ENGR 102 – Programming Practice
Mini Project 3
Fall 2018

Tags: Tkinter, GUI Widgets, Layout, Clusters

Analyzing restaurants data may allow discovering (i.e., clusters) restaurants which provide similar menus and have similar ratings. Likewise, foods may be grouped together based on the restaurants' menus. In this mini project, you are going to develop a clustering analysis tool to help users better understand and compare restaurants meals.

How should it look like?

Your program will have a graphical user interface (GUI) which will function and look like as explained below.

1. Firstly, the GUI will be small and show the main buttons as shown in **Figure 1.**

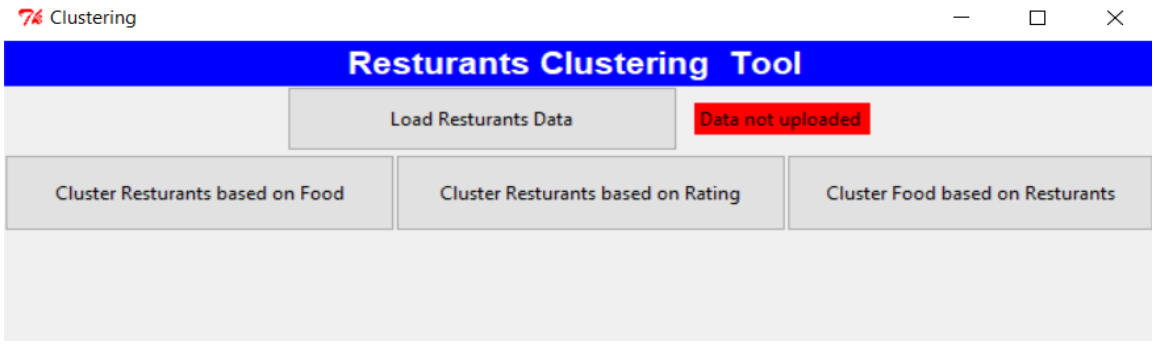

Figure 1. Main GUI window.

2. Once the user loads the data, the GUI window will expand as below**.**

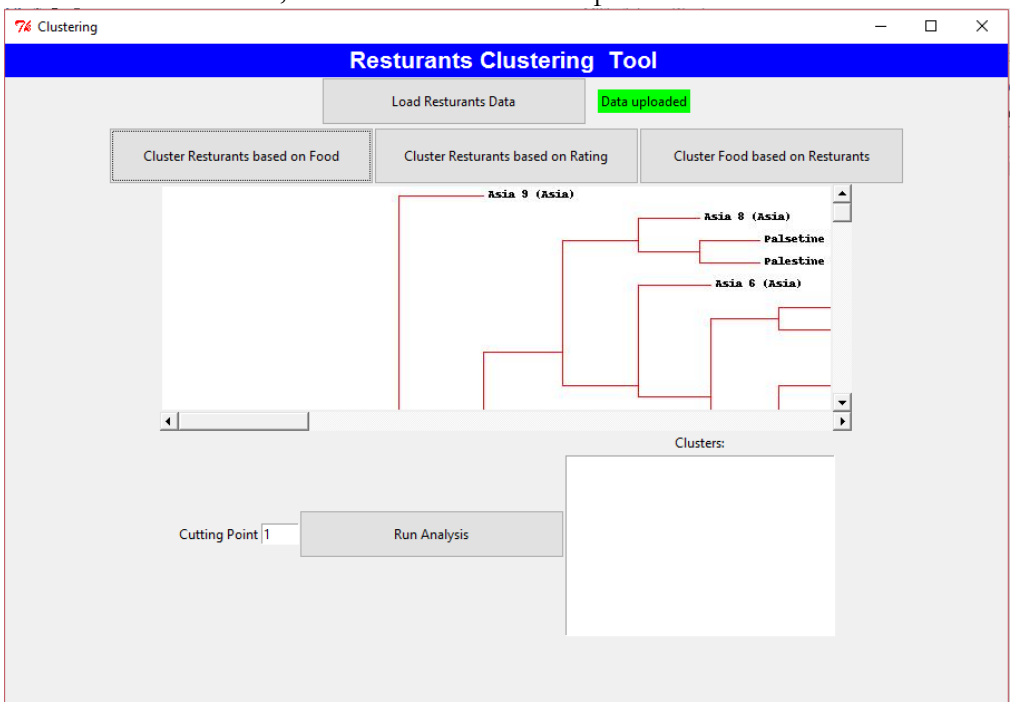

Figure 2. GUI window expanded once the data is uploaded.

