Scientific Programming
Partial 2016/12/19

Stefano Teso and Toma Tebaldi

December 19, 2016

Assignment

Given:

- The two `*.tsv` files, describing a set of protein **interactions** occurring in multiple organisms.
- A user-provided list of **organism identifiers**.
- A user-provided `float` **threshold**.

write a Python program that, for each of the given organisms:

- **Prints** how many (distinct) proteins, how many (possibly duplicated) interactions, and the list of hubs in that organism. (We define as hubs those proteins whose number of interaction partners is in the 99%-percentile.)
- **Plots** a histogram of the number of interactions for each protein in that organism.

Interactions with score below the **threshold** should be discarded.

interactions.tsv

```
Interaction InteractionTypes ConfidenceValues
340098 "psi-mi:""MI:0407""(direct interaction)"
466046 "psi-mi:""MI:0794""(synthetic genetic interaction defined by inequality)"
344378 "psi-mi:""MI:0796""(suppressive genetic interaction defined by inequality)"
718382 "psi-mi:""MI:0915""(physical association)"
783952 "psi-mi:""MI:0403""(colocalization)" 1.0
...
```

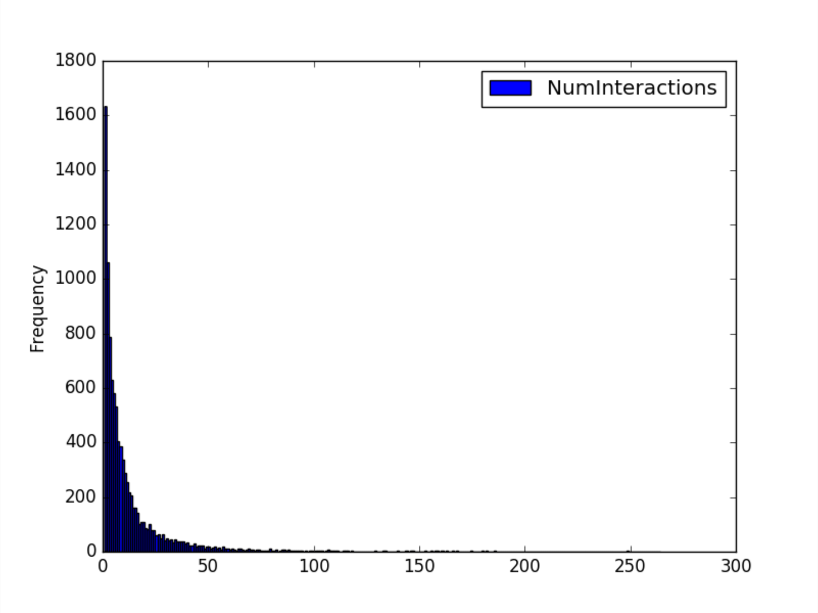
interactors.tsv

```
Interaction InteractorA InteractorB TaxidA TaxidB
340098 entrez gene/locuslink:3621 entrez gene/locuslink:126961 taxid:9606 taxid:9606
466046 entrez gene/locuslink:851967 entrez gene/locuslink:856607 taxid:559292 taxid:559292
344378 entrez gene/locuslink:855228 entrez gene/locuslink:854976 taxid:559292 taxid:559292
718382 entrez gene/locuslink:9675 entrez gene/locuslink:5591 taxid:9606 taxid:9606
783952 entrez gene/locuslink:5719 entrez gene/locuslink:5708 taxid:9606 taxid:9606
...
```

Expected Output

```
$ python exam.py
write space-separated taxa: 9606
write a threshold: 0
taxid:9606 has:
     9934 interactors (unique)
     130344 known interactions (possibly duplicated)
[’entrez gene/locuslink:10236’ ’entrez gene/locuslink:10291’
 ’entrez gene/locuslink:10768’ ’entrez gene/locuslink:10856’
 ’entrez gene/locuslink:10980’ ’entrez gene/locuslink:1915’
 ...
 ’entrez gene/locuslink:84365’ ’entrez gene/locuslink:84419’
 ’entrez gene/locuslink:9045’ ’entrez gene/locuslink:9349’
 ’entrez gene/locuslink:9861’]
```

Expected Output



Suggested Structure

We *suggest* to write five functions:

1. A function that takes a threshold, reads the two `tsv` files, and returns a single `DataFrame` with the contents of the two files (aligned based on the `Interaction` ID).
   Interactions whose score is below the threshold should be discarded.
2. A function that takes the full `DataFrame` with the interactions (interactor pairs), and returns a new `DataFrame` with two columns: the interactors, and the corresponding confidence.
   Hint: you can rename the columns of a `DataFrame` with `df.columns = ['column1', 'column2', ...]`.
   Hint: to concatenate vertically two `DataFrame`'s, use the `pandas.concat([df1, df2])` function.

Suggested Structure

We *suggest* to write five functions:

1. A function that takes a `DataFrame` with the interactors and the confidences, and returns a new `DataFrame` with the interactors and the number of interactions in which they participate. (Duplicates should be counted.)
2. A function that takes a `DataFrame` of interactors and interaction counts, as well as an organism ID, and plots a histogram. The number of bins should be chosen appropriately.
3. A function that implements the program.

Instructions

Send a copy of the `.py` file to my email address:

```
teso _AT_ disi _DOT_ unitn _DOT_ it
```

with the subject:

“`sciprog midterm {your name} {your matricola}`”

Send the program as an attachment to the mail.