Group 11 Milestone 2 Report

Data Summary:

Name	Type	No. of Samples	No. of Features	Source(s)
COVID World	Time Series	156	7	https://corona-api.com/timeline
Ix APIs	Time Series Numerical, Categorical	20230	5	https://ix.br/agregado/ https://www.de-cix.net/en/locations/germany/frankfurt/statistics https://www.ams-ix.net/ams/documentation/total-stats https://portal.linx.net/api/lans/throughput?start=1592390700000&end=1592477100000 https://www.msk-ix.ru/en/traffic/ https://www.dataix.ru/en/statistics/ https://tools.franceix.net/pops https://www.netnod.se/ix-stats/sums/ https://www.hkix.net/hkix/stat/aggt/hkix-aggregate.html https://www.nix.cz/en/technical#traffic

Steam	Time Series	936 (Daily)	3	https://steamdb.info/app/753/graphs/
Socialblade	Time Series	12/ 52 Monthly/Weekly	100,000 (Pre.pro~ 2 feature) *	https://www.socialblade.com
P*rnhub	Time Series	109 (Daily)	2	https://www.pornhub.com/insights/coronavirus-update-may-26 **
Twitch	Time Series	94 (Monthly)	5	https://twitchtracker.com/statistics
Apple Mobility	Time Series		8	https://www.apple.com/covid19/mobility
Google Mobility	Time Series		8	https://www.google.com/covid19/mobility/
Playstation	Time Series		4	https://gamstat.com/playstation/
Google Trends	Time Series	* 262 (Weekly)	2	https://trends.google.com/trends/

Data Overview:

Covid World Data:

The general focus of the AMI project is to utilize the machine learning pipeline to learn about the impact of Covid-19 on various metrics. Our group focuses on the impact of Covid-19 on online traffic. This justifies the need for Covid-19 data.

In general, two approaches can be taken when analyzing the Covid-19 impact. In the first approach, we would be concerned with regional analysis. While this offers a larger dataset, and a more versatile approach, it suffers from differing data collection methods by country. For this reason, our project focuses on the worldwide impact, and requires a Covid-19 statistic that captures the daily world pandemic situation. The only draw back here is that first world countries

were hit with the Covid-19 pandemic first, while poorer nations are now seeing a flare-up in cases. As there exists a disparity in internet usage determined by the wealth disparity between countries, we may see an issue with Covid-19 data from the past few weeks.

IX data:

Internet exchange data is the main source of internet traffic information. This data gives us a complete overview of internet traffic as a whole, over the course of the Covid-19 pandemic. In terms of gross internet traffic, this data gives the most complete picture of internet traffic caused by any medium or motivation- business, education, entertainment, information, etc.

This dataset consists of a series of datasets, one for each Internet Exchange. For each Exchange, the dataset contains a time series of the data flux (bitrate).

The IX dataset contains the features Timestamp, Date, Bitrate, Internet Exchange, Type. Here, Type and Internet Exchange are categorical features, where Type indicates whether the dataset for one internet exchange uses the average or the maximum bitrate as a feature. Internet Exchange denotes which exchange the data came from.

Steam, Playstation:

Gaming is one of two major sources of entertainment-related internet traffic (alongside streaming video). In addition to gross internet traffic data, we decided that it might be interesting to observe specific use-cases as well. Even if the trend in internet traffic may be mild, this trend may still hide underlying changes in consumption.

The goal of our group is to find not just the total impact, but also the granular impact of the pandemic on internet traffic.

The Playstation features comprise of the date, the number of PS3, PS4, and Vita players.

Socialblade, Twitch:

These two represent the second major source of entertainment- streaming. Socialblade delivers Youtube statistics, especially directly concerning channels and channel views. Taking a general statistic that sums up the top Youtube channels in the world, by country, by week, we can find general viewing trends. The top 250 channels by country are taken, as these represent the most consistent source of views in non-pandemic times.

In the absence of data concerning total Youtube numbers, this approach yields the most reliable approximation of the Covid-19 impact on Youtube.

Additionally, we gathered data on Twitch viewing habits. However, this data is nonoptimal and will likely remain unused, due to the monthly and not minimum weekly resolution.

The Twitch dataset contains features on the average active streamers, average concurrent channels, average concurrent viewers, and the total time watched. Socialblade contains features on the daily, weekly, and monthly subscribers and video views, as well as the weekly change in subscribers and video views.

P*rnhub:

During initial research, we found that explicit content experienced the largest impact due to the pandemic. As lude content is known to be one of the major sources of internet traffic, even during non-pandemic times, an increase here should have an effect on overall internet traffic. Luckily, P*rnhub has been very forthcoming with their numbers during the Covid-19 crisis and offers a chart cataloging the daily viewing numbers.

