Georgios Karagiannis | H00226912

gk13@hw.ac.uk

Abstract

The aim of this coursework was to develop a simple Python-based application, which analyses and displays document tracking data from a major web site.



Data Analysis of a Document Tracker

Assessed Coursework 2 - Industrial Programming (F21SC)

# Table of Contents

[Table of Contents - 1 -](#_Toc437198097)

[1. Introduction - 2 -](#_Toc437198098)

[2. Requirements’ checklist - 3 -](#_Toc437198099)

[3. Design Considerations - 4 -](#_Toc437198100)

[4. User Guide - 5 -](#_Toc437198101)

[Countries screenshot - 6 -](#_Toc437198102)

[continent screenshot - 6 -](#_Toc437198103)

[browsers screenshot - 7 -](#_Toc437198104)

[browsers name screenshot - 7 -](#_Toc437198105)

[avid users screenshot - 8 -](#_Toc437198106)

[5. Developer Guide - 9 -](#_Toc437198107)

[Classes - 9 -](#_Toc437198108)

[class MyGui - 9 -](#_Toc437198109)

[class MyMethods - 9 -](#_Toc437198110)

[My Main Program - 10 -](#_Toc437198111)

[UML Class Diagram - 12 -](#_Toc437198112)

[6. Testing - 13 -](#_Toc437198113)

[Errors I found out - 13 -](#_Toc437198114)

[7. Reflection on programming language and implementation - 14 -](#_Toc437198115)

[Libraries - 15 -](#_Toc437198116)

[8. Coursework and Course Conclusion - 17 -](#_Toc437198117)

[Appendix - 18 -](#_Toc437198118)

# 1. Introduction

The aim of this coursework was to develop a simple, data-intensive application using the scripting, object-oriented language Python.

Python is a widely used general-purpose, scripting, object-oriented, high-level programming language. Its design philosophy emphasizes code readability, and its syntax allowed me to express concepts in fewer lines of code than would be possible in languages such as C#. The language provides constructs intended to enable clear programs on both a small and large scale.

The learning objective of this coursework for me, was to develop proficiency in advanced programming concepts, stemming from both object-oriented and functional programming paradigms, and to apply these programming skills to a concrete application of moderate size like this.

Python supports multiple programming paradigms, including object-oriented, imperative and functional programming or procedural styles. It features a dynamic type system and automatic memory management and has a large and comprehensive standard library. Python interpreters are available for installation on many operating systems, allowing Python code execution on a wide variety of systems.

Because of this coursework I developed personal abilities in using Python as a “glueware” to build, configure and maintain a moderately complex application and deepen the understanding of integrating components on a Linux system.

For the implementation I used ***Python 3*** as installed on our school’s Linux lab machines and also a range of libraries. Specifically the ***json***, ***tkinter***, and ***matplot*** libraries. I used the json library for parsing, the tkinter library for GUI functionality and the matplot library for visualising the results. One other option it was to use the ***pandas*** library for processing the input data but searching about that library I notice that ***pandas*** is extremely helpful for bigger application that that application and decide not to spend time to learn how ***pandas*** works. I chose to implement the processing function myself. I used basically my MacBook for the implementation. Especially I installed Ubuntu 14.04 LTS as virtual machine and IDLE with Python 3.5. Furthermore I installed ***Anaconda*** which is a completely free Python distribution and includes more than 300 of the most popular Python packages for science, math, engineering, and data analysis. I will explain more about libraries later.

I chose to work in a Linux OS machine because I would like to become more familiar with that OS. I had to use as main tool an OS like that since my undergraduate studies in which I used Solaris which is basically a UNIX operating system.

The *issuu.com* platform is a web site for publishing documents. The web site tracks usage of the site and makes the resulting, anonymised data available to a wider audience. For this exercise, I used one of these data sets to perform data processing and analysis in Python.

Firstly I will discuss which of the requirements I have delivered and which I have manage to implement. After that I will provide the basic design of my code. Furthermore I will provide a user guide and a developer guide. Finally I will show the results of the test I have done and how Python’s features helped me to implement the application.