How should it work?

1. At the beginning, the user will load the data by pressing on **Load Restaurants Data** button. The user will be able to choose the data from a file dialog window as shown in **Figure 3.** The data is provided to you in a folder called ***data*** which consists of excel files that shows different restaurants, each file having the restaurant's each meal's code, name and meal health rating. The rating, in each file, given in column D is the overall rating for the restaurant itself.

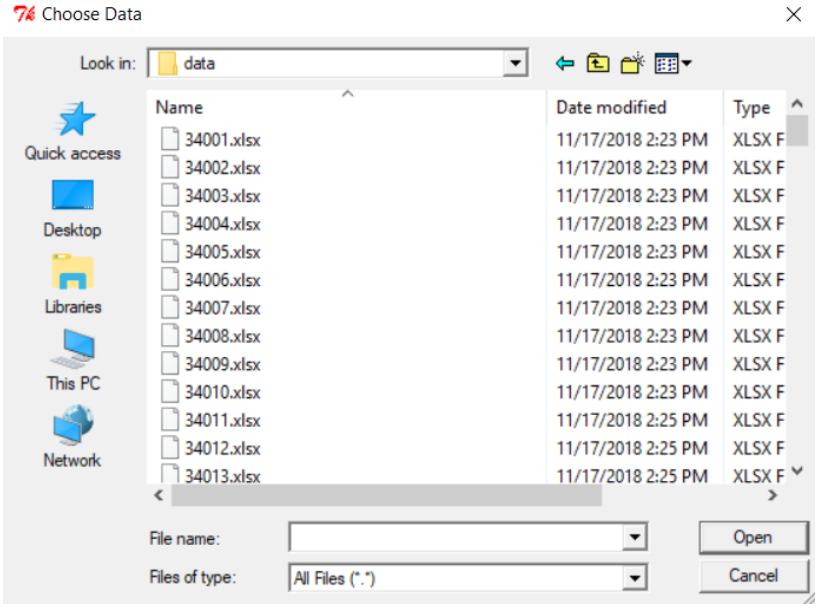

Figure 3. File dialog window

2. After the user loads the data, the label in red next to the loading button which indicates **Data not uploaded** should turn into green indicating **Data uploaded.**
3. As a next step, the user will trigger the hierarchical clustering analysis based on food, ratings or restaurants by pressing one of the three buttons (Figure 4). For example, if the user chooses the **Cluster Restaurants based on Food** button, then the GUI window will expand to show a dendogram, on a Canvas widget, of the clustered data based on food (**Figure 4).**
4. **Note:** in cluster step you should use the **clusters.py** file, which is also provided to you. You will find the Hierarchical clustering method and the dendrogram drawer as methods in that file.

Detailed explanation:

Clustering restaurants based on Food: This feature will allow users to identify restaurants that have the same menu offered in their food selections. You may build a matrix representation, similar to blogs example, where rows represent restaurants and columns represent foods. An individual cell at row i and column j may be populated with 1 if restaurant represented by row i has the food represented by column j. Otherwise, this cell may be populated with 0. When the ***Cluster Restaurants based on Food*** button is clicked, clustering task will be initiated by using hierarchical clustering, and the resulting dendrogram will be displayed on a Canvas widget below the button

(**Figure 4**). Note that labels in this dendrogram include both restaurants' names and their country code in parenthesis (e.g. Turkish 9 (Turk)). As a part of this project, you are already provided (in clusters.py file) with functions that generate an image file for a dendogram. You may consider displaying this image on a canvas with horizontal and vertical scroll bars.