Apple and Google Mobility:

The more people move, the larger the percentage of mobile data related internet traffic they cause. For this reason, we decided to find data concerning the mobility of people during the corona pandemic. Apple and Google, through the prevalence of Android and iOS, have the most complete dataset concerning mobility, which led us to choose these as sources.

The features contained in this data set are the date, census FIPS code, and the percent changes in retail and recreation, grocery and pharmacy, parks, transit stations, workplaces, and residential.

Google Trends:

Since Google is the biggest search engine, looking into it's search Data allows us to analyse Trends in keywords related to our research.
This dataset contains the percentage change in weekly searches for various keywords related to topics in our research like Covid-19, entertainment or Home-Office.

Data-Processing steps:

TO-DO:

All numerical, time-series data needs to be z-scored (Zero-mean, std-dev scaled). This, however, is an easy step and can be done prior to input in the training cycle. In addition, data points for weekly time series may need to be interpolated between weekly values, in order to match the granularity of the Covid-19 dataset.

In addition, datasets with multiple features will need to be optimized. Then Covid-19 dataset contains 7 features, of which three each are the daily/ cumulative deaths, recovered, and new cases. Naturally, some of these are highly correlated. For examples, cumulative deaths correlates heavily with cumulative cases, as the number of deaths is a function of the Covid-19 death rate and the virus's infection rate.

Location for Preprocessing:
Preprocessing code can be found on the gitlab, under src/data_collection. Preprocessed data can be found under /res/dataset/processed.

Covid-19:
Covid-19 data required no cleaning or completion, as the project is purely world-centric. Originally, we were using an API that allowed us to collect the major features for every country. However, the API recently broke, and no longer offers data beyond a certain, small range. This data would have needed cleaning, as many countries update their infection counts irregularly. We would have used the by-country metrics in order to limit the scope of the project to the developed world, where data is more readily accessible and accurate. The final world Covid-19 data contains the features of total and daily recoveries, deaths, and cases, as well as the daily dates.
As mentioned in the TO-DO, the Covid-19 data will need to be optimized, as several features are heavily correlated.

Socialblade:
The Socialblade data was scraped from Socialblade, a website that tracks daily, monthly, and weekly activity of the most influential channels on several big social media platforms such as YouTube, Twitter, Instagram, Facebook, etc. We analyzed the available data for all categories and found that only YouTube and Twitter data was sufficiently well tracked to be of any use. Due to time constraints, we focused on YouTube first. In case we find the time during the preprocessing phase, we might add the twitter data later.

Since there are no official stats for YouTube's traffic, we want to use the weekly/monthly view counts as a traffic indicator. YouTube or Socialblade does not offer such a view count. Instead, Socialblade offers the statistics for the top 250 subscribed channels per country. Combined with more than 240 country categories available, we can track the data for > 50.000 very influential channels.

There are several problems with this. First of all, the top channels right now do not have to include channels that were successful in 2018 for example. Also, if traffic was to mainly increase on mostly unknown channels we would have no chance of tracking it. Also, the data obtained

from the website is heavily biased. In 2018, there are about 30.000 tracked channels in the main categories, whereas in 2020 we track more than 50.000. It remains to be seen if we can balance this during the remaining preprocessing.

There were multiple problems with obtaining the data. Since Socialblade is protected against DDoS attacks and bots by Cloudflare, we wrote a custom scraper to get the data. The scraping happens in three steps:

1. Get all countries available on Socialblade.
2. For each country, get the list of the top 250 subscribed channels.
3. For each of these > 50.000 channels, scrape all available data from the stats page.

The two main issues were:

- Bypass the Cloudflare bot challenge request.
- Bypass the Cloudflare IP block after issuing too many requests.

We solved the Cloudflare bot challenge with the cloud scraper module, a module that can respond to the Javascript request from Cloudflare to imitate a real browser. We then wrote a scraper that wraps around the cloud scraper to process the data from the websites HTML code. This was a little more complex than we initially anticipated. The data is saved inside Javascripts and not accessible via HTML tags. The scraper takes care of these problems.

