Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220551

Atlantes: Automated Health Related & COVID-19 Data Management for Use in Predictive Models

George VANGELATOS[a,1], Haralampos KARANIKAS[a] and Sotiris TASOULIS [a]

[a] *Department of Computer Science and Biomedical Informatics, University of Thessaly, Lamia, Greece*

Abstract. The scientific community, having turned its interest, almost entirely, to the treatment and understanding of COVID-19, is constantly striving to collect and use data from the countless available sources. That data, however, is scattered, not designed to be combined, collected in different time periods and their volume is constantly increasing. In this paper, we present an automated methodology that collects, refines, groups and combines data for a large number of countries. Most of these data resources are directly related to COVID-19 but we also choose to include other types of variables for each country, which may be of particular interest for researchers working in understanding the COVID-19 pandemic. The presented methodology unifies critical information regarding the pandemic. It is implemented in Python, provided as a simple script that extracts data, in the form of a daily time series, in a short period of time, directly available to be incorporated for analysis.

Keywords. COVID-19, open-source, health data collection, data management

1. Introduction

In late 2019, an infection of unknown origin was detected with manifestations of respiratory diseases as it began to spread rapidly in Hubei Province, China, especially in its largest city, Wuhan. Shortly afterwards, the World Health Organization designated coronavirus 2019 (COVID-19) and by March, the outbreak was a "pandemic" [1].

Although there has been significant forecasting and research, there is little data available to actively monitor current data trends and their statistical and practical significance. In addition, in most cases this data is available from different sources, with different formats and even different units of measurement for different time periods. For this reason, in this paper, we introduce a unified methodology in the form of an open-source Python script that is able to aid in the collection of all information relevant, and / or possibly relevant, to COVID-19 and their integration into a coherent data set with universal names, units of measurement and pre-specified time frames. This way, we expect to enforce data analytics and machine learning research towards deploying appropriate prediction and forecasting tools for COVID-19 outbreaks [9]

[1] Corresponding Author, George Vangelatos, Department of Computer Science and Biomedical Informatics, University of Thessaly, Papasiopoulou 2-4, Postal Code: 35131, Lamia, Greece; E-mail: gvangelatos@uth.gr

Our methodology supports the PrescITs' project Knowledge base (KB). The PrescIT project[2] aims to develop a Clinical Decision Support System (CDSS) platform to support safe e-Prescription, using per-disease rule-based algorithms to prescribe treatment [2] and knowledge extraction from reliable sources, such as large drug databases, scientific literature and the pharmaceutical market. One future research direction that the KB could support, is the investigation of relations among the COVID-19 pandemic and the chronic disease prescription.

2. Methods

Our methodology includes various steps that are consolidated via an open-source[3] implementation in the form of a single Python function. Below is an overview of the workflow followed by the algorithm during its execution

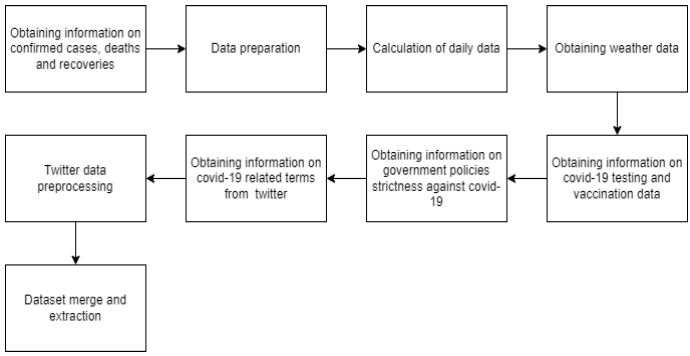

Figure 1. Algorithm execution flow.

When called, the function first retrieves data from a repository [3] about cumulative COVID-19 cases, cumulative recovered and cumulative deaths and transforms the dataset into a time series. Continuing the time series is enriched by calculating the daily changes of the cumulative variables by subtracting the next day from the one before. At this point the nearest weather station to the center of the chosen country is accessed and hourly weather data [4] is retrieved and transformed into daily average values.

The next step in the process is to retrieve information about the selected country's COVID-19 testing, vaccinations and the strictness of the policies adopted [5][6][7]. The stringiness of the policies adopted is rated on a scale from 1 through 5, and includes quantification of policies such as restrictions on gatherings, work from home, facial coverings etc. At this point new variables are calculated combining the ones retrieved as well as preprocessing is used to treat missing values and clean up data. That is achieved either by linear interpolation or corrections of mistakes in the data values that are gathered from the data sources such as duplicates entries for certain dates.