# 2. Requirements’ checklist

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 | Python usage | | | Done! |
| 2-a | Views by country using mathplot | | | Done! |
| 2-b | Views by continent using mathplot | | | Done! |
| 3-a | Views by browser | | | Done! |
| 3-b | Views by browser (only the main browser name) | | | Done! |
| 4 | Top 10 readers (by time spent reading documents) | | | Done! |
| 5-a | A function that takes a document UUID and returns all visitor UUIDs of readers of that document. | *Done!* |
| 5-b | A function that takes a visitor UUID and returns all document UUIDs that have been read by this visitor | *Done!* |
| 5-c | A function to implement the “also like” functionality | *Done!* |
| 5 | 5-d | An “also like” list of documents based on readership profile for sorting the documents | *Done!* |  |
| 5-e | An “also like” list of documents, using a sorting function, based on the number of readers of the same document | *Done!* |  |
|  | 5-f | A list of top 10 document UUID | *Done!* |  |
| 8 | GUI usage | | | Done! |
| 9 | Command-line usage | | | NOT! |

My program’s core logic has implemented in Python, uses libraries for input, data processing and visualisation. I used json library for parsing, the pandas library for processing the input data, the tkinter library for GUI functionality and the matplot library for visualising the results.

My application is able displays a histogram (using mathplot) of countries of the viewers, also I provide a histogram of countries by continent. Furthermore, my application displays a histogram of all browser identifiers of the viewers and also a histogram that illustrates only the main browser name. The application is simple and easy to use because has a GUI in order to perform all the operations. The only thing that I did not implement is to provide a command-line interface

# 3. Design Considerations

I will explain how I chose to design the code, how I chose to read the data and to illustrate the programs results. I will also illustrate a UML diagram of my program.

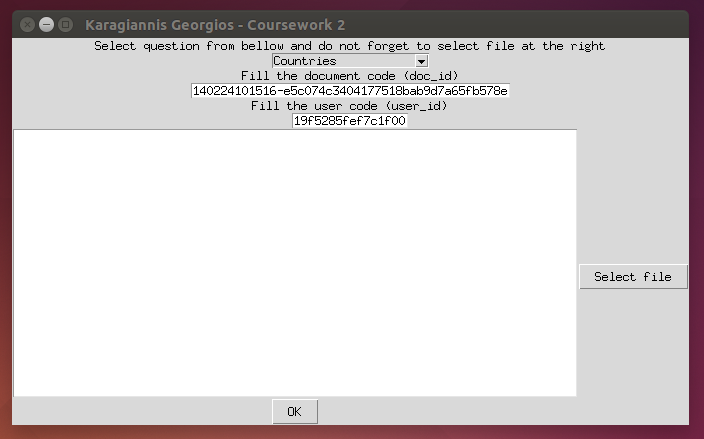
To begin with, I separate my code in two classes, one class is my GUI and the other class has the methods for the first 4 questions. The methods for the question 5 are in the main body of the program. I used the libraries that we show in the class and after a research I decide to use and some other libraries in order to achieve the requirements. The GUI is very simple. I all cases I chose to real the JSON file line by line. I tried to avoid the multiples loops as I could. I use lists and dictionaries. I tried to make my application as fast as I could in order to handle a big JSON file.

# 4. User Guide

I will provide a user guide which helps users to operate my application.

That is the main windows of my application. The user can select what he wants to see, my menu is based on questions of the coursework.

User’s computer must have all the libraries I used for the implementation



In this area the user should first select what he wants to see, the countries of the visitors, the continents etc.

After that the users must fill the document ID and the user ID (depends on the question)

