

Scientific Programming

Midterm simulation

Introduction

Luca Bianco - Academic Year 2017-18
luca.bianco@fmach.it

Available material

in: /usr/local/sciprolab1/



lab



python-3.6.3-docs-
html



theory_slides



BiopythonTutorial.
pdf



Matplotlib.pdf



numpy-ref-1.13.0.
pdf



pandas.pdf

Problem 1

The file `biogrid-interactors.tsv` has 5 columns separated by a tab (`:raw-latex:``\t```):

```
Interaction InteractorA InteractorB TaxidA TaxidB
```

An example with the header follows:

```
Interaction InteractorA InteractorB TaxidA TaxidB
499073 entrez:8945 entrez:155945 9606 11676
```

Each line of the file describes an interaction, the first data line reports interaction `499073` involving interactors `entrez:8945` (belonging to organism `9606`) and `entrez:155945` (belonging to organism `11676`).

Write the following functions:

1. `loadData(filename)` that loads the data and returns a dictionary with the number of interactions for each organism. Remember to skip the first line. **Note: an interaction where TaxidA is different from TaxidB, will increase the count of both organisms, while if TaxidA and TaxidB are the same organism the interaction will increase the count by one).**
2. `printData(interDict, minCount)` that gets the dictionary created by `loadData` and:
 - a. prints the organisms having a number of interactions `>` of `minCount`;
 - b. prints the total number of organisms present and the average number of interactions per organism;

Problem 1: answer

Calling

```
myDict = loadData(myfile)
printData(myDict,2000)
```

should print:

```
Organisms with more than 2000 interactions:
Taxid: 316407    interactions: 171101
Taxid: 10116     interactions: 5561
Taxid: 9606      interactions: 329889
Taxid: 6239      interactions: 8662
Taxid: 284812   interactions: 70451
Taxid: 7227      interactions: 48591
Taxid: 36329     interactions: 2543
Taxid: 559292   interactions: 673581
Taxid: 10090     interactions: 38748
Taxid: 3702      interactions: 42591
Total number of organisms: 61
Avg interactions x organism: 22928.59
```

Problem 2

The file `biogrid-interactions.tsv` has 3 columns separated by a tab (`:raw-latex:~\t~`):

Interaction	InteractionTypes	ConfidenceValues
-------------	------------------	------------------

An example with header follows:

Interaction	InteractionTypes	ConfidenceValues
783952	psimi:MI:0403 (colocalization)	1.0
701836	psimi:MI:0915 (physical association)	
551345	psimi:MI:0799 (additive genetic interaction defined by inequality)	3.937113975
1199912	psimi:MI:0799 (additive genetic interaction defined by inequality)	0.2259

Each line represents an interaction, the first data line describes interaction `783952` that is a `psimi:MI:0403 (colocalization)` and has confidence value `1.0`. **Note that confidence values are not always present, like in the second data line.**

Write the following functions:

1. `loadInteractions(filename)` that loads the tab separated value file in a dictionary (hint: use `Interaction` as the key) and prints the total number of interactions present. Remember to skip the first line.
2. `findByTerm(term, interDict)` that gets the dictionary created by `loadInteractions` and prints the number of interactions with the keyword `term` in the `InteractionType`.

Ex. considering the 4 entries above, `findByTerm("genetic", interDict)` would print:

```
2 entries have keyword "genetic" in the interactionType
```

Problem 2: answer

Calling

```
myDict = loadInteractions(myfile)
findByTerm(myDict,"association")
findByTerm(myDict,"colocalization")
findByTerm(myDict,"interaction")
```

should give

```
Loaded 1370394 interactions
337484 entries have keyword "association" in the interactionType
44057 entries have keyword "colocalization" in the interactionType
988853 entries have keyword "interaction" in the interactionType
```

Problem 3

The two tab separated files of the previous problems, [biogrid-interactors.tsv](#) and [biogrid-interactions.tsv](#) have a common column "Interaction".

Write a python program that loads both files and:

1. Writes the complete information (i.e. Interaction InteractionTypes ConfidenceValues InteractorA InteractorB TaxidA TaxidB) for the entries having ConfidenceValues > the mean ConfidenceValue to a tab separated value file. Prints the number of written entries and the mean ConfidenceValue of the global dataset;
2. Reports the average ConfidenceValues for each InteractionType and produces a boxplot of all the Confidence values;

Hint: load the two files as pandas DataFrames and merge them on the "Interaction" column. **Hint1:** you can use DataFrame.to_csv to write a DataFrame to a text file (choose the appropriate separator!)

Problem 3: answer

The mean ConfidenceValue is 675.49
98 entries have a ConfidenceValue > 675.49

Mean Confidence per InteractionType:

InteractionTypes	
psimi:MI:0403 (colocalization)	0.607030
psimi:MI:0407 (direct interaction)	2.613836
psimi:MI:0794 (synthetic genetic interaction defined by inequality)	124005.029496
psimi:MI:0796 (suppressive genetic interaction defined by inequality)	2.319607
psimi:MI:0799 (additive genetic interaction defined by inequality)	2.516134
psimi:MI:0914 (association)	NaN
psimi:MI:0915 (physical association)	7.945797

Name: ConfidenceValues, dtype: float64

Box plot of confidence values