Finally, a plethora of repositories' subdirectories are accessed to retrieve information about daily values of the usage of COVID-19 related terms in twitter [8] and are added to the time series, after which all the collected, created and cleaned up datasets are merged in one which is returned through the function.

[2] The PrescIT project: https://www.prescit.com/
[3] GitHub: https://github.com/gvangelatos/atlantes

The function that implements the aforementioned methodology is entitled "creatCovidDataframe ()". It is written using the Python programming language utilizing several of its available packages. The function when called automatically collects, prepares, creates and consolidates data up to one day before the date it was called.

It is important to note that particular weight has been given to the data processing step, in order to automatically extract data that are immediately available for any further data analysis. During execution the algorithm takes care of any discontinuity in the data, as well as the uniformity of the time intervals between the entries of the data set. Time intervals uniformity is achieved by carefully selecting repositories with daily available data or calculating daily data from repositories that offer a higher sample rate, such as hourly.

3. Results

The function "creatCovidDataframe()" accepts only one argument, which is the country for which data will be exported. At the time of writing the function supports a total of 121 countries which are available on GitHub. During execution the function prints on which part of the algorithm it is, as well as the percentage on which it is currently on as shown in Figure 2.

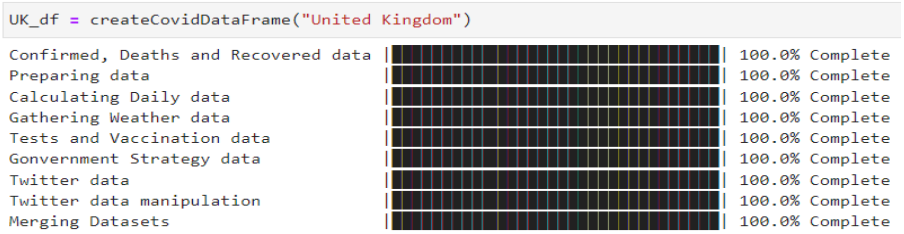

Figure 2. Algorithm runtime output.

The data set exported by the function, consists of daily data, with the first date starting on 01/01/2020 and the last being the day before the one on which it was called. In the last dates in some of the variables that make up the time series, due to a linear interpolation that is applied to the dataset and because their sources do not have new data yet as they are updated every other day, the value of these variables remains constant for 2 and / or 3 days. That is the result of the linear interpolation function applied that carries forwards the data values to fill all the available dates.

In total the function generates a data set which is a time series with sampling time per day and 98 parameters / variables for each day. These 98 variables combine to form the following categories:

- Daily weather data
- Government measures and data on austerity measures
- Global usage terms for COVID-19 on Twitter
- Country statistics
- Daily and cumulative data COVID-19
- COVID-19 Testing and Vaccination data
- Daily admissions to hospitals and intensive care units (some countries)

Comparing our presented methodology with other tools available right now as Open-source solutions, the inclusivity and availability of data are evident. While other solutions

offer only region specific data or provide tracking for only one variable throughout numerous countries, our presented tool is able to access COVID-19 related data for any one of the 121 countries in minutes as shown in Figure 3.

```
--- Maximum observed execution time: 376.6089494228363  seconds ---
--- Minimum observed execution time: 118.51333141326904 seconds ---
--- Average observed execution time: 211.48482256218537 seconds ---
```

Figure 3. Algorithm calculated execution times.

In addition, most solutions for accessing data require extra programming steps in order to access their remote repositories (CSSEGISandData) or they work by manually downloading the datasets (ourworldindata) and integrating them into the code. While some limit the data categories strictly to COVID-19 related some of the datasets are not even in a daily series format. The function presented is designed to fully automate this process accessing, merging and preprocessing data to integrate it into a coherent dataset with predefined units of measurement and time frames.

Finally, in Figure 4 we present two from the many interesting visualizations resulting from variables contained within the exported dataset from "creatCovidDataframe()" showcasing the huge scientific interest that exploration of variables whose impact hasn't yet been accurately measured such as social media and weather data pose.

