

CS5590 APS -Python Programming-LAB1

Mohamed Hamdalla. Class ID:15



October 13, 2019

UMKC

**Introduction**

Python is a broadly utilized universally useful, elevated level programming language. It was at first planned by Guido van Rossum in 1991 and created by Python Software Foundation. It was fundamentally produced for accentuation on code meaningfulness, and its linguistic structure enables software engineers to express ideas in less lines of code.

Python is a programming language that gives you a chance to work rapidly and incorporate frameworks all the more proficiently.

python has many advantages that includes, but not limited to, that it’s an open source and network improvement and can support third-party modules. It’s also portable across operating systems.

It can be evoked for many applications includes GUI based work area applications (Games, Scientific Applications), web structures and applications, and enterprise and business applications.

All the world-wide organization such as Google, Yahoo, YouTube, and so many other organizations uses python.

**Objectives:**

\*In this project Python will be evoked to perform multiple achievement by applying the simple basic statements of python to do essential operations for the end user such as:

1. finding the longest substring without repeating characters from a given string.
2. Combine the grades of all the subjects for different students in one file in an organized manner.

\*Then a python code that uses classes to simplify the lengthy codes will be developed for the library management system.

**Approaches**

In this project multiple approaches will be used for different tasks. The advanced approached that will be used include:

\* Web scraping, which is one of the features of python will be used to fetch a table from a given link.

\* Machine leaning tools, Nearest Neighbor (k-nearest neighbors (kNN)), Naive Bayes Classifier, and Support Vector Machines (SVM) will be also used in this project for classifications and clustering different features of a given data sets.

\* Natural language processing, Natural Language Toolkit (NLTK), will be used to show the most important words and sentences of a given text.

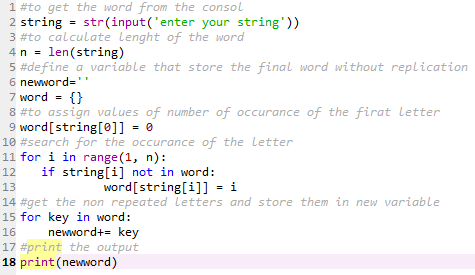
**Workflow and Code Evaluation**

As I mentioned in the previous sections, multiple tasks, 8 tasks, will be done in this project. In this section I will present each task individually and its output. The tasks performed in this project is as follows:

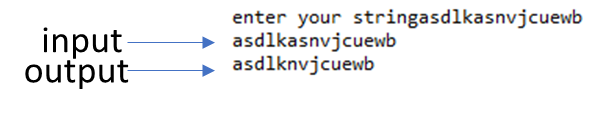
1. Write a program to find a longest substring without repeating characters from a given string input from the console.

Sample Input: ‘ababcdxa’ / Sample Output: abcdx

The code starts by asking the user to enter the word, calculate number of characters in the word, then define a new variable. Check of the character excise in the variable, if yes don’t repeat it other wise append it to the variable. Finally print the output as shown in Fig.

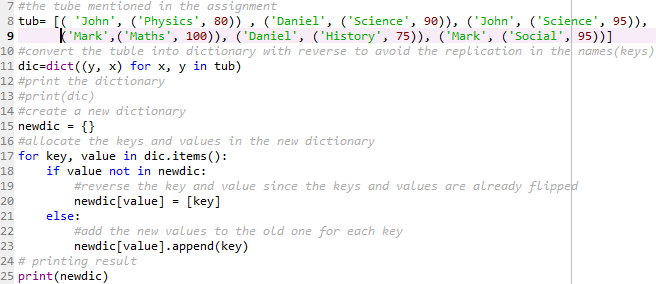


If you run the code it will ask you to enter the string and shows the output as shown in the Fig.



1. Suppose you have a list of tuples as follows:[( ‘John’, (‘Physics’, 80)) , (‘ Daniel’, (‘Science’, 90)), (‘John’, (‘Science’, 95)), (‘Mark’,(‘Maths’, 100)), (‘Daniel’, (’History’, 75)), (‘Mark’, (‘Social’, 95))]Create a dictionary with keys as names and values as list of (subjects, marks) in sorted order.{John : [(‘Physics’, 80), (‘Science’, 95)]Daniel : [ (’History’, 75), (‘Science’, 90)]Mark : [ (‘Maths’, 100), (‘Social’, 95)]}

The code starts by define the tuble that has the data mentioned in the task. The tuble then is converted into dictionary. Another dictionary is created, and the key and values of the first dictionary is allocated in the second one without replication. The final dictionary is then printed to show the results.

