HSLU: Lucerne University of Appplied Sciences and Arts

DVIZ Main Project

Kevin Häusler

BSC Artificial Intelligence and Machine Learning

Tutor: Prof. Elena Nazarenko, PhD

December 10, 2024

Table of Contents

1 Work Table

Due to me (Kevin Häusler) being the sole member of this project I have omitted the "done by" column.

Date	**Hours**	**Task description**
22.10.2024	1	Downloading 9k Dataset from Shodan
07.11.2024	1	Setup the project, convert the json data to xlsx, setup latex environment
07.11.2024	1	Configure the initial streamlit setup
08.11.2024	1	Start writing the report + setup Github
15.11.2024	2	Research about Shodan
23.11.2024	1	Writing more functions in the code
08.12.2024	1	Downloading new 32k Dataset from Shodan
08.12.2024	2	Analysis of the new 32k Dataset

2 About the Project

2.1 What is the project?

The project is something akin to a cyber security analytic about the devices listed on shodan.io based on their location being Rotkreuz. It is supposed to show insights about what is accessible and possibly insecure.

There are also dynamic parts where you can filter the data.

2.2 Data

The data is generated from shodan.io. I do have a lifetime membership that grants me enough credits to request and download 10'000 entries. I did filter the data to be only from Rotkreuz which resulted in a little bit over 9000 datapoints that I am using for this project.

Why I am not using the entire 32'000 dataset is explained in my report where I go indepth about how I got the data.

2.3 Tools

Here is a list of the tools I am using:

PyCharm
Python 3.11
UV
Streamlit
Shodan CLI
Docker
Github
LaTeX

I have setup the project in pycharm with UV so it is easier to distribute, there is also a docker file in case you have docker installed to easily run this project without having to create your own environment.

For the main library I have decided to use Streamlit, I have tried it out in the past but never for a real project so I wanted to get more indepth experience with it. It has a great documentation and it lets me easily create dataframes and charts for this project.

The whole report is also written in the PyCharm IDE with LaTeX and uploaded to the Github repository. It is written in the report.tex file and compiled to report.pdf.

This setup is also useful in case I need to work from different machines.

2.4 Shodan

Shodan.io or more commonly referred as just Shodan is a search engine tool used to provide a more comprehensive view of the internet, more specifically it crawls, scans and collects information from a wide range of devices that are public to the internet.

This includes servers, IOT devices, webcams, routers even industrial control systems.

This data can be browsed and used to "help" assess cybersecurity risks. It does not guarantee that your systems are secure or insecure, so do not mistake it as being a complete cybersecurity assessment.

I was lucky enough to get a "lifetime" membership for 5 USD, this gives me access to multiple tools and 100 query credits.

1 Query credit corresponds to 100 results which is why my initial dataset to try out the feasibility of this project only contains 9000 datapoints. For this project I decided to sign up for a freelancer membership so I get 10´000 query credits which I can use to get the entire dataset of Rotkreuz and maybe implement an interactive tool as well with the Shodan API.

3 Motivation

The main motivation of this project was to use shodan.io and streamlit to create something interesting. I have rarely used my shodan membership which felt bad even though I only paid 5 USD for a lifetime membership.

Streamlit has been something I have been trying out for a personal project but I never really got into it as much as I would have liked due to other commitments.

3.1 Cyber Security

Another relevant motivation would be cyber security. This semester I am taking the Introduction to Cyber Security class which has been surprisingly fun and interesting. While this project does not equal to a real cyber security assessment I do want to try my best to create something that is informative and useful in that regard.

4 Research

4.1 A Study on Internet of Things Devices Vulnerabilities using Shodan

There is a short research paper about the vulnerabilities of IOT devices where the author used shodan as a search engine to get the required data for his research.

4.2 Guides on how to use Shodan

There are many interesting and informative articles on medium.com and cyber security blogs about how to leverage Shodan for cyber security or more "nefarious" purposes like gaining access to unsecured webcams and NAS.

4.3 Using Shodan to hunt down Ransomware Groups

One of my favorite cybersecurity researcher and youtuber John Hammond is affiliated with the cybersecurity firm Huntress, which is how I came upon a blog post about using Shodan to hunt down Ransomware groups.