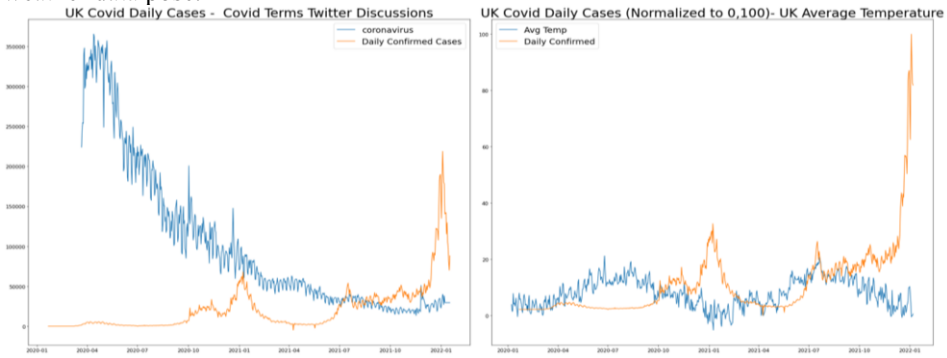

Figure 4. Daily social media and weather data plots.

4. Discussion

As countries experiment with ways to limit the spread of COVID-19, and as analysts continue working on unlocking the secrets of mitigating the spread of the pandemic, they will need more and more data. Thus a way to acquire clean and trustworthy data in a fast and reliable way is of outmost importance in the race against the spread.

Furthermore, understanding the inner workings of the progression of the pandemic proves troublesome while some areas seem to be doing a better job than others adopting the same measures; the reasons behind this still remain undiscovered. As a result a widening of the search for variables explaining that phenomenon is required, something that call for variables like temperatures and twitter metrics being available for the same time frames as the COVID-19 data and being accurately and frequently collected.

5. Conclusion

The data generated by our methodology is largely ready to be used in data analytics and predictive models, helping to mitigate the problems faced by COVID-19 and other data analysts. The implemented function tries to solve the problems of the data being scattered, not designed to be combined, in different time periods and that its volume is constantly increasing by offering a fast, automated and trustworthy solution to acquiring data and removes most of the need for pre-processing.

Finally, within the exported data, are included variables, still unknown whether they do offer any insight to the pandemic, or ones whose impact hasn't yet been accurately measured due to the lack of data. As such, government policies adopted, temperature values, social media metrics and country related statistics are included in the dataset as a possibility to hold the key to further our understanding of the pandemic.

Moving forward, we need to overcome limitations in order to be able to improve Atlantes. Such limitations are finding sources that deliver data daily or switch to sources with an even greater level of credibility. In addition, looking into the future we aspire to incorporate even more variables in the collection of Atlantes such as cases by gender as well as use KB to support is the investigation of relations among the COVID-19 pandemic and the chronic disease prescription by incorporating related data.

Acknowledgement

This research has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (project code: T2EDK-00640).

References

[1] WHO Director-General's remarks at the media briefing on COVID19- 03/2020 [Google Scholar]
[2] Karanikas H, et al., Development Of Prescription E-Protocols For Medicines And Integration On The Greek National E-Prescription System, Value Heal. 18 (2015) A385. doi:10.1016/j.jval.2015.09.835.
[3] Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Inf Dis. 20(5):533-534. doi: 10.1016/S1473-3099(20)30120-1
[4] Meteostat: The Weather's Record Keeper (Source: Meteostat)
[5] Mathieu E, Ritchie, H, Ortiz-Ospina E, et al. A global database of COVID-19 vaccinations. Nat Hum Behav (2021). https://doi.org/10.1038/s41562-021-01122-8 .
[6] Hasell J, Mathieu E, Beltekian D, et al. A cross-country database of COVID-19 testing. Sci Data 7, 345 (2020). https://doi.org/10.1038/s41597-020-00688-8
[7] Hale T, Angrist N, Goldszmidt R, et al. A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). Nature Human Behaviour. 2021). https://doi.org/10.1038/s41562-021-01079-8
[8] Banda JM, Tekumalla R, Wang G, et al. A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration [Data set]. (2022). In Epidemiologia (Version 104, Vol. 2, Number 3, pp. 315–324). Zenodo. https://doi.org/10.5281/zenodo.6336166
[9] Painuli D, Mishra D, Bhardwaj S, et al. Forecast and prediction of COVID-19 using machine learning. (2021). Data Science for COVID-19, 381–397. https://doi.org/10.1016/B978-0-12-824536-1.00027-7