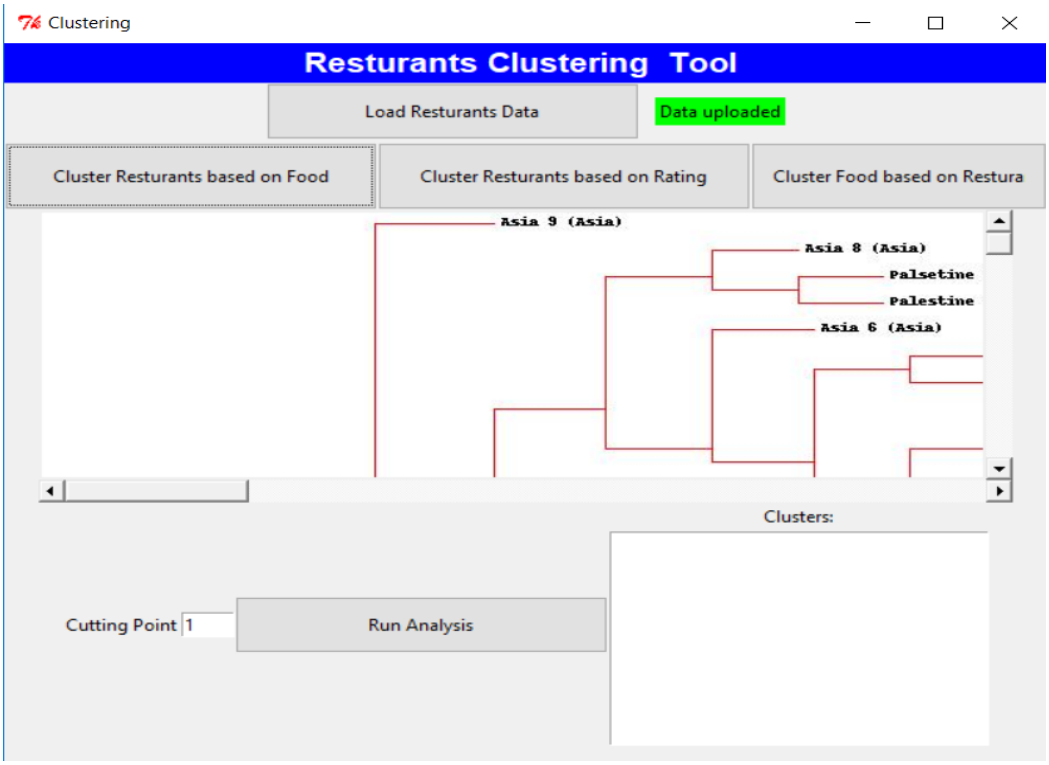

Figure 4. GUI when ***Cluster Restaurants based on Food*** button is clicked.

Below the dendrogram, there will be a set of widgets that will allow the user to analyze the clusters at different levels. ***Cutting point*** specifies the level at which the hierarchical clustering tree will be "cutoff".

More specifically, consider the following example: dendrogram (Figure 5) with the actual data points represented by letters A through H. Table 1 shows the list of clusters that will be included in the analysis for some values of cutting levels/points.

When hierarchical clustering is represented by a dendrogram, cutting a dendrogram (refer to Figure 5) that shows clusters levels at a certain level gives a set of clusters. Cutting at another level gives another set of clusters. Depending on our cutting level we get the clusters.

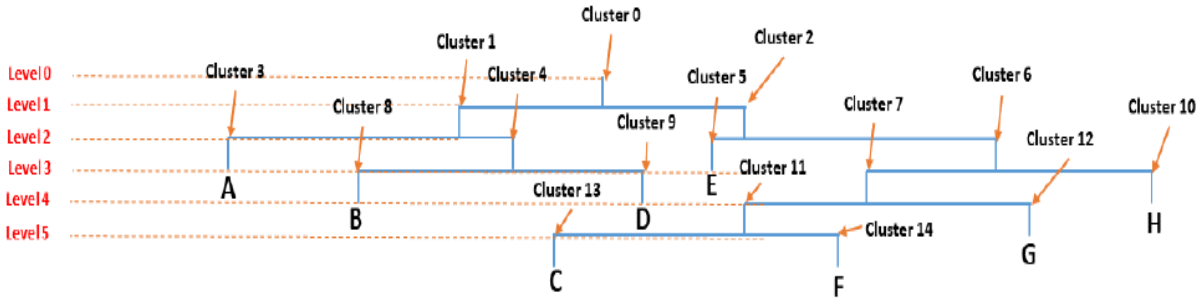

Figure 5. An example dendogram with different cutting levels.

