Word frequency histogram

Program processes text and show the its "word histogram" -- counts every word and show this count.

Task

Use any language you choose to create an app that will get a text file as input and provide its word histogram as output. – display only words that appear more then 1 time in the text.

The program will be tested with the following text:

“Progforce, more than anything else, is a league of extraordinary talent - sought, refined, and dedicated to providing the most impeccable intelligence and service. Our team of professional software developers are

specially chosen through a process of selection based not only on training, but conception and creative application. Our people don't just plug in the numbers. Our people create custom solutions for custom needs."

The output should be:

of => 3
and => 3
our => 3
custom => 2
people => 2
the => 2

rules:

1. the process should be case insensitive (we don't care about Capital or Lowercase).
2. A word should be at least 2 characters long.
3. Ignore special characters like ".,:" etc. one exception of this rule is words like: don't.

Customizable settings

User can config:

typeOfInput / typeOfOutput

Type of input/output data.

Variants:

- ***console*** -- input text manually from console or output to console.

Note: for ending of enter data it must be clicked enter-button twice

- ***file*** -- input/output text from/to file.

The second parameter is file-path.

Note: if there is no input-file then it'll throws exception (this case showed in test_14). Output-file is generated automatically.

- ***folder*** -- input/output text from/to files in folder.

The second parameter is folder-path.

Note 1: if there is no input-folder then it'll throws exception. Output-folder and files are generated automatically.

Note 2: out-files have the same name as in-files.

So, as it can be concluded, that there are nine types of i/o data -- it showed in tests 1-9.

minWordSize

The minimum size of word for taken into account.

Note: Word is the combination of any letters, digits and sign "".

minWordsCount

The minimum count of words for taken into account.

caseSensitive

Case sensitivity.

Default settings

```
typeOfInput : file>>src\studyJava\ProgForce\files\text_in.txt
typeOfOutput : file>>src\studyJava\ProgForce\files\text_out.txt
minWordSize : 2
minWordsCount : 2
caseSensitive : false
```

So, by default it works according to “task-requires”.

Test cases

Shortly: it was tested 16 cases. More details are in file: `WordHistogramDemoTest.java`. Here is short contents of tests.

Test 1

Data input and output manually (console-console), other options are default

Test 2

Console-file, other options are default

Test 3

Console-folder, other options are default

Test 4

File-console, other options are default

Test 5

File-file, other options are default

Test 6

File-folder, other options are default

Test 7

Folder-console, other options are default

Test 8

Folder-folder, names of folders are different, other options are default

Test 9

Folder-file, other options are default

Test 10

File-file, minWordSize=1, minWordsCount=1, caseSensitive=false
This case counts usage of all words in the text.

Test 11

File-file, minWordSize=2, minWordsCount=1, caseSensitive=false
This case counts usage of all words (with size>=2) in the text.

Test 12

File-file, minWordSize=2, minWordsCount=2, caseSensitive=true
This case make program sensitive to letter-case, so words “word” and “wOrD” are different and counts separately.

Test 13

Defaults.
Case: in-file is empty (out-file has only header).

Test 14

Defaults.
Case: in-file is absent.
Out: exception.

Test 15

Folder-folder, names of folders are the same, other options are default.
Program create new folder, based on [in-folder name] + "_out"

Test 16

File-file, minWordSize=1, minWordsCount=2, caseSensitive=false
This case counts usage of all repeated words in the text.