While it is not entirely relevant to my project it has been a very interesting read and I can recommend it for people that are interested in this.

https://www.huntress.com/blog/using-shodan-images-to-hunt-down-ransomware-groups

5 Data

5.1 Getting the Data

As mentioned I will be using Shodan to get the data about all the devices that are scanned in Rotkreuz. This is done with their search engine tool on https://www.shodan.io/search?query=country

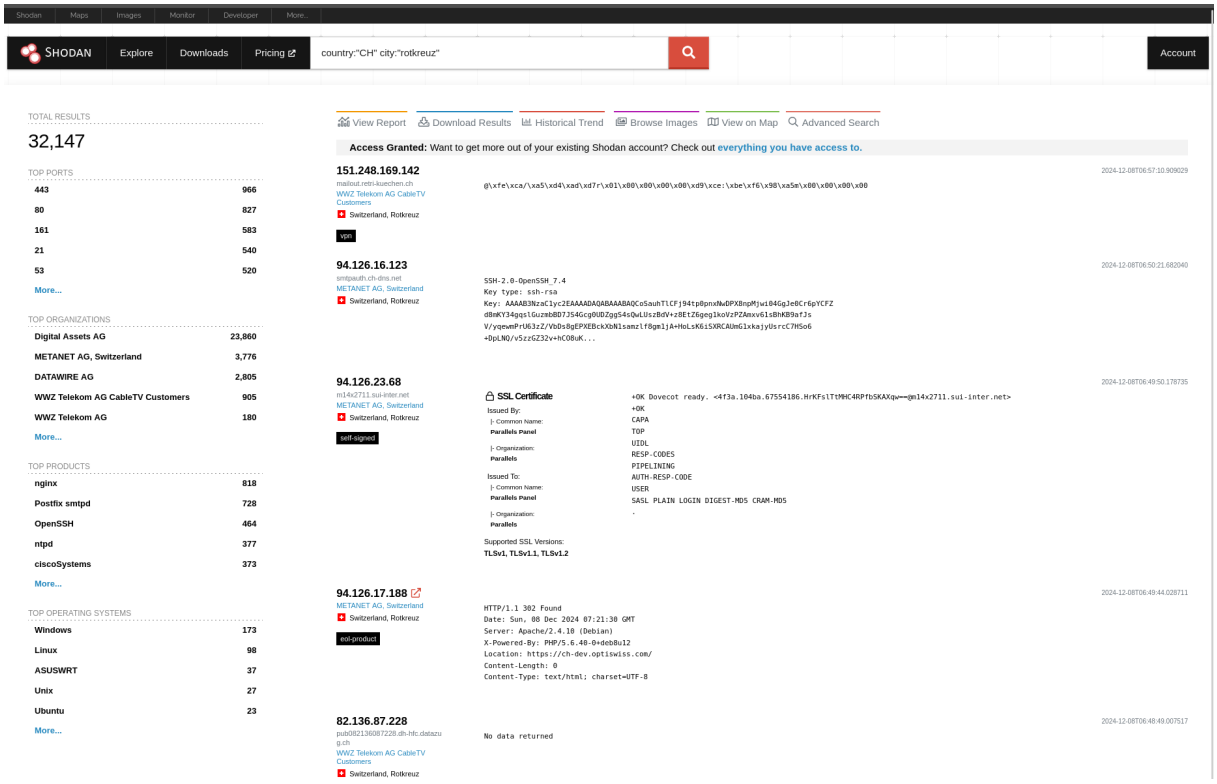

Figure 1: Shodan Search Results for Rotkreuz CH

Here we can see 32'147 results which I downloaded.