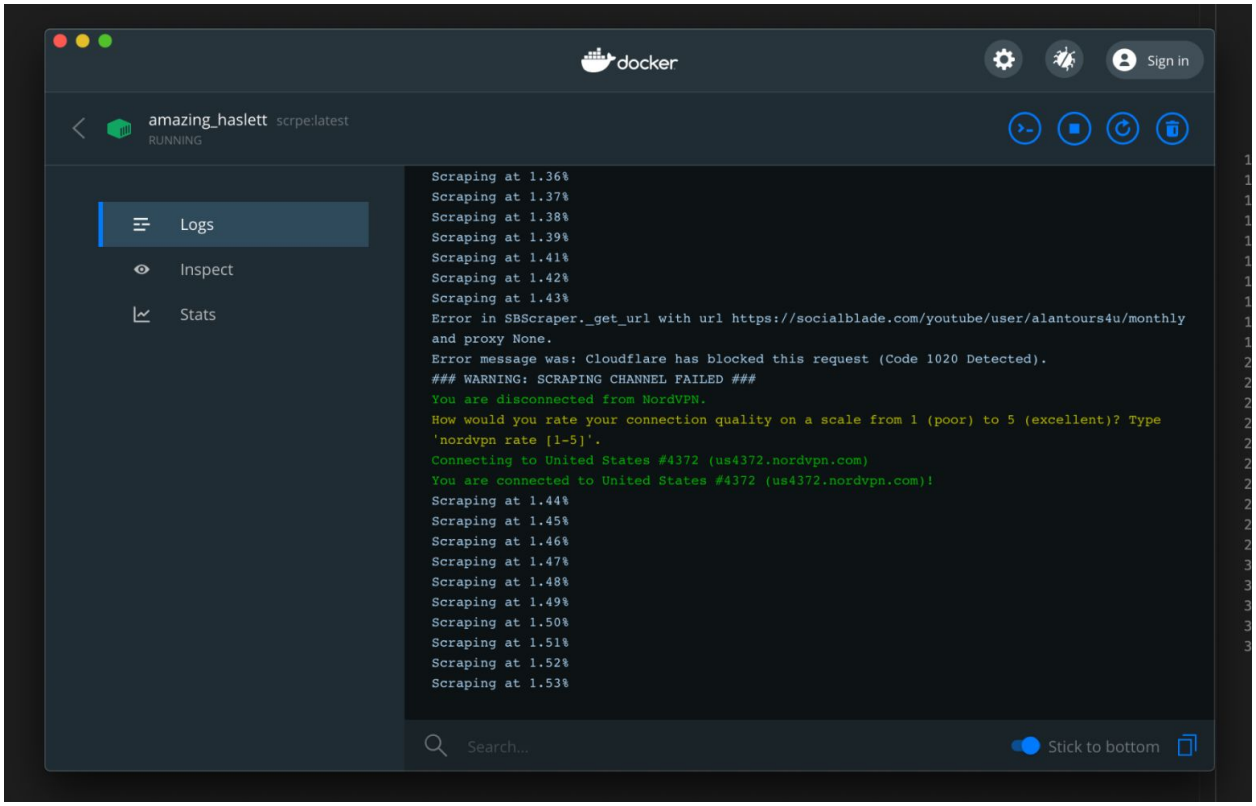
IP blocking was an even bigger issue. In order to prevent this, we connect to a VPN account and change our server as soon as Cloudflare starts blocking our requests. Even then it takes a huge amount of time to scrape all channels. We addressed this by dividing the list of links into smaller work packages, dockerized the scraping process and are able to scrape with up to 6 VPN connections (maximum number of simultaneous connections for our VPN account).