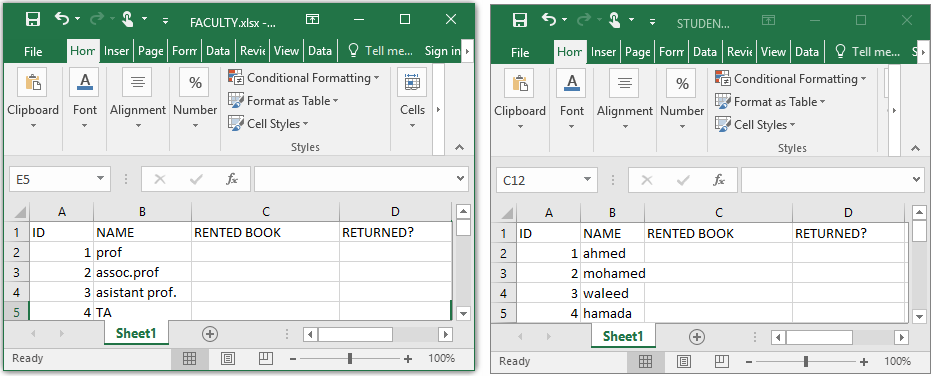


The output is organized as mentioned in the task and shown in the following Fig.



1. Write a python program to create a library Management System (eg: Student, Book, Faculty, Department etc.)

In this task I will use classes to divide the task into simpler tasks. I will have student and faculty classes. I will also have a class for the library that contains subclasses for borrowing and returning the books. The code will have two excel sheets containing the person (student or faculty) ID and Name. the code will ask the user for his id and check the name of the person of it is exist in the data base of the library (excel sheets). Once it found him it will ask him to choose the book to borrow or if he already has browed book it will ask him if he wants to return it. The excel sheets will be updated accordingly. The excel sheets looks as shown in following Fig.