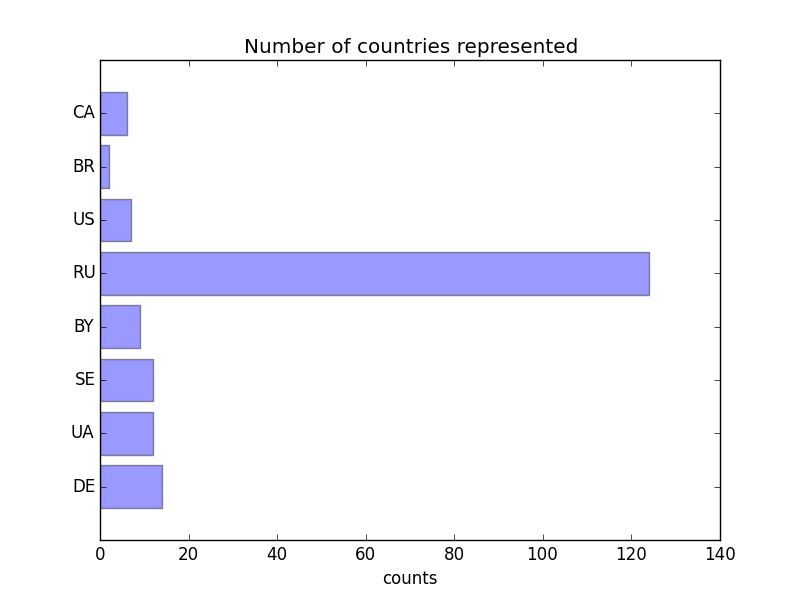
The user should select the JSON file that he wants to analyze.

At the end the user should press OK in order to see the result. Some questions have as result a histogram and some others a list with data that is illustrated at the big white area of the window

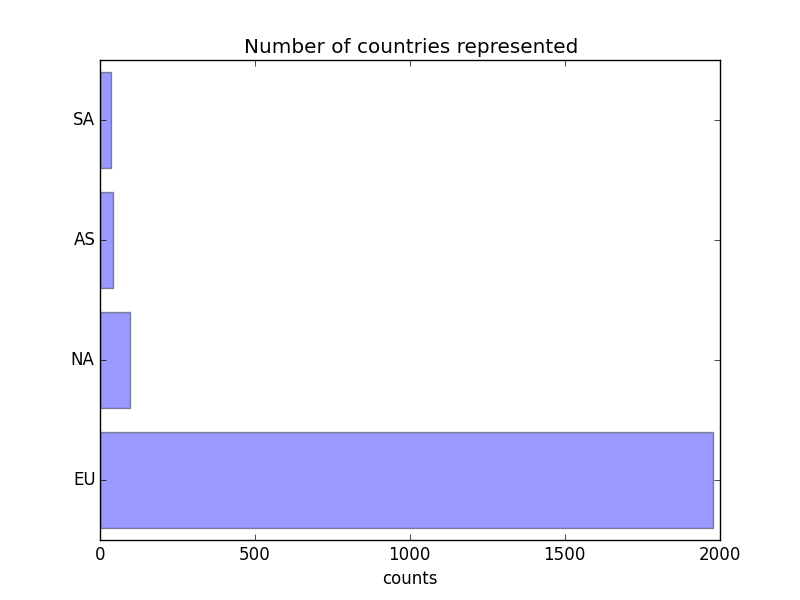
I believe that my GUI is very simple and the usability of the application is high. The user has only to follow 4 steps.

Some screenshot of the results are following:

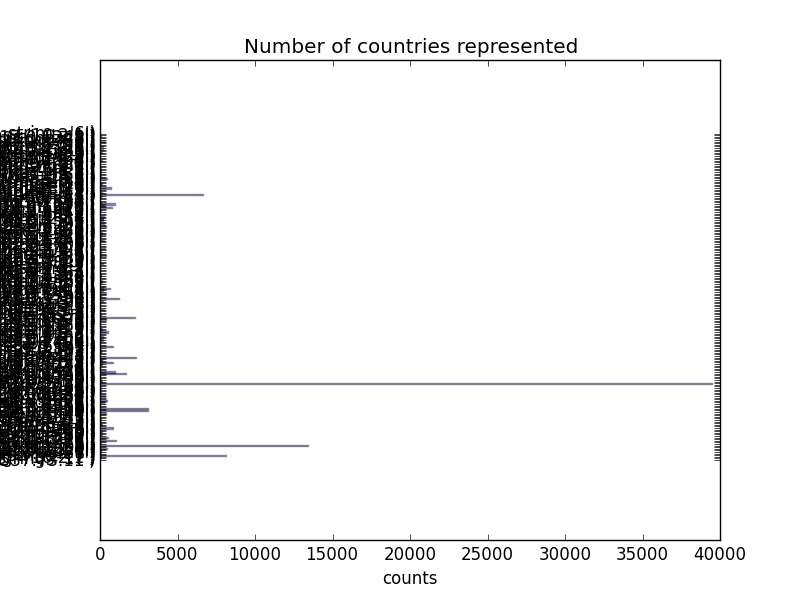
### Countries screenshot



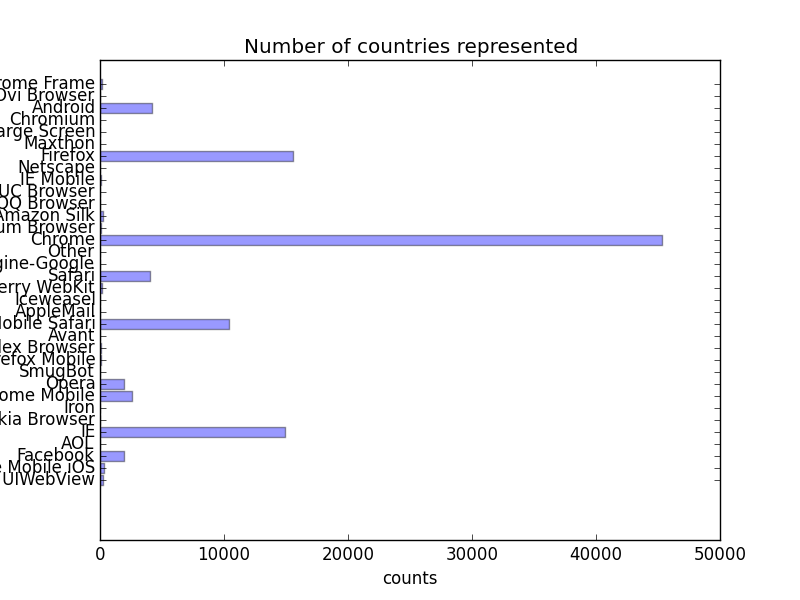
### continent screenshot



### browsers screenshot

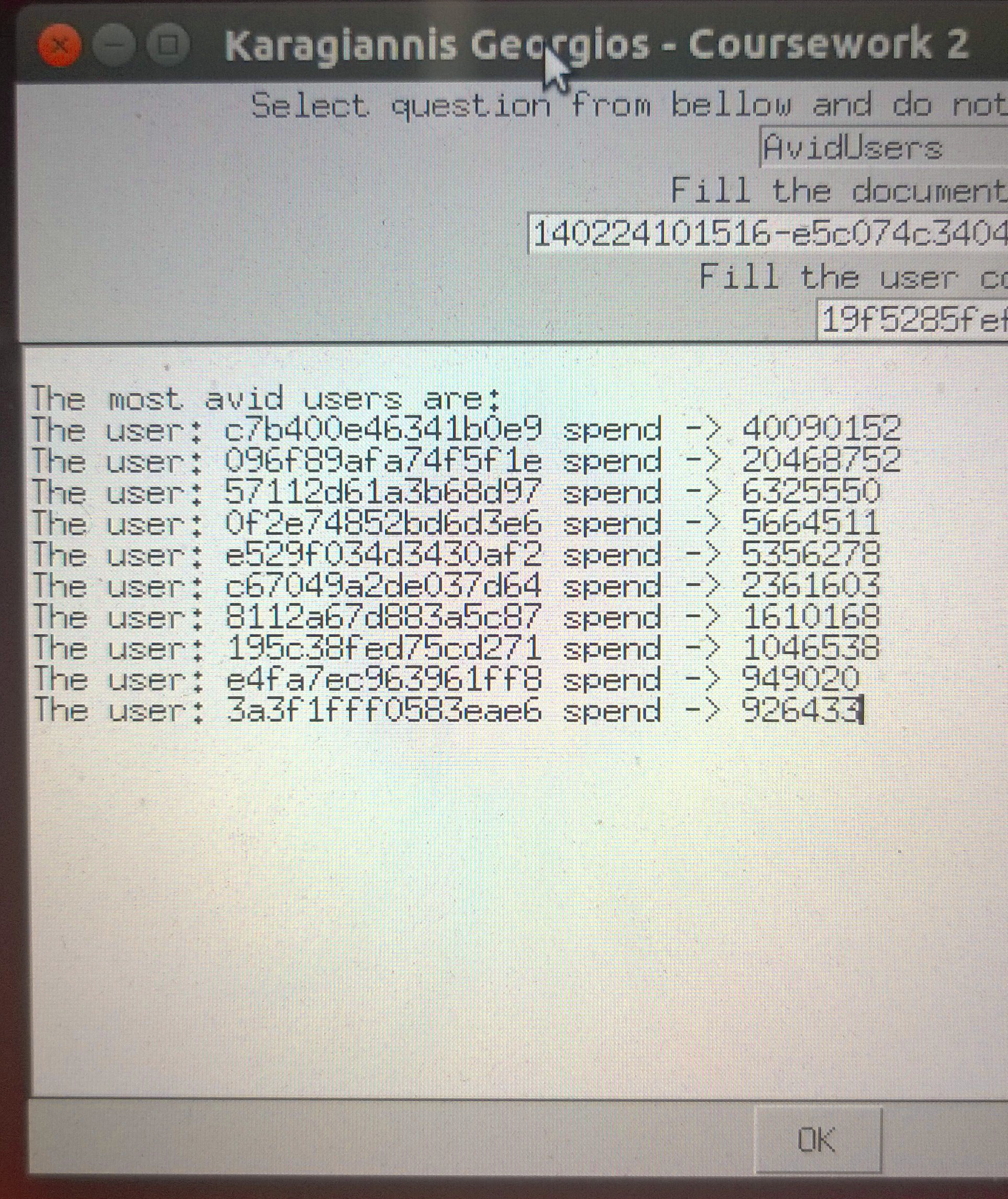


### browsers name screenshot



### avid users screenshot

(Problem with screenshot and the data. I took a real picture of my screen)



The users takes messages for the bad usage of the application.

# 5. Developer Guide

Firstly I will describe my application design and I will explain every class I have built, every method I implement and how my methods works.

## Classes

### class MyGui

That is the class which is responsible for the GUI, in this class I have the methods which implement the GUI using the tkinter library and the methods for the data input. For this class I took real help from the python’s sample you gave us.

#### def \_\_init\_\_

This method is just the constructor of the class, creates the windows.

#### def widget

This method has the fields of the basic window of my application, has the labels with the text, the buttons and the fields. I implement everything using the tkinter library.

#### def launch

This method takes the input of the user and passes the arguments in to the method launchProgramm in order to have functionality.

#### def selectfile

This method is in order to build the button for the file selection of the users. I have import a specific function of the tkinter, the filedialog in order to open a windows and to select a JSON file. This method just creates a global variable, which has the path for the file and the filename.

#### def printinthearea

This method exists in order to have access and to print results in the main text area of the GUI.

### class MyMethods

This class has the methods for the implementation of the four first questions. The methods for the question five are not here. They should be here but I had some problems at my code design and I choose to have the catch the requirements and to have a lack in the code design.