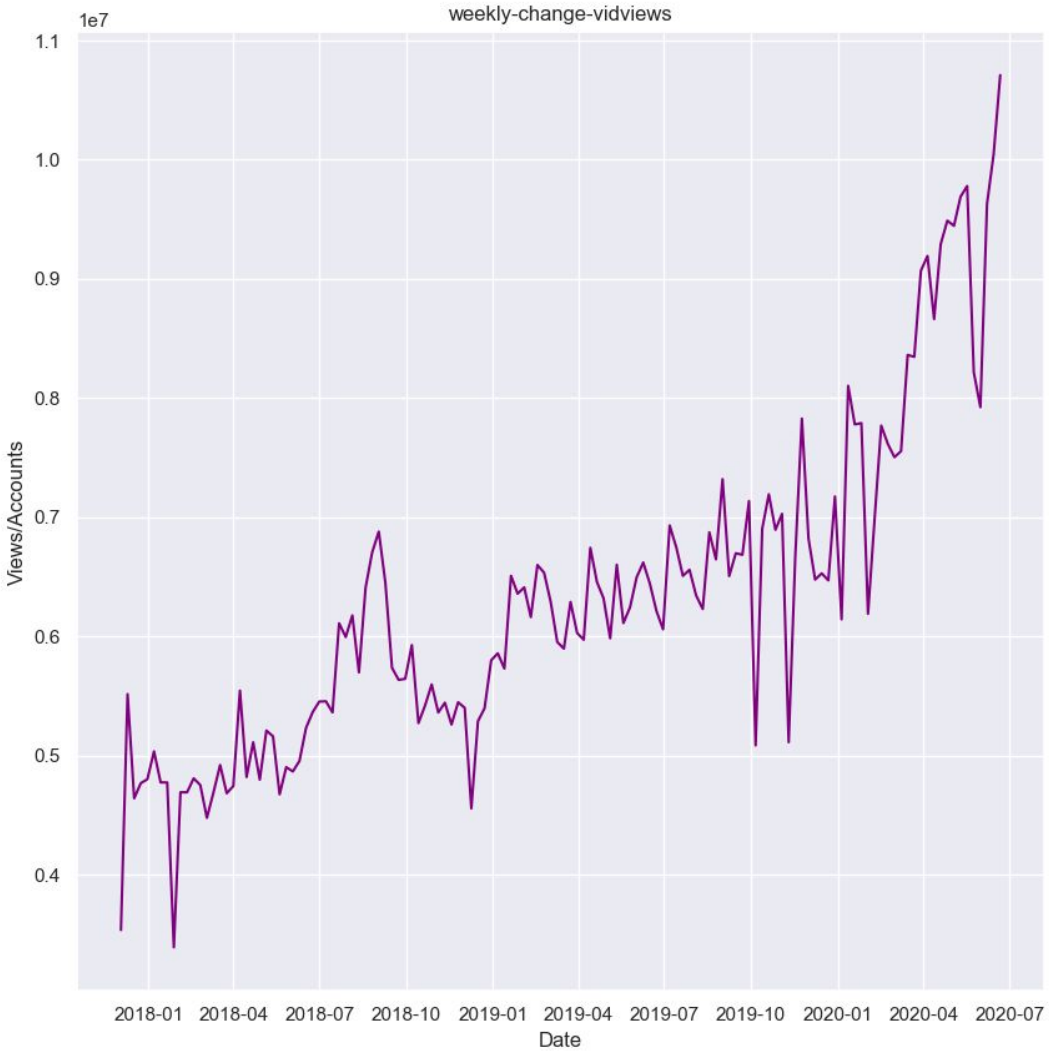
The final workflow was as follows:

- Scrape the available countries, scrape their top lists and save the links into work packages by running work_packages_creator.py.
- Start containers for the work packages with start_containers.py

The complete code can be found in social_scraper_lite, all Docker related files are located in the Docker subfolder. Please note that the duplicate files in Docker are intentional to be able to load them into the containers.

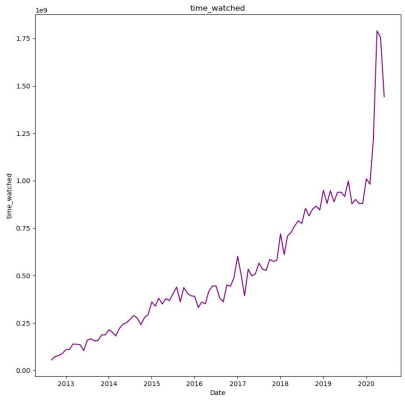
In a nutshell, the crawler sends requests until Cloudflare blocks the IP address used. It then switches over to another VPN server, leaving Cloudflare clueless of prior activity.

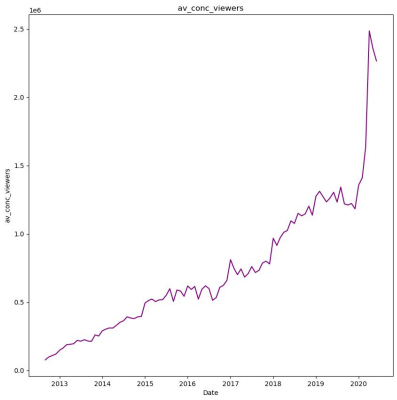

docker
Sign in
amazing_haslett scrpe:latest
RUNNING
Logs
Inspect
Stats
Scraping at 1.36%
Scraping at 1.37%
Scraping at 1.38%
Scraping at 1.39%
Scraping at 1.41%
Scraping at 1.42%
Scraping at 1.43%
Error in SBScraper._get_url with url https://socialblade.com/youtube/user/alantours4u/monthly and proxy None.
Error message was: Cloudflare has blocked this request (Code 1020 Detected).
WARNING: SCRAPING CHANNEL FAILED
You are disconnected from NordVPN.
How would you rate your connection quality on a scale from 1 (poor) to 5 (excellent)? Type 'nordvpn rate [1-5]'.
Connecting to United States #4372 (us4372.nordvpn.com)
You are connected to United States #4372 (us4372.nordvpn.com)!
Scraping at 1.44%
Scraping at 1.45%
Scraping at 1.46%
Scraping at 1.47%
Scraping at 1.48%
Scraping at 1.49%
Scraping at 1.50%
Scraping at 1.51%
Scraping at 1.52%
Scraping at 1.53%
Search...
Stick to bottom