Cutting level/point	Included Clusters
0	Cluster 0
1	Cluster 1, Cluster 2
2	Cluster 3, Cluster 4, Cluster 5, Cluster 6
3	Cluster 3, Cluster 8, Cluster 9, Cluster 5, Cluster 7, Cluster 10
4	Cluster 3, Cluster 8, Cluster 9, Cluster 5, Cluster 11, Cluster 12, Cluster 10
5	Cluster 3, Cluster 8, Cluster 9, Cluster 5, Cluster 13, Cluster 14, Cluster 10

Table 1. Clusters of the dendogram in Figure 4 with different cutting levels.

When the user clicks on the ***Run Analysis*** button, a listbox will list all the included clusters depending on what the user specifies **(Figure 6).** Assume that there are n included clusters. Then, the clusters will be numbered from 0 to n-1 starting from left-most one to the right-most one in the hierarchical tree. For instance, assume that our clustering hierarchy is the one in **Figure 5**, and the cutting level is 2. Then, the included clusters would be: Cluster 3, Cluster 4, Cluster 5, and Cluster 6. In the listbox, these clusters will be renamed and numbered starting from 0 as follows: Cluster 0, Cluster 1, Cluster 2, and Cluster 3. Besides, next to the cluster name, in parenthesis the following pieces of information will be listed as well: (i) the restaurant Type (ASIA, TURK, AMR and PAL) with ***most*** number of restaurants under that cluster, (ii) the number of restaurants under that cluster, and (iii) the accuracy of clustering which is computed as the ratio of the restaurants that belong to the most popular restaurants type under that cluster. For instance, assume that a cluster contains 10 restaurants in total (2 from ASIA, 3 from TURK, and 5 from PAL). In this cluster, the most popular restaurant type is PAL, and the accuracy is computed as (number of PAL restaurants / number of all restaurants in this cluster) which is 50%. The restaurants under a cluster includes all those leaf level (i.e., nodes without any children) data points. For instance, in Figure 4, cluster 0 includes all data points from A to H. Cluster3 includes only A, while cluster 2 includes E, C, F, G, and H.

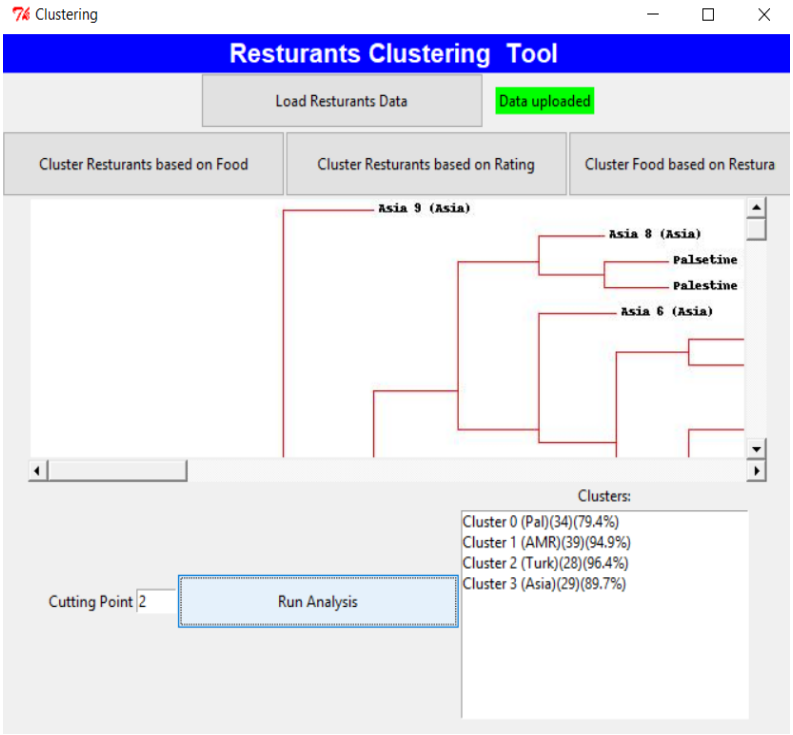

Figure 6. Run Analysis button clicked with ***Cluster Restaurants based on Food*** button clicked with cutting point of 2.