#### def \_\_init\_\_

It is just the constructor of the class and only has a dictionary because I use it in order to find the continent based on the country code.

#### def countries

This method is in order to calculate and to plot the histogram with the countries of the visitors for a specific document. Takes as arguments the filename and the document ID, firstly reads the JSON file line by line and if the document ID is that the user chose then append the country code in a list, after that using the library Collection I make a dictionary which has as key the name of the country and as value how many times that country exist. At the end I just illustrate that dictionary as histogram using the mathplot library. If there in nothing in the dictionary I print a message.

#### def continent

This method is in order to calculate and to illustrate a histogram with the continent of the reader for a document. Takes as arguments the file name and as the previous method (maybe I could use the previous method and not to write the same code again. If I had a separate method for the histograms I could use the previous method for that) calculate a dictionary with the countries and how many times users for each country read the document. After that using the dictionary cntry\_to\_cont (as the sample we show in the class) I create a new dictionary which has as key the continent code and as value how users for that continent. I just compare each country with the other dictionary and I increase the value each time. If there in nothing in the dictionary I print a message.

#### def browsers

This method is for the calculation and the histogram of the browsers, the full name. I read the JSON file line by line and I just append each browser name into a list, after that using the collections. Counter I create a dictionary in order to illustrate.

#### def browsersname

Same as the previous methods but this time I use a library in order to give me the main name of the browser based on JSON file’s string. I used the user\_agent in order to append into a dictionary the family of the browser. DOES NOT WORKS AT THE LAB! I could split the string but I didn’t.