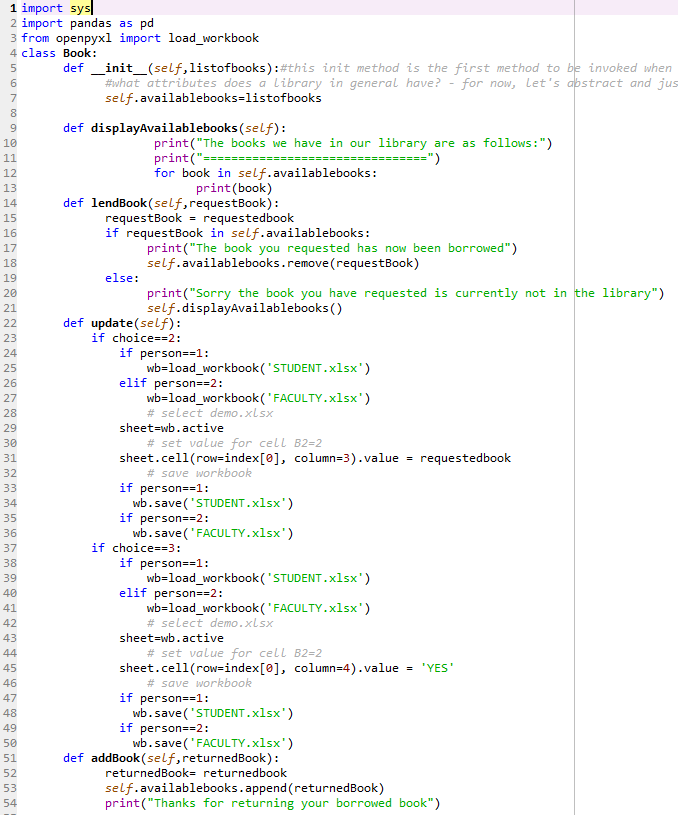


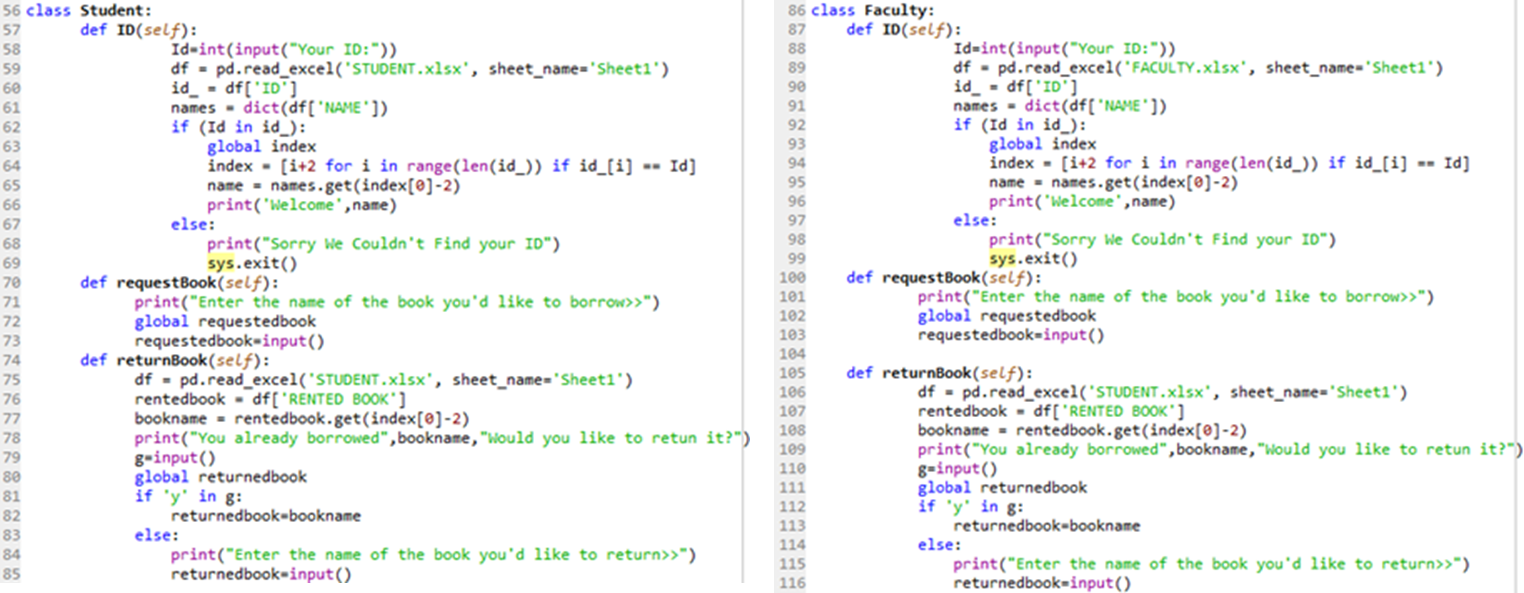
The main code is shown in the following Fig.



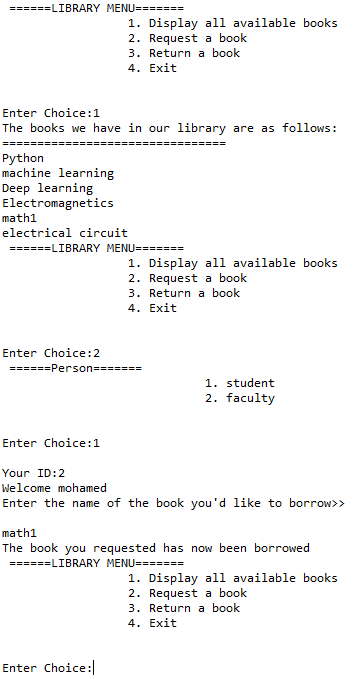
In the first line the we store all the possible books to the library. Then will have the while loop to keep the program running until the user terminate it.

Once the user chose his category, the code will call the classes needed. The used classes (three class) in this code is shown in the following Fig.

