Software Applications 2019/2010

Project Specification

Andrea Giovanni Nuzzolese

Input

The file gene_table.txt contains summary annotation on all human genes, based on the Ensembl annotation[1]

For each gene, this file contains:

- gene_name based on the HGNC nomenclature[2];
- gene_biotype for example protein_coding, pseudogene, lincRNA, miRNA etc[3];
- chromosome on which the gene is located;
- strand on which the gene is located;
- transcript_count the number of known isoforms of the gene.

The incipit of the file is the following:

```
gene_name,gene_biotype,chromosome,strand,transcript_count
TSPAN6,protein_coding,chrX,-,5
TNMD,protein_coding,chrX,+,2
DPM1,protein_coding,chr20,-,6
SCYL3,protein_coding,chr1,-,5
C1orf112,protein_coding,chr1,+,9
...
```

[1] http://www.ensembl.org/index.html

[2] http://www.genenames.org/

[3] A more detailed description of the biotypes can be found at http://vega.sanger.ac.uk/info/about/gene_and_transcript_types.html

General Goal

Write a program able to:

1. parse the dataset;
2. analyse the data;
3. compute relevant outcomes based on arithmetic and statistical operations;
4. present the user with relevant outcomes.

The program must be designed and implemented in Python by using the Object-Oriented paradigm. Additionally, the dataset management and presentation must rely on Pandas and Flask, respectively.

Detailed Instructions

The program is composed of three main parts, consisting of:

Part 1. the part with classes and their associated methods for:

a. reading the dataset;
b. providing a registry of analytical operations that can be performed on the dataset;
c. coordinating the analytical operations with respect to user inputs received from the user interface (cf. Part 3). Those operations are only coordinated by this part. In fact, the actual execution of such operations is delegated to another part (cf. Part 2).

Part 2. the part with classes and their associated methods for:

a. analysing the data with respect to a specific objective. Each class creates a snapshot of the dataset, thus it implements a specific objective. The following are the objectives that the classes should implement:
    i. recording the numerical metadata consisting of the number of rows and columns of the dataset;
    ii. recording the general semantics of the dataset, i.e. the labels of the columns;
    iii. recording the number of genes for each biotype. The list should be sorted in ascending order;
    iv. recording, given a certain biotype as input, the list of associated genes;
    v. recording the number of chromosomes in the dataset;
    vi. recording the number of genes for each chromosome. The list should be sorted in ascending order;
    vii. recording, for each chromosome, the percentage of genes located on the + strand;

viii. recording, for each chromosome, the percentage of genes located on the - strand.

The classes that implement the objectives should be designed by using inheritance appropriately.

Part 3. the part that implements the Web-based user interface (UI). Such a UI provides a list of choices, where each choice enables an analytical objective (cf. Part 2). However, the interaction between the UI and the analytical objectives is always mediated by the software Part 1. The list of choices can be represented as a list of hyperlinks. Each hyperlink is associated with a dedicated view (i.e. a Web page) that shows the outcomes of the analytical operation requested by the user. Again the dedicated view provided by the Web page does not communicate directly with the Part 2, but it communicates with the Part 1. Accordingly the Part 1 is responsible for (i) forwarding the request to the appropriate class/method of Part 2 and (ii) returning the result to the view.

The three parts should be implemented as three separate components, i.e. three Python modules consisting in three separate files.
The software must be described into a project document by using UML diagrams in order to point out what are:

- the use cases;
- the structure of classes and components;
- the behaviour of activities.