#### def avidusers

This method is for the calculation of the most avid users. I read the JSON file line by line and I build a dictionary which has as key the user ID and each time I add the time of that user as value. At the end I sort that dictionary and I print the top ten users and the time of each one at the GUI and at the terminal.

## My Main Program

#### def launchprogramm

This method takes as argument the question the user chose to see, the file that the users wants to analyse and the user ID and document ID and each time calls the method which calculates the results.

The following methods are for the question five (5) only

#### def readfile

It is a method which just reads the JSON file. I call this method in from the other methods.

#### def readersOfdocument

It is the method for the 5-a. Takes a document UUID and returns all visitor UUIDs of readers of that document in a dictionary. The id of the users and how many times this user has read that document.

#### def documentsReaded

It’s the implementation of the 5-b. This method takes a visitor UUID and returns all document UUIDs that have been read by this visitor. Uses the def readfile in order to read the file and creates a list with the documents.

#### def alsoLike

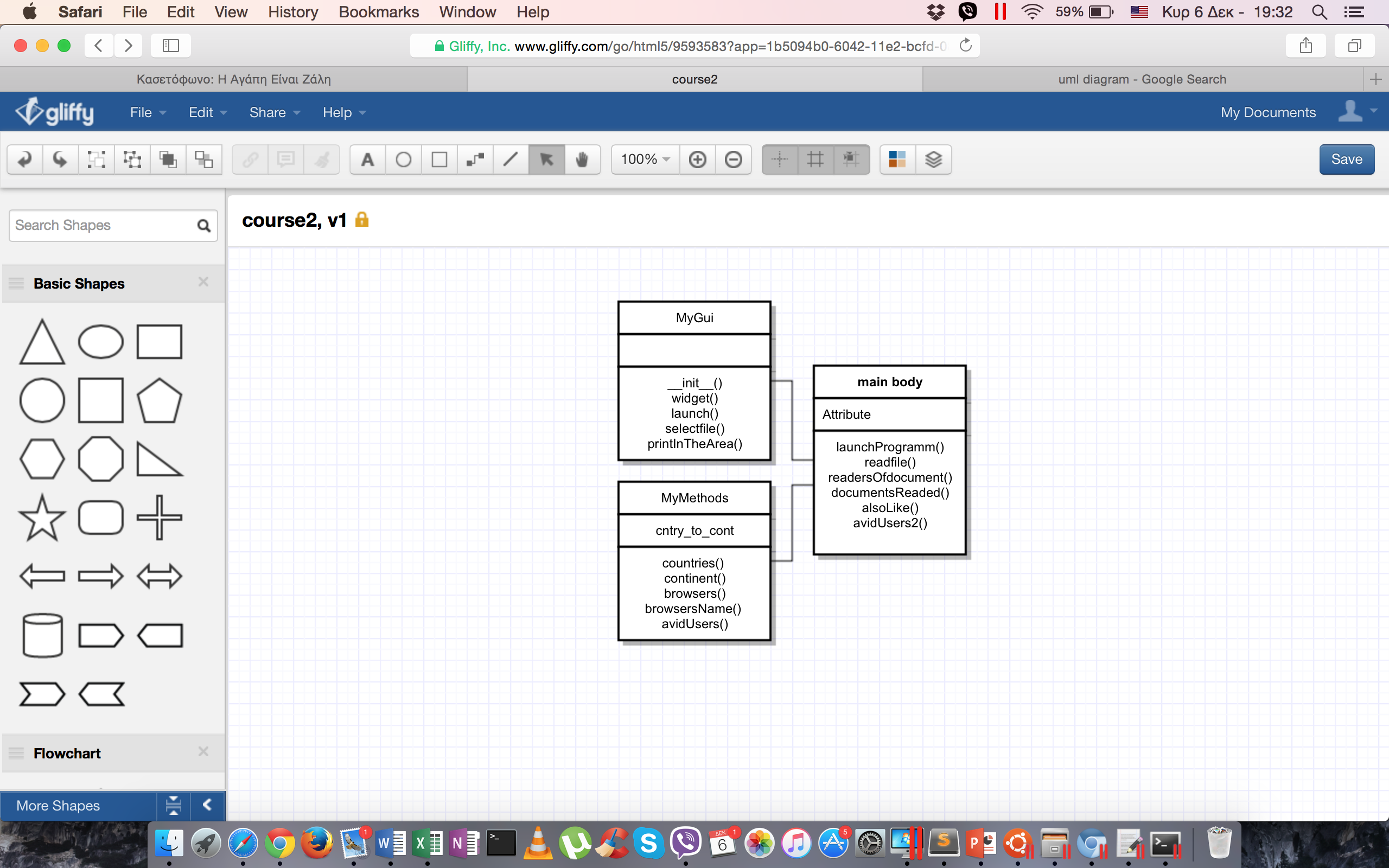
This method takes as arguments the user ID and the docID and I use the other methods in order to make the list of “liked” documents. I have 2 lists, the first on is for the visitors and the second is for the visited documents. The list with the visitors has the visitors of a specific document using the previews method. For each one I store the documents he read and I set that list, I removing all elements which matches documents ID. I have a new list the list temp which has these elements. For each document now I have to find the users and how many users where. I sort and I print that list. For the next question I use a sorting function in order to print the Top10 documents by page read time of readers.