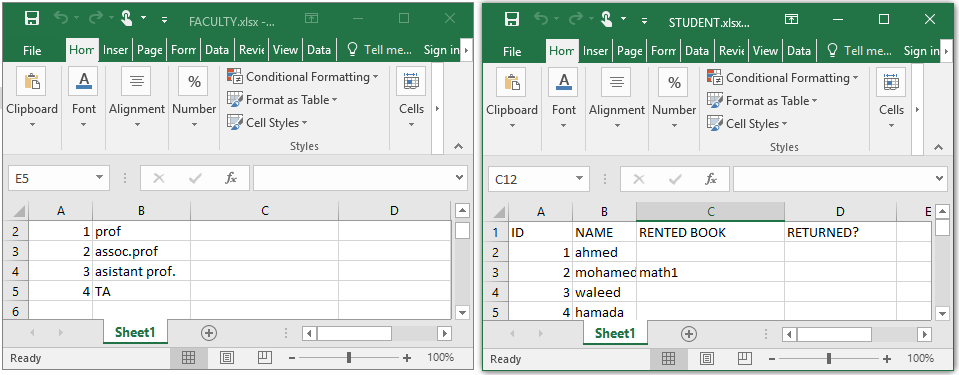




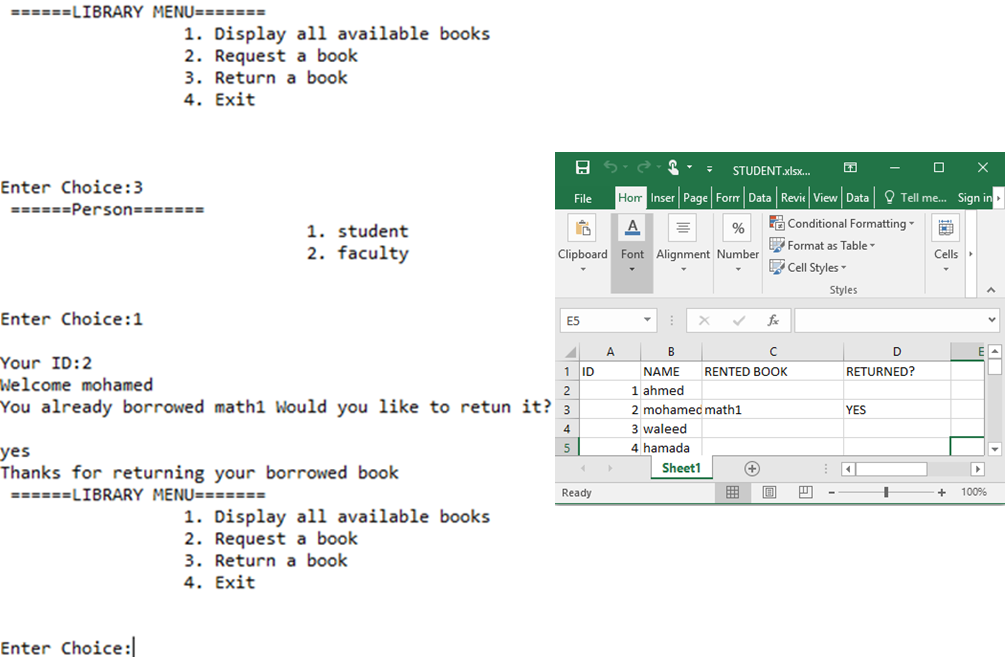
Once you run the code it will ask some question then it will allow the user to borrow or return a book as shown in the following Fig.



The excel sheet of the STUDENTS should update the status of the student whose name is Mohamed and his ID is 2. The bowered book should be mentioned in his history as shown in the following Fig.



If we continue the code to return the book, the following process will happen.

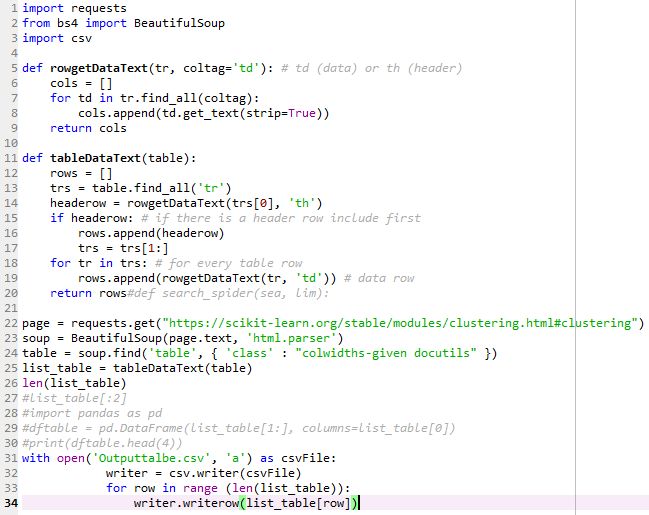


1. Go to https://scikit-learn.org/stable/modules/clustering.html#clusteringand fetch comparison of the clustering algorithms in scikit-learn.