Clustering restaurants based on their rating:

This feature will allow users to identify restaurants that have similar ratings. You may build a matrix representation, similar to blogs example, where rows represent the restaurants and columns represent food meals. An individual cell at row i and column j may be populated with the meal rating, if the restaurant represented by row i has the meal represented by column j. Otherwise, this cell may be populated with 0. When the middle button **(*Cluster Restaurants based on Rating*)** is clicked, the clustering task will be initiated by using hierarchical clustering, and the resulting dendrogram will be displayed on a Canvas widget **(Figure 7)**. Note that labels in this dendrogram include both restaurants names and their rating (in parenthesis) that can be obtained from the data files.

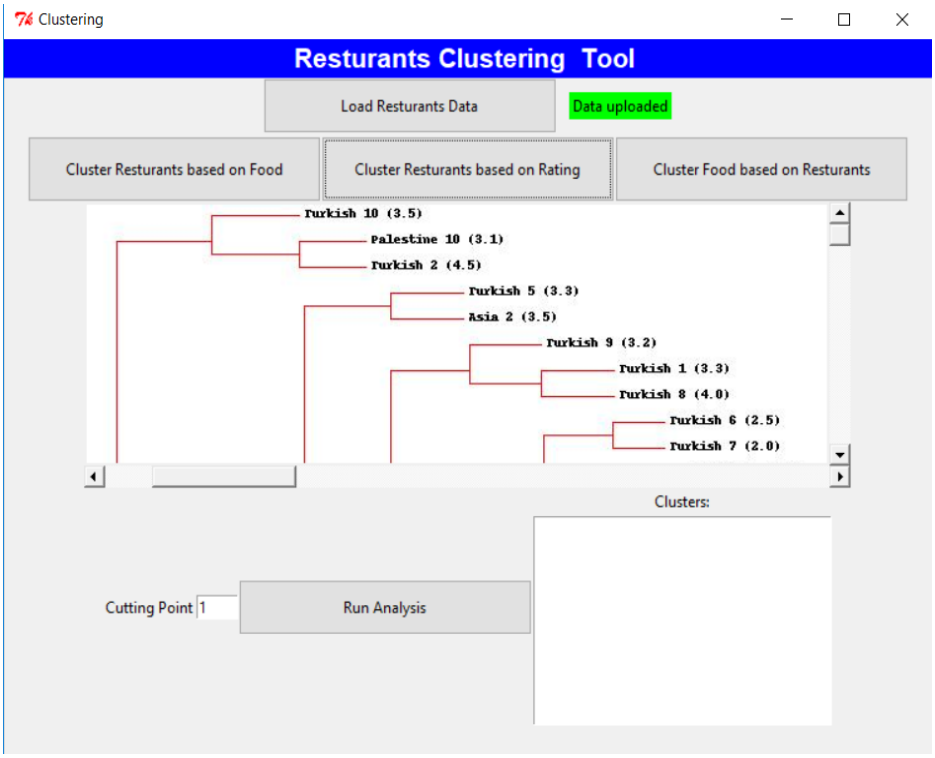

Figure 7. GUI when ***Cluster Restaurants based on Rating*** button is clicked.

In the analysis part, the user will be able to investigate whether clusters represent high Rating (>=4,0), poor Rating(<2,50), and mediocre Rating(2,5<=Rating<3,5). That is, instead of Restaurant type-based analysis, user will perform a RATING-based analysis here. Accuracy will be computed in a similar manner based on the ratio of most popular category in each cluster as explained for the restaurant type-based analysis. An example case is provided below in **Figure 8**. Cluster 0 has 43 number of restaurants with a poor Rating in this case.