#### def avidUsers2

This method is only for the shorting by time. Returns a dictionary with top10 users by the time. I used this method as sorting function for the 5 - e

Also at the main body I open the GUI and I call the launchProgramm with arguments the input of the GUI

## UML Class Diagram



# 6. Testing

The main problems are with bad inputs so when something like that happens I illustrate a message to user. I tried to check the data and to compare it with the result for a small file. I tried to catch all the errors.

I tried the application using a JSON file with 100K lines in order to see how fast it is.

I tried to press OK without a file as input.

I tried to press OK having an incorrect document ID in order to illustrate the histogram with the countries and the histogram with the continents.

I tried to run the application in a «slow» computer. I gave to my virtual machine only one core of my processor and 512mb RAM in order to see how fast my application is.

### Errors I found out

If the JSON file has not the structure it supposed to have then my application has problem. One other error I found the last time and I did not fixe it is about the question 5, if the userID and the documentID does not exists in the JSON file I should provide a message.

# 7. Reflection on programming language and implementation

Python is a multi-paradigm programming language: object-oriented programming and structured programming are fully supported, and there are a number of language features which support functional programming.

The core philosophy of the language is summarized by the document "PEP 20 (The Zen of Python)"

* Beautiful is better than ugly
* Explicit is better than implicit
* Simple is better than complex
* Complex is better than complicated
* Readability counts

Python has a large standard library and that is one of Python's greatest strengths, the official repository of third-party software for Python, contains more than 65,000 packages offering a wide range of functionality, including:

* graphical user interfaces, web frameworks, multimedia, databases, networking and communications
* test frameworks, automation and web scraping, documentation tools, system administration
* scientific computing, text processing, image processing

I had not at all experience with Python, I think that it is an easy to learn programming language. Python has a smaller number of syntactic exceptions and special cases than C#

My code design is not as good as I would like to be much better. I started implement my code having no experience at all in Python. My first code design was bad and I continue with that in order not to lose at least the requirements. I use the same code in many parts as the methods, I have double times a function which returns a top10, I could use another class for the question five and not to have my code at the main program. I should use a much better object oriented design.

The best I should do is first to understand in deep how the libraries I used works and to understand better the requirements and after to start the design of the code and the implementation. I am not happy at all with myself having that design, I believe that having a good design is the most important.

I start implement the method AvidUser with a much more different design which was making my program extremely slow. First of all I had to make a list with all the users but that list may have double time some users so I had to ***set*** the list. A have a variable in order to summarize the total page read time of each user in the list. In order to do that I have a double loop, I search each user in the list with the users in my data, I try find that user in my data, when I will find the user I will add his reading time and I will continue the search because maybe he is multiple times in the data. After that I add in a dictionary as key the username and as value the sum of the reading time. I set my variable as zero in order to continue with the second user. But not all the users have as event\_type the pagereadtime so I had to check about that. At the end I have to short that dictionary based in the value and to print the top10 user. The results was fine but the complexity of the algorithm big and the program really slow.

That was an example of my bad code design.

I also could use a separate method in order to plot the histograms instead of having the same code multiple times.