Twitch:

Twitch data required no additional data cleaning or completion. All data was as it should be. Interestingly, a clear picture can be made detailing the impact of the pandemic: viewing numbers spike heavily at the start of the pandemic

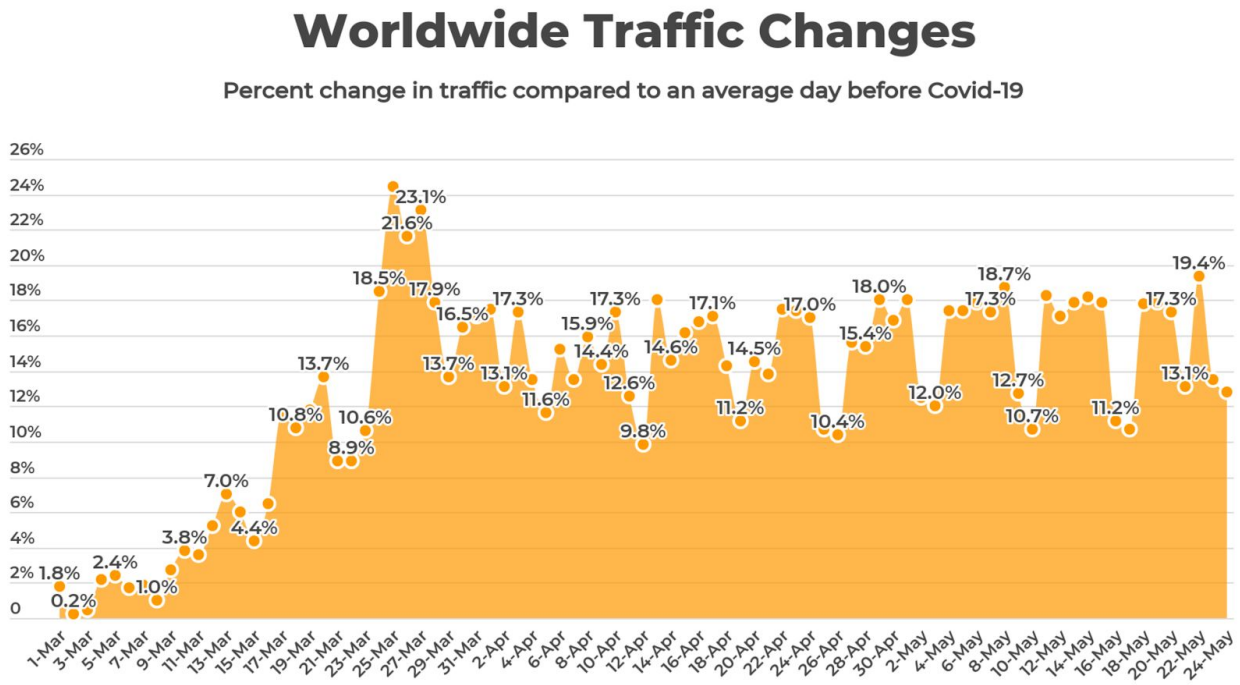




.

P*rnhub

Data from this service also required no fixing or completion. The data was readily available in .csv format, and displays a clear trend that can be attributed to the pandemic.



IX datasets:

Several IX datasets were only available as graphs in image format. In order to extract relevant information, we utilized the Graphreader library, which allowed us to gain information to the resolution of the image. While this creates a slight error, the IX dataset contains enough information to infer rough trends.

As the scope of our project is concerned with modeling the world-wide Covid-19 impact, the IX dataset will be merged along all internet exchanges, with the hope of creating an approximation based on the most used Internet Exchanges.

Apple and Google Mobility:
This dataset was added relatively recently. As of yet, we have not further inspected the core of the data. Once we have done so, the dataset will likely be compressed into the minimum number of features needed to describe the rough mobility of people during the pandemic.

Google Trends:
Data in this dataset required no processing or completion, it was already available in .csv format after extraction. We extracted this data using the pytrends library in Python and saved it in .csv format.