Doğukan Ertunga Kurnaz
21702331
Section 1

Answers:

1. Perceval is easy to use tool for gathering data from various sources. It has many useful features such as built-in sleep-for-rate-limit support for some of the platforms such as Github. Because the Github API has a rate limit, we cannot fetch data from the 1 GET request to the API. I tried it on my own to see what happens. I saw that I could only fetch the bit of the data for the large repositories. So, the first requirement which is gathering data would be harder without this tool's absence. We know that from the industry, most APIs enforce us to pay subscription fees after some number of using it. With Perceval, we can define multiple API keys to bypass it. It is a very useful feature for the industry, especially for data scientists who are trying new things in their homes.

2. I love Elastic so I chose a repository from them which is ECS. So, all the requirements for the homework have been met. Repository data was fetched on 1 April 2022.

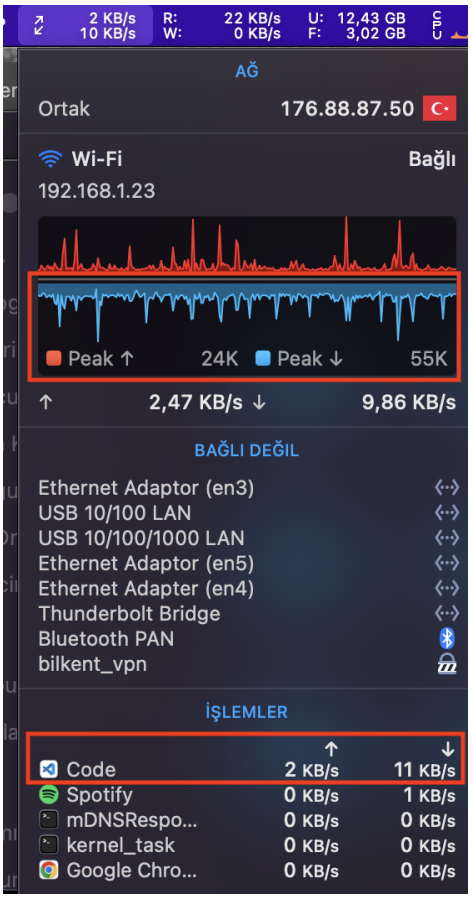
Code output for questions 2*, 3, 4**, 5, and 6:

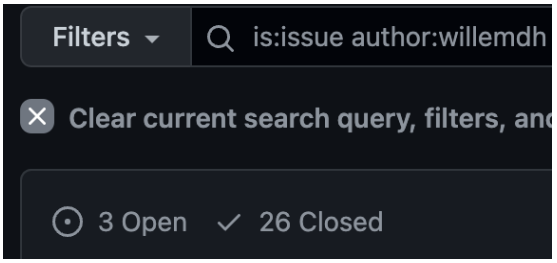
```
Today's date is:  2022-03-31 21:49:39.326971+00:00
----------------Part 2----------------
Date of first opened issue is:  2018-05-31 14:40:59+00:00
Date of last opened issue is:  2022-03-27 16:33:10+00:00
The repository's age is:  1396  days
Issues created in the last 30 day is:  6
It has 517 issues
It has 30 contributors
----------------Part 3----------------
Open Issue Count 130
Closed Issue Count 387
----------------Part 4----------------
Top 5 users who opened the most number of issues
webmat :  67
ebeahan :  45
vbohata :  40
willemdh :  29
MikePaquette :  22
----------------Part 5----------------
No of closed issues that did not have any assignee is:  283
Average day of resolution for issues without an assignee is:  334
----------------Part 6----------------
No of closed issues that had at least one assignee is:  104
Average day of resolution for issues with at least one assignee is:  256
```

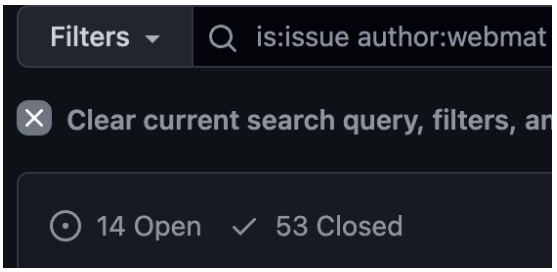
7. As you can see from the results if an issue has at least one assignee it is solved in a shorter time period. I think the reason behind this is that someone has an issue assigned to his/her, he/she does not want to take another issue that has no assignee on it. Additionally, a reminder e-mail has been sent automatically by GitHub if someone assigns you an issue. It improves the visibility of the issue assigned to you. There may be some sleeping issues with no assignee on it for the backlogging. But, the best practice is not to be.

Extras:

During the tests, I saw that even if I put multiple API keys, there is still a limiting factor for fetching data. Sleep-for-rate-limit option is necessary however it very slows down the process. ~500 issues fetched within the ~700 seconds which is a lot. We can see the spikes in the network activity to observe when the script is working or sleeping. Other than that, I also try fetching all data from the CLI tool of Perceval, but again it resulted in the same. Also, I exported the data it generated and saw that there is a huge number in terms of size when we consider it is just formatted text.







*Timezone is GMT+0 in the screenshot. Turkey is in the GMT+3 timezone.
**I verified the results by looking into GitHub issues.