It was not necessary in some methods to use an extra list in order to build the dictionary.

The design of the question five (5) should be better but I chose to implement the task as it is and I had not time make it better. I was trying to make a separate class for the task 5, that class should have the methods and to call the class’s methods from my main program.

I also should catch more exception, I should fix the errors with the bad JSON file.

My application is a bit slow in the questions with the browser name.

## Libraries

* JSON (JavaScript Object Notation) is a lightweight data-interchange format. It is easy for use to read and write and also easy for machines to parse and generate. I have to use this library in order to read the data from the file.
* Tkinter is Python's standard Graphical User Interface package. It is the most commonly used one. There are several popular GUI library alternatives available, such as wxPython, PyQt (PySide), Pygame, Pyglet, and PyGTK. I had to implement a very simple GUI and I choose to stay in tkinter because that library we used in class.
* Matplot is a plotting library for the Python which produces publication quality figures. Matplotlib tries to make easy things and I could generate plots with just a few lines of code. One other choice it was Plotly python library which is a collaborative graphing and analytics platform. I choose to stay in matplot as we talk in the class
* Pandas is a library written for the Python for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables, time series etc. I chose NOT to use it because the size of application is not big enough and I had not enough time in order to learn how to use that library. I chose to think my own ways to implement the function for the analysis.
* For the 3-(b) question in my own laptop I used one more library, the ***user-agents*** (pypi.python.org/pypi/user-agents),especially I import: ***from user\_agents import parse***. I will explain why.
* For the 3-(b) I had to distinguish the browsers and to keep only the main of the browser. The first thing I thought it was to cut the string and to keep only the first word until the first (/) but it is not correct. I made a small search about the user agent string and I found out that the browser’s main name is not the first work (en.wikipedia.org/wiki/User\_agent) and the word “Mozilla” is used to indicate compatibility with the Mozilla rendering engine. I have as comment in the code both of the choices.
* I used the library collections in order to create directly dictionaries from lists. I used that library in the method which illustrates the browsers. I used the counter. A Counter is a container that keeps track of how many times equivalent values are added.
* I used the os.path in order to give me the path of the file that the user decided to analyse.
* I used the operator for the question 5 as the example we show in the class.
* I used the RE library (regular expression) in order to scan through a string, looking for any location where this RE matches.

# 8. Coursework and Course Conclusion

Python is a modern, scripting language and object oriented programming language that carefully incorporates features found in the most common industry and research languages.

It was the first application I have ever implement for in Linux platform using IDLE, Linux terminal and Python. It was a challenge for me and I had the opportunity to learn more about the technological advances scripting languages provide.

To summarize, in this project the most important thing for me was to gain skills using a scripting language like Python, to learn how to glue software together, I used some libraries and having no previews experience in Python but I think it was easy to learn comparing to C#.

As I working about the coursework I notice that Python has a very active online community. Python’s code is much shorter than C#’s code and also Python is good for web scripting

I finished my undergraduate studies at 2009 having a BSc in Information Science but the last 6 years I had lost the connection with programming. It was a bit difficult for me to start again programming but I am happy with the result. Choosing that course I tried to learn advanced programming techniques in a general purpose languages (C#) and in a scripting languages (Python). I implement software for different platforms, C# - Visual Studio for Windows and Python - IDLE for Linux. I tried as I could to gain skills from this course in order to use these skill in future working field.

Course’s notes and samples helped me a lot because some times in the class I couldn’t understand some parts of the course because of my poor English language. I tried to make the coursework as good as possible and also I tried to understand in deep the Python and how the scripting languages works. I tried to have a good class structure and data, I know that my application is just OK and I wish I had more time to spend on it. To be honest if I had the opportunity to implement again the same application I would use a complete different way. I delay to understand what I had to do and how to do.

# Appendix

The following website helped me to implement the coursework

* [www.macs.hw.ac.uk/~hwloidl/Courses/F21SC/index.html](http://www.macs.hw.ac.uk/~hwloidl/Courses/F21SC/index.html)  
  Course’s website helped me a lot because some times in the class I couldn’t understand some parts of the course because of my poor English language
* stackoverflow.com - codeproject.com  
  That websites was very helpful for me, I found some samples of code that helped me to understand some things and to implement the coursework