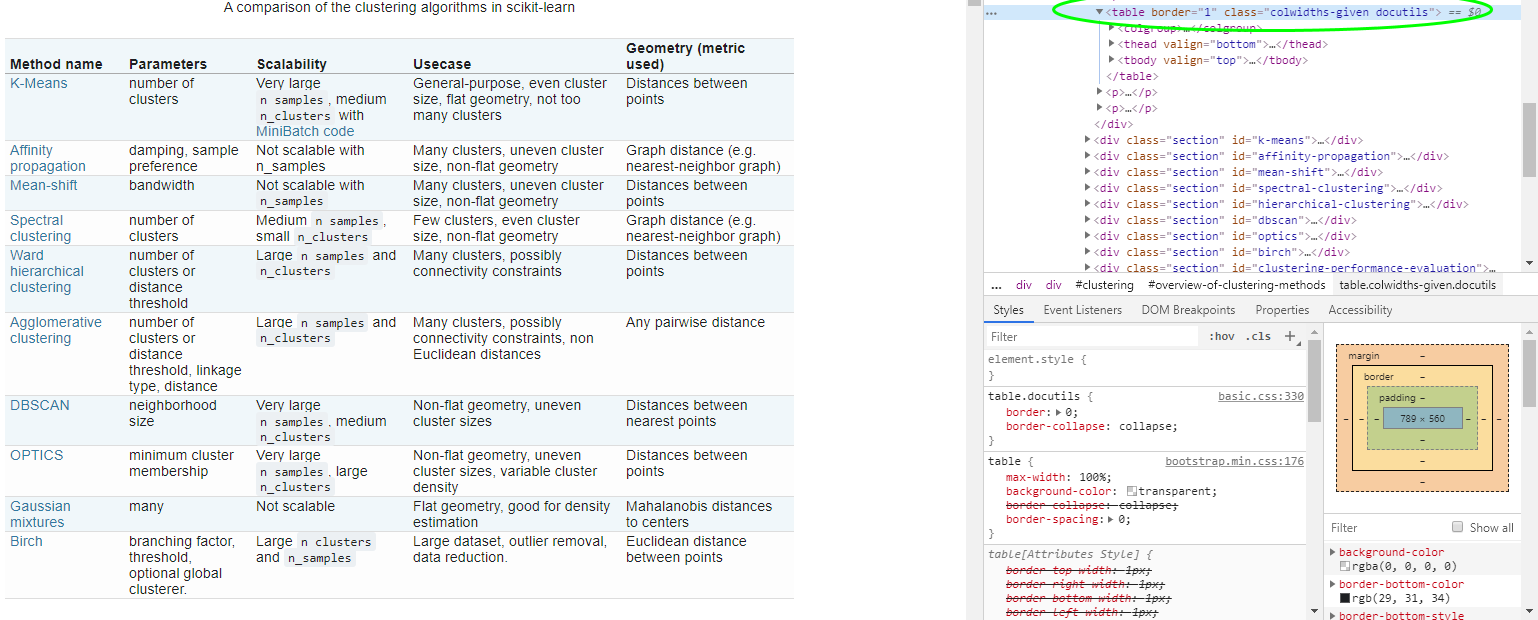
Hint: Use BeautifulSoup package.

As mentioned in the task I will use BeautifulSoup package to read the data of a comparison table from a URL and store the data into excel sheet.