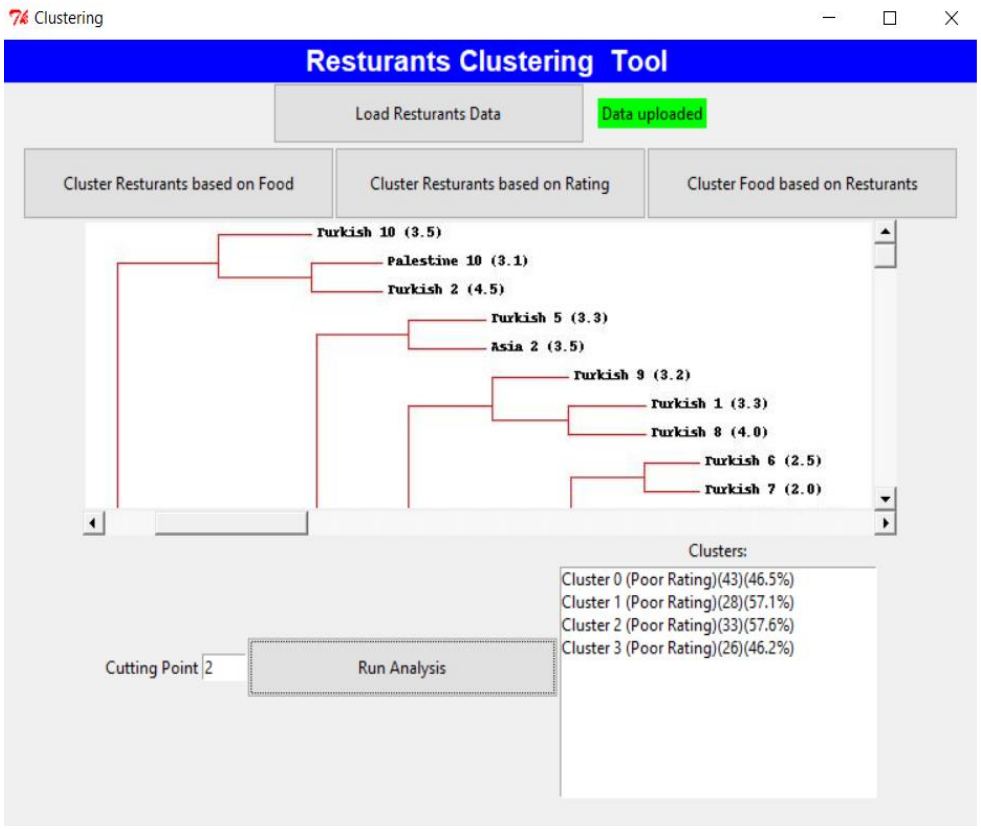

Figure 8. Run Analysis button clicked with ***Cluster Restaurants based on Rating*** button clicked with cutting point of 2.

Clustering food based on restaurants: This feature will allow users to identify food meals that are found in the restaurants' menus. You may build a matrix representation, similar to blogs example, where rows represent food meals and columns represent restaurants. An individual cell at row i and column j may be populated with 1, if the restaurant represented by column j has the meal represented by row i. Otherwise, this cell may be populated with 0. When the right-most button (***Cluster Food based on Restaurants***) is clicked, the above clustering task will be initiated by using hierarchical clustering, and the resulting dendrogram will be displayed on a Canvas widget **(Figure 9).**