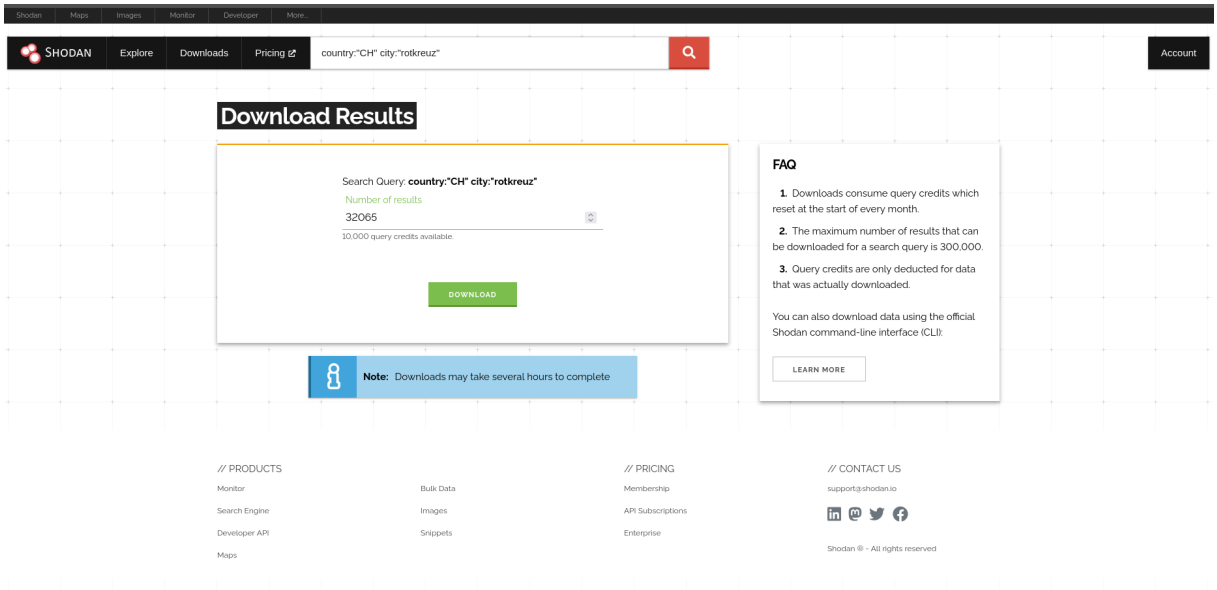

Figure 2: Downlaoding the search results

5.2 Preparing the Data

The download returns a *.json.gz archive which I have renamed to shodan_data_unprepped.json.gz for this project.

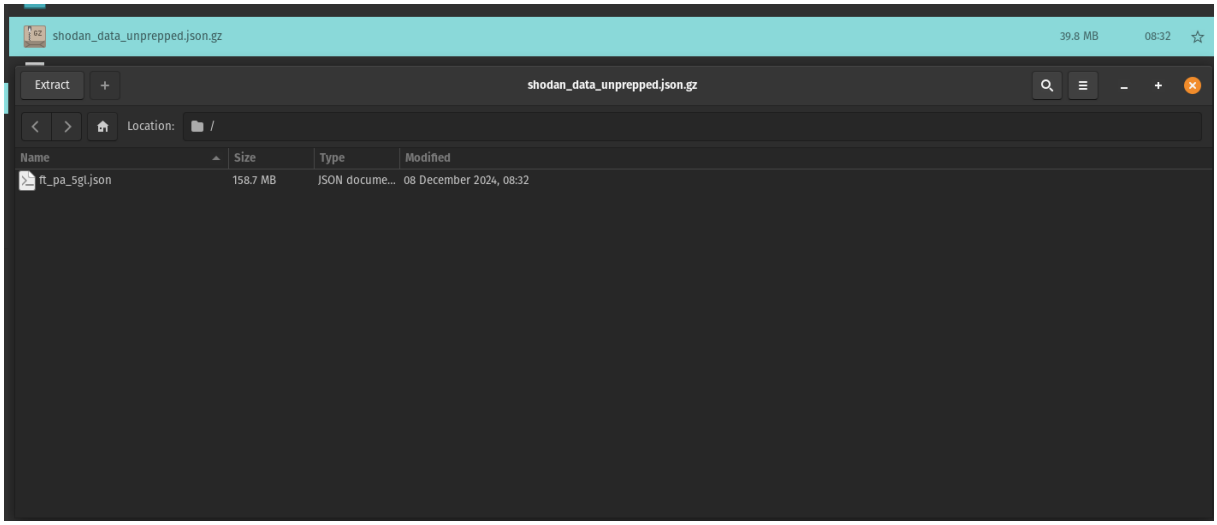

Figure 3: The Downloaded Archive
https://www.shodan.io/search?query=country

We can convert this with the following command: "convert shodan_data_unprepped.json.gz xlsx to an xlsx file which we rename to shodan_data_32k.xlsx.

5.2.1 Unexpected Issue

But what is this? We suddenly have a lot (27’768) of N/A entries in our dataset.

Figure 4: 27´000 N/A with the 32k Dataset

When I compare this with the 9k Dataset that I downloaded on 22.10.2024 we only have 4’643 N/A entries.

Figure 5: Only 4’643 N/A with the 9k Dataset

An indepth look at the Shodan search engine shows an organisation called Digital Assets AG with 23'860 datapoints, and while they have a lot of open ports they do not return any data.