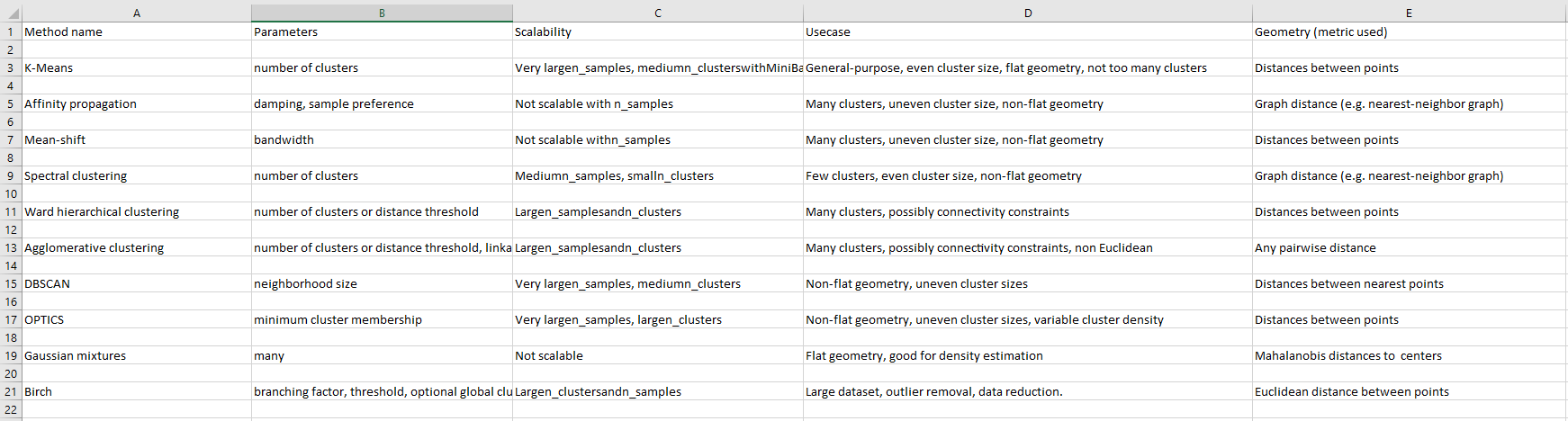
The code of that do this task is shown in the following Fig. where two functions were used to read the rows and columns of the table.



The class of the table was extracted form the webpage provided in the task as shown in the following Fig.



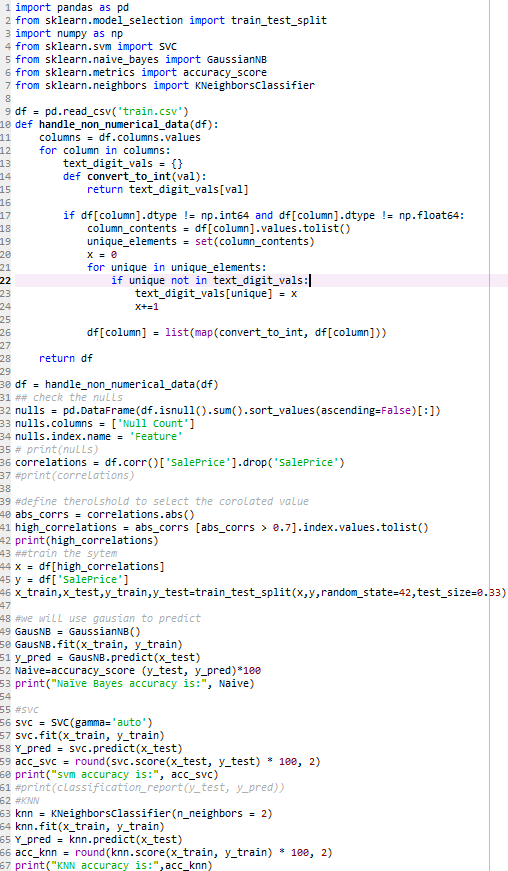
The output excel file is shown in the following Fig.



1. Pick any dataset from the dataset sheet in the class sheet or online which includes both numeric and non-numeric features a. Perform exploratory data analysis on the data set (like Handling null values, removing the features not correlated to the target class, encoding the categorical features, ...)

b. Apply the three classification algorithms Naïve Bayes, SVM and KNN on the chosen data set and report which classifier gives better result.

The following code will have a function that convert the nonnumeric data in the data set onto numeric and apply all the requited classifiers and show the accuracy of each classifier. The data set used in this task is the same data set used in the class for the house prices.