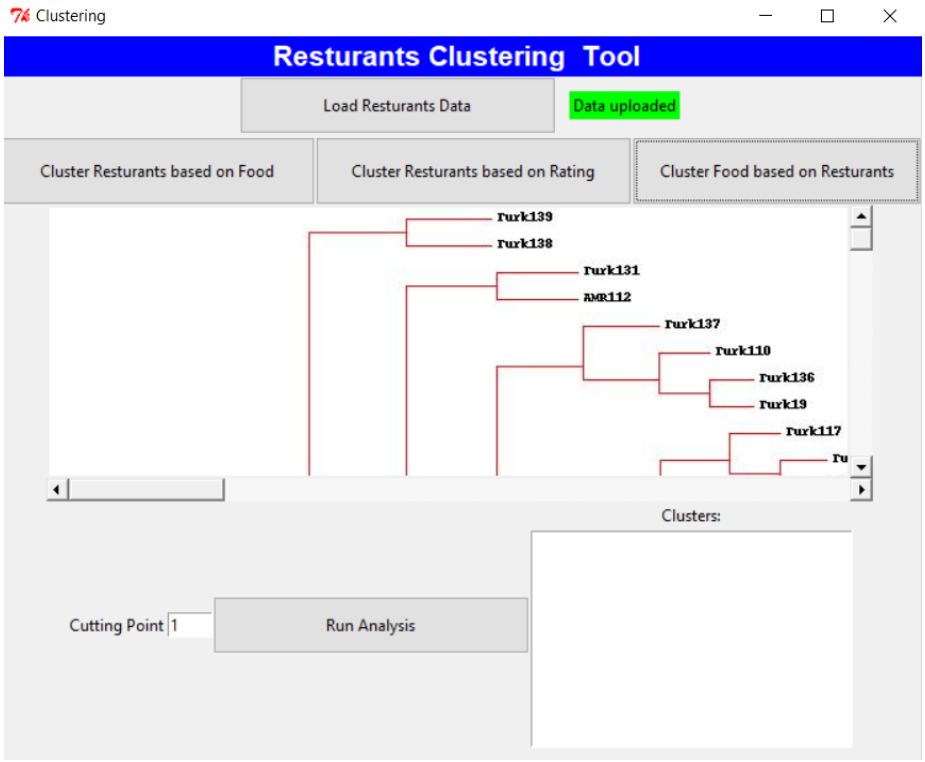

Figure 9. GUI window when the ***Cluster Food based on Restaurants*** button is clicked

The analysis part **(Figure 10)** will be very similar to the one in clustering restaurants based on meals. The only difference here would be that instead of restaurants, meals will be listed as cluster members. For each meal, the restaurants type information will be obtained through the restaurants who have the meal. A sample report is provided in Figure 10.

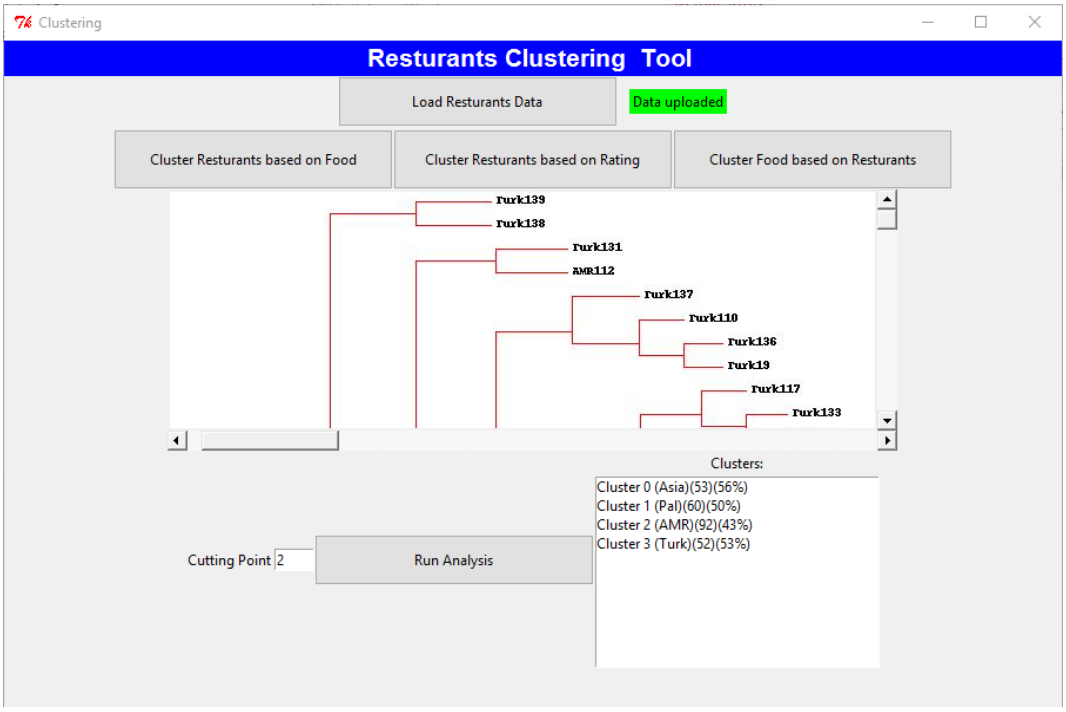

Figure 10. Dendogram and Clustering analysis when the ***Cluster Food based on Restaurant*** button is clicked

Can you provide any further pointers that may be helpful? :

Implementation Notes / Hints:

- **You need to use at least 3 classes in this project.**
- **In order for you to understand better how the GUI should work, please check the GIF image that will be provided with the project documents.**
- Take a look at the data before starting in order for you to have an insight of how to work with data and give the right variables to the clustering function that are in **clusters.py**.
- The following link contains examples of using **Canvas:**

  http://www.tutorialspoint.com/python/tk_canvas.htm
- You may use **xlrd** module to read Excel files (install it on PyCharm in the same way you did the other modules). Please see the following reference:

  https://www.blog.pythonlibrary.org/2014/04/30/reading-excel-spreadsheets-with-python-and-xlrd/
- The following link contains examples of using tkFileDialog.

  https://pythonspot.com/tk-file-dialogs/
- When you visualize your results, you may use pillow:

  https://pypi.org/project/Pillow/
- To add a scrollbar on canvas, the following example may be useful:

  https://stackoverflow.com/questions/7727804/tkinter-using-scrollbars-on-a-canvas

Warnings:

- You **CANNOT** use place for geometry, only grid and pack are allowed.
- Do not talk to your classmates on project topics when you are implementing your projects. Do not show or email your code to others. If you need help, talk to your TAs or myself, not to your classmates. If somebody asks you for help, explain them the lecture slides, but do not explain any project related topic or solution. **Any similarity in your source codes will have serious consequences for both parties**.
- Carefully read the project document, and pay special attention to sentences that involve "should", "should not", "do not", and other underlined/bold font statements.
- If you use code from a resource (web site, book, etc.), make sure that you reference those resource at the top of your source code file in the form of comments. You should give details of which part of your code is from what resource. Failing to do so may result in **plagiarism** investigation. Last but not the least, you need to understand code pieces that you may get some other resources. This is one of the goals of the mini projects.

Even if you work as a group of two students, each member of the team should know every line of the code well. Hence, it is important to understand all the details in your submitted code.

How and when do I submit my project?

- Projects may be done individually or as a small group of two students (doing it individually is **strongly** recommended for best learning experience). If you are doing it as a group, only **one** of the members should submit the project. File name will tell us group members (Please see the next item for file naming details).
- Submit your own code in a single Python file. Name it with your and your partner's first and last names. As an example, if your team members are Deniz Barış and Ahmet Çalışkan, then name your code file as deniz_baris_ahmet_caliskan.py (Do not use any Turkish characters in file name). If you are doing the project alone, then name it with your name and last name similar to the above naming scheme.
    - Those who do not follow the above naming conventions will **get** 10% **off** of their project grade.
- You are given two weeks instead of two weeks to work on this project. Submit it online on LMS by **17:00 on 12th December, 2018, Wednesday**.

    Late Submission Policy:

    - -10%: Submissions between 17:01 – 18:00 on the due date
    - -20%: Submissions between 18:01 – midnight (00:00) on the due date
    - -30%: Submissions after which are up-to 24 hours late.
    - -50%: Submissions which are up-to 48 hours late.
    - Submission more than 48 hours late will not be accepted.

Grading Criteria?

GUI Design	Reading and Parsing Excel File and uploading data properly	Hierarchical clustering for 3 different cases	Cut Off Analysis
(15)	(15)	(45)	(25)

Your code should be efficient, easy to follow and track. Therefore, from your overall grade, we will deduct points by the specified percentage for the following items:

- Inappropriate/Cryptic variable names and method names (10%)
- Classes and objects are not used properly (30%)
- Insufficient commenting (10%).

Have further questions?

If you need help with anything, please use the office hours of your TAs and the instructor to get help. **Do not walk in randomly (especially on the last day) into your TAs' or the instructor's offices. DO NOT LEAVE YOUR QUESTIONS TO THE LAST MINUTE. Make an appointment first. This is important. Your TAs have other responsibilities. Please respect their personal schedules.**

IMPORTONT NOTES:

Note 1: Plagiarism:

- Zero tolerance
- Cases will be referred to the Ethics Committee
- Both parties (provider and receiver) are responsible
- Process:
    - Automated computerized checks for pre-filtering
    - Human review for confirmation
    - Referral to the Ethics Committee if true positive