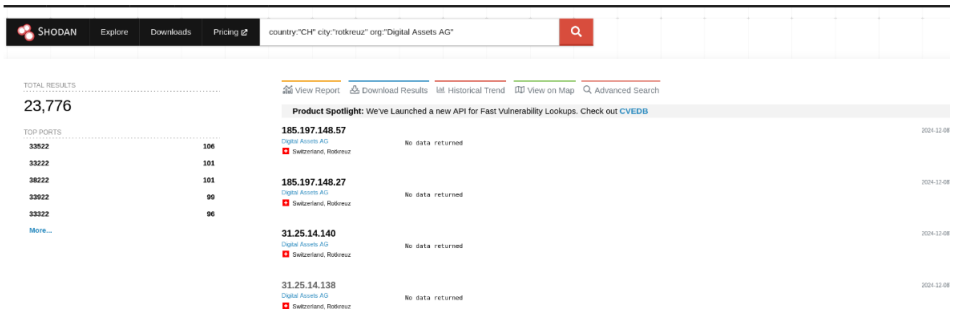

Figure 6: Digital Assets AG Rotkreuz CH Search

An overview of one of these IP can be looked at here https://www.shodan.io/host/185.197.148.27

When we look at the company itself we can see that they have 77'280 servers in Germany

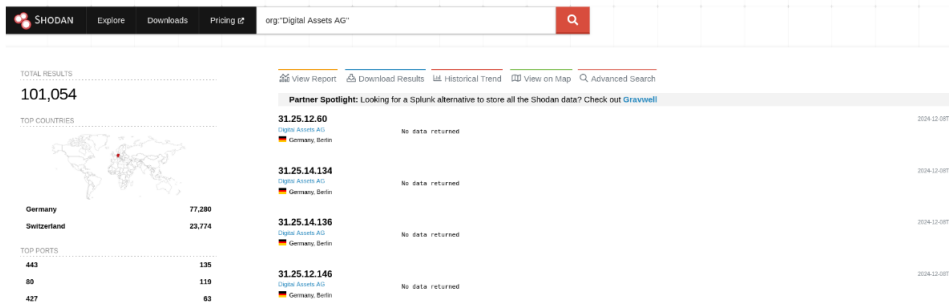

Figure 7: Shodan Digital Assets AG

The ASN lookup confirms that it is part of the Google Cloud Platform. While I do not have any confirmation it is highly likely that they are renting or colocating servers (in this case VPS) in the new CKW Datacenter in Rotkreuz.

To be safe I will use the 9k dataset from 22.10.2024 for this project.

5.3 Analyzing the Data

Now that we have our data we an start analyzing it. First I want to check some general information like: - What are the most common ports? - What are the most common used Software - Are there any weird/interesting findings that stand out?

5.3.1 Port Distribution

A closer look at the Port distribution shows that the main use is web (port 443 and 80) and email services (port 25). There are some concerning ports like DNS (53) that are not recommended to be accessible publicly.

This can be exploited by using the exposed DNS servers to DDoS other systems by sending tons of fake DNS responses in a DNS amplification attack.

Thankfully most of the exposed DNS ports are from web hosting companies and datacenters that should know the risks and how to prevent them. There are still 19 open DNS Ports from WWZ ISPs that could be residential that are worrisome.

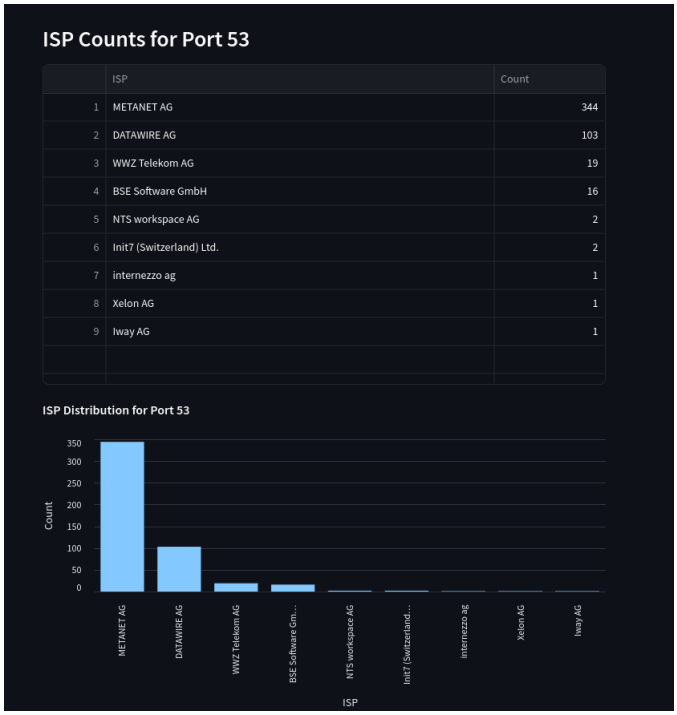

Figure 8: Port 53 Distribution by ISP