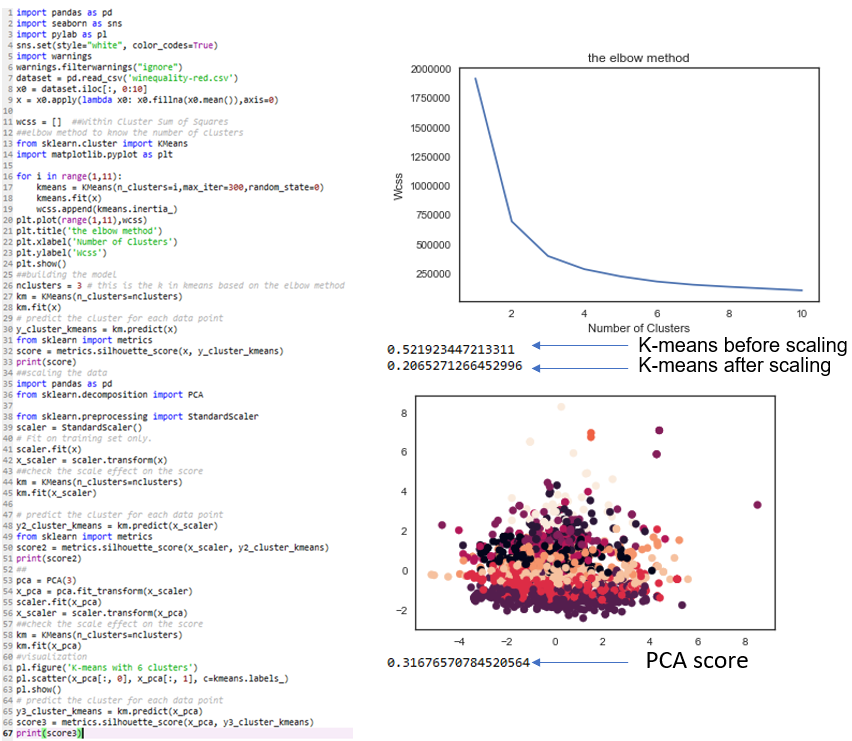


The output of the code shows the most correlated features to the target and shows the accuracy of each classifier. SVM has the best accuracy among the other classifiers.



1. Choose any dataset of your choice. Apply K-means on the dataset and visualize the clusters using matplotlib or seaborn. A. Report which K is the best using the elbow method.

b. Evaluate with silhouette score or other scores relevant for unsupervised approaches (before applying clustering clean the data set with the EDA learned in the class)

in this code I used the red-wine quality dataset provided in the class. I excluded the target of this data set to change the problem to unsupervised clustering. The following Fig. shows the code and the results. According to elbow method, I chose K=4 to have 4 clusters which is almost the same clusters in the target column in the red wine data set.

1. Write a program in which take an Input file, use the simple approach below to summarize a text file:Link to input file: https://umkc.box.com/s/7by0f4540cdbdp3pm60h5fxxffefsvrwa. Read the data from a file

b. Tokenize the text into words and apply lemmatization technique on each word.

c. Find all the trigrams for the words.

d. Extract the top 10 of the most repeated trigrams based on their count.

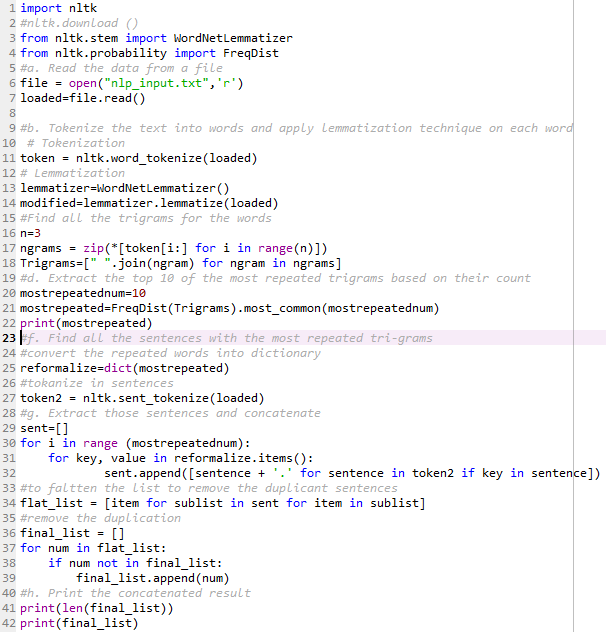
e. Go through the text in the file

f. Find all the sentences with the most repeated tri-grams

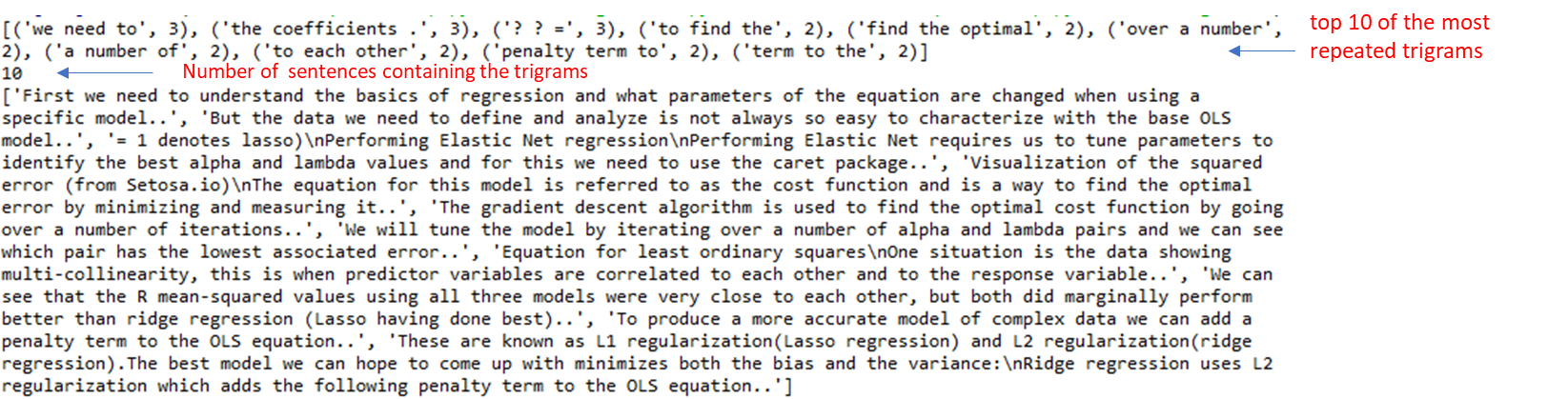
g. Extract those sentences and concatenate

h. Print the concatenated result

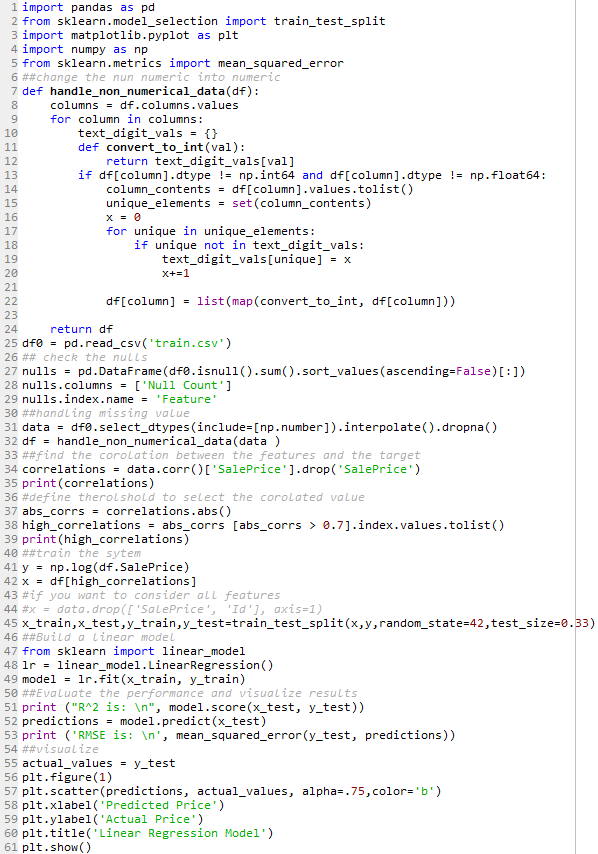
the following code will do the requested tasks one by one and print the results as follows:



The output is as follows:



1. Create Multiple Regression by choosing a dataset of your choice (again before evaluating, clean the data set with the EDA learned in the class). Evaluate the model using RMSE and R2.



The output of this task is shown in the following Fig.:

