Corona Datenverarbeitung

Prüfungsstudienarbeit des 4. Semesters  
Prüfer: Markus Eider M.Sc.

**Studiengang:**  Angewandte Informatik

**Semester:** SS 2020

**Datum der Abgabe:** 06.05.2020

|  |  |  |
| --- | --- | --- |
| Name | Matrikelnummer | E-Mail |
| Philipp Muhr  Michael Mican  Maximilian Seitz | 00692629 00692390 00692807 | philipp.muhr@stud.th-deg.de michael.mican@stud.th-deg.de maximilian.seitz@stud.th-deg.de |

Erreichte Punktezahl: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Note: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Erstkorrektur Unterschrift: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Zweitkorrektur Unterschrift: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Inhalt

[1 Einleitung 3](#_Toc39577157)

[1.1 Motivation 3](#_Toc39577158)

[1.2 Problemstellung 3](#_Toc39577159)

[1.3 Lösungsansatz 4](#_Toc39577160)

[2 Methodologie 4](#_Toc39577161)

[2.1 Datenbeschaffung 4](#_Toc39577162)

[2.2 Datenverarbeitung 6](#_Toc39577163)

[2.3 Probleme 8](#_Toc39577164)

[3 Ergebnisse 9](#_Toc39577165)

[4 Schlussfolgerungen 10](#_Toc39577166)

[5 Anhang 10](#_Toc39577167)

# Einleitung

## Motivation

Das Coronavirus versetzt die gesamte Welt in einen Ausnahmezustand. Bereits im März 2020 wurde eine weltweite Pandemie ausgerufen. Bisher ist zudem unklar warum es in verschiedenen Ländern unterschiedlich bedrohlich ausfällt. Diese Situation führt zu Maßnahmen, wie Maskenpflicht in der Öffentlichkeit, Ein-/Ausreiseverbote oder sogar zu Ausgangssperren innerhalb von „Hot Spots“. In dieser Studienarbeit soll die Ausbreitung des Virus genauer unter die Lupe genommen werden. Außerdem werden die Gesundheitssysteme, der Gini-Koeffizient und die aktuellen Google Trends der Länder analysiert, um Klarheit gegenüber der Expansion zu erhalten. Durch Vergleich der verschiedenen Daten inkl. automatisierter Visualisierung sollen neue Kenntnisse bezüglich COVID-19 gewonnen werden. Die Problemstellung wird im nächsten Unterkapitel genauer veranschaulich.

## Problemstellung

Damit die Ausbreitung des Corona Virus visualisiert und analysiert werden kann, müssen geeignete Datensätze ausfindig gemacht und aufbereitet werden. Da es sich um die Ausbreitung eines Virus handelt können dementsprechend Daten vom Europäischen Zentrum für die Prävention und die Kontrolle von Krankheiten (ECDC) bezogen werden. Jedoch sollen Korrelationen zu verschiedenen anderen Faktoren festgestellt werden. Daher reichen die Daten des ECDCs allein nicht aus und es müssen ebenfalls andere Informationen gewonnen werden. Diese Arbeit wird nicht nur die Ausbreitung des Virus aufzeigen, sondern auch die Zusammenhänge zwischen Corona und den Zustand des Gesundheitssystems verschiedener Länder. Weiterhin wird die Einkommens- und Vermögensverteilung untersucht, um einen möglichen Zusammenhang zwischen COVID-19 und den Zustand der Gesellschaft bezüglich Armut und Reichtum festzustellen. Außerdem werden Google Anfragen visualisiert, um aufzuzeigen, wie der Verlauf der Interesse der Bevölkerung war. Im folgenden Kapitel wird unsere Vorgehensweise knapp beschrieben.

## Lösungsansatz

Da es sich beim Corona Virus um eine Pandemie handelt muss der Zeitverlauf möglichst Welweit betrachtet werden. Die Suche nach Geodaten ergab, dass die Corona Verbreitungsdaten vom ECDC bezogen werden. Die Google Anfragen Daten werden direkt von Googles API empfangen. Außerdem werden die Gini-Koeffizienten (Einkommens- und Vermögensverteilung) der verschiedenen Länder von der „World Bank Group“ als csv Tabelle angefragt. Die Datenverarbeitung, als auch die Visualisierung wird mit Python gehandhabt. Es wurden aus Übersichtsgründen mehrere Skripte erstellt, die den Download, die Verarbeitung und die Darstellung vollautomatisiert durchführen. Das Programm ist so konzipiert, dass es durch Ausführen einer Windows .bat Datei gestartet wird. Die Python Bibliothek matplotlib unterstützt das Zeichnen von Graphen, mit der es möglich ist Diagramme und auch Weltkarten zu erstellen. Durch diese werden unsere finalen Berechnungen schlussendlich visuell dargestellt. Im nächsten Kapitel werden die Techniken genauer erläutert und welche Probleme zur Zeit der Entwicklung auftraten.

# Methodologie

## Datenbeschaffung

Vom ECDC konnten Daten im csv Tabellen Format heruntergeladen werden. Dieser Datensatze beinhaltet den Anstieg der Corona Fälle in verschiedenen Ländern weltweit. Diese Daten bilden die Grundlage dieser Studienarbeit, da alles in Vergleich auf die Corona Fälle gesetzt wird. Ursprünglich sollten die Daten von der „World Health Organization“ (WHO) bezogen werden. Allerdings befanden sich die Daten der Corona Fälle auf der Website der WHO im pdf Format und waren daher deutlich schwieriger zu verarbeiten als die des ECDCs.

Die Datenbeschaffung der Google Suchen erfolgte direkt über die offizielle API. Da sehr viele Anfragen an die Schnittstelle von unserem Skript gestellt werden, wird ein „429 TooManyRequests“ Fehler abgeschickt und die Übertragung bricht ab. Ergo wurde in das Skript eine Abfrage eingebaut, die überprüft, wann dieser Fehler auftritt und der Download wird in diesem Fall um eine Minute pausiert. Danach wird überprüft bei welchem Datensatz abgebrochen wurde und an dieser Stelle werden dann die restlichen Anfragen an die Google API gestellt.

Zusätzlich wurden Daten zum Gini-Koeffizienten besorgt, um aufzuzeigen, wie sich COVID-19 auf Länder mit unterschiedlich starker Einkommensverteilung auswirkt. Diese Daten kamen von der World Bank wiederum in csv Form. Problem bei diesem Datensatz war, dass diese Tabelle im UTF-8 BOM Zeichenformat erstellt wurde. Folglich muss diese zum normalen UTF-8 Format konvertiert werden, um die Daten überhaupