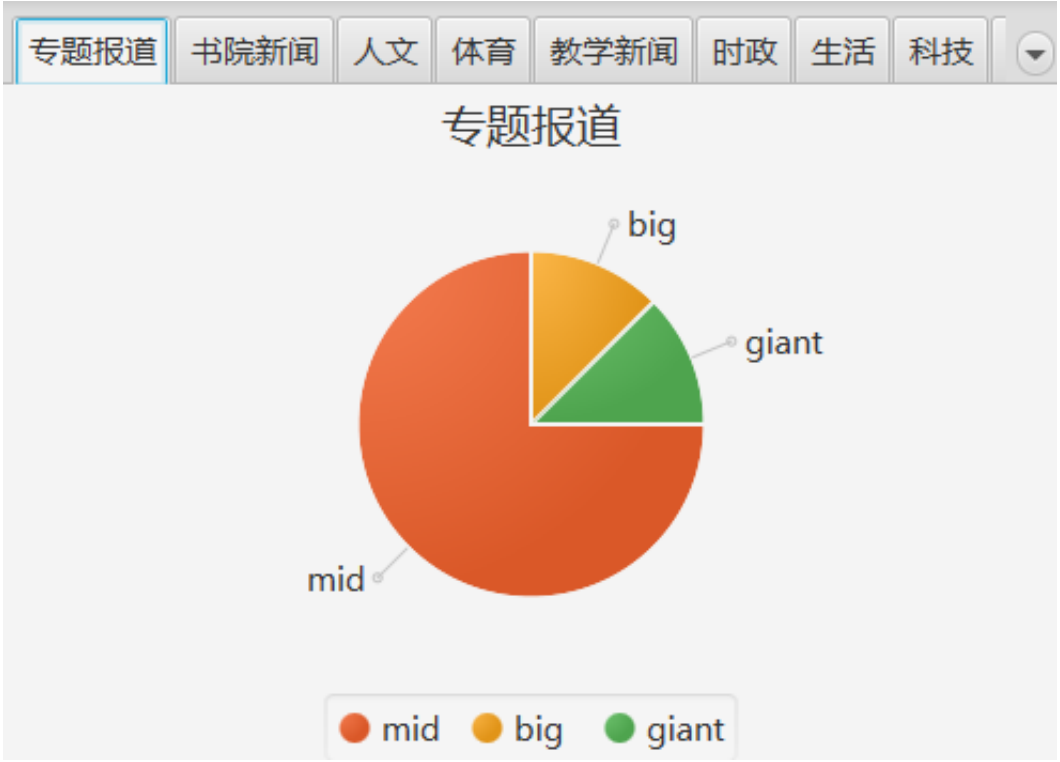


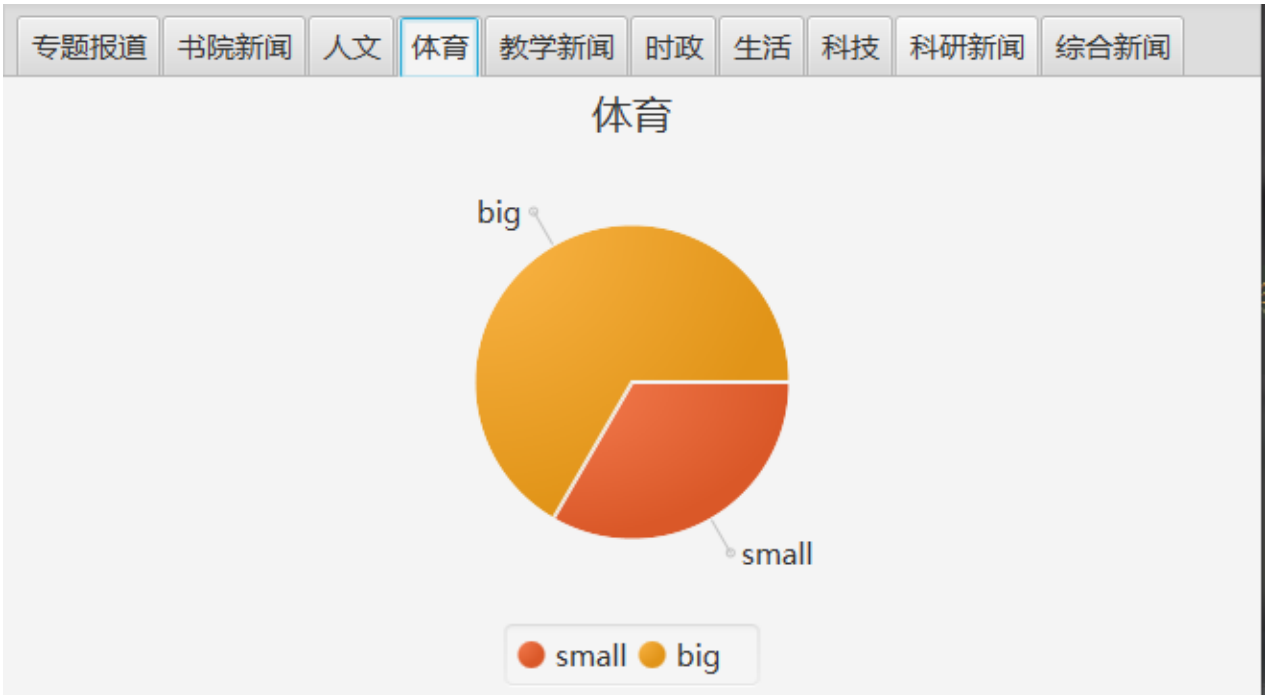
3.Draw the second charts (File size statistics)

- You should draw a bar or a pie chart for each genre, which means in this example there will be 10 charts.
- The number of charts required is dynamically calculated based on the data determined from the directory structure starting at the root. You should dynamically add the charts to the pane at runtime.
- for each chart:
    - the name of each part should be the string in the corresponding intervalNames.
    - the value of each part should be the proportion of the number of current interval files to the total.
    - if the ratio is 0, which means there no file is in this interval this would not be shown in the generated chart.
    - the title of the chart should be the corresponding name in first chart.

The statistics logic has been provided for you to use in the file Counter.java. Tou should use this file to generate the data.

One possible result is shown following:





4.Change the root and intervals settings interactively

1. Change the root setting:

   Use a DirectoryChooser to choose the new root, and then statistics it (know that the intervals has not been changed) and refresh your UI.

2. Change the intervals setting:

   Use a FileChooser to choose a new yml file, which file will contain the 'Intervals' and 'IntervalNames' (or more information, but you only pay attention to these two), and then statistics by this new setting (know that the root has not been changed) and refresh your UI.

3. Change both two settings:

Use a FileChooser to choose a new yml file, which file should have the information as default config file have, and then statistics it and refresh your UI.

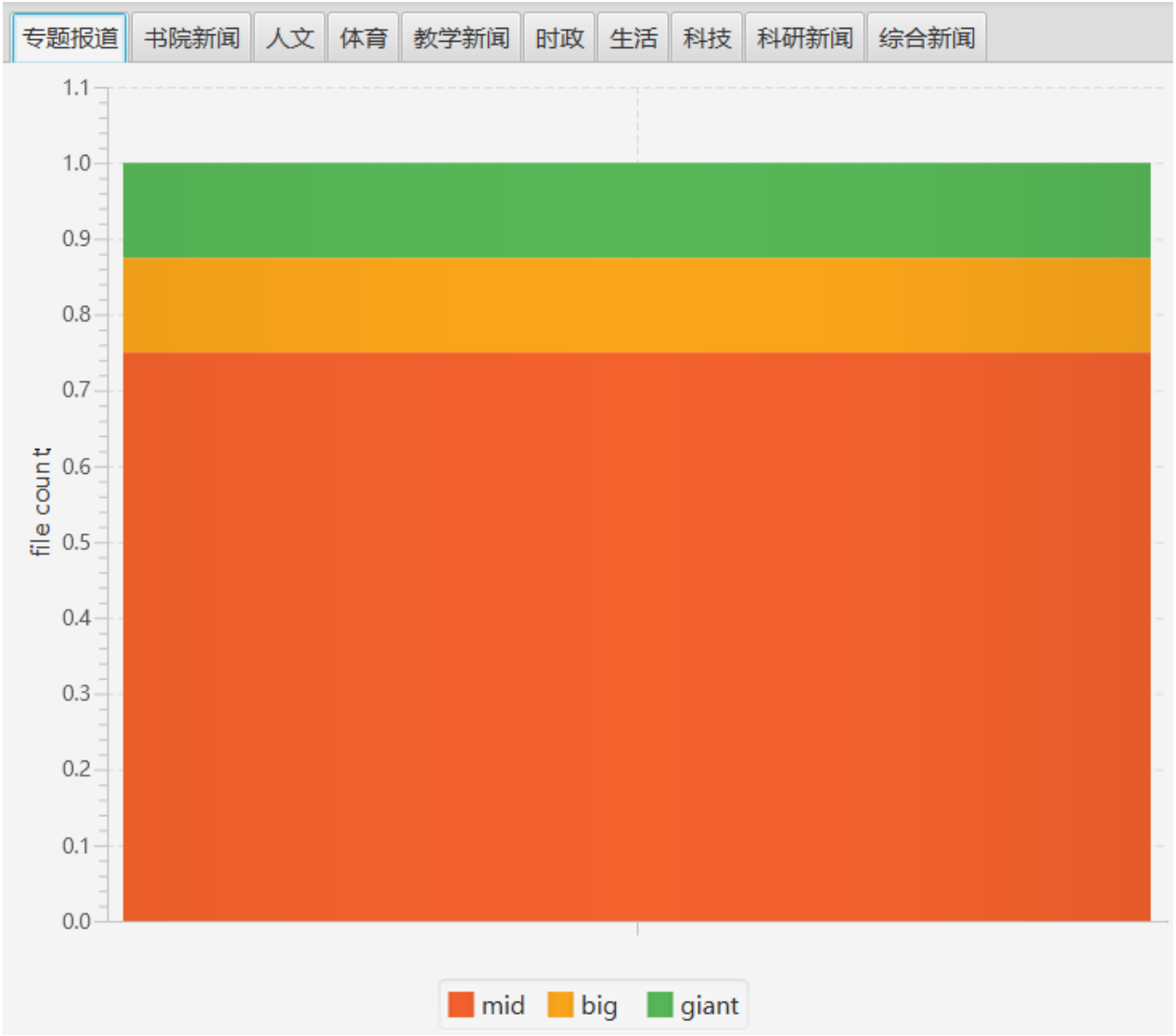
The FileChooser should just accept *.yml files

These interactions can be trigger by buttons or other modules.

(optional / bonus marks) 5.change the chart type for 'intervals' charts by interaction

User can change the chart type (e.g. switch between a pie chart, a bar chart, a Pareto chart etc).

Following is a possible stacked bar chart:



Other

- When some error was caused by the user, you can alert the user.

  For example: user choose a yml file which does not contain the info we need,then you can alert that the yml file is wrong.
- Any reasonable interactions is acceptable
- Some reading materials for pareto chart:

- https://support.minitab.com/zh-cn/minitab/18/help-and-how-to/quality-and-process-improvement/quality-tools/supporting-topics/pareto-chart-basics/
- https://en.wikipedia.org/wiki/Pareto_chart
- https://baike.baidu.com/item/%E5%B8%95%E7%B4%AF%E6%89%98%E5%9B%BE/8735273?fr=aladdin

Submission Requirements

Submission of Assignment 2:

(1) You should submit the whole project and compress the folder to a .zip file.

(2) **No Chinese characters** are allowed to appear in your code.

(3) The assignment should be submitted before the deadline **(Apr.21. 16:00pm). Late submissions within 24 hours after the deadline (even a few minutes) will incur a 50% penalty**, meaning that you can only get 50% of the score, which you could get if the assignment was submitted before the latest deadline. **Assignments submitted after the latest deadline will not be graded** (meaning your will get a zero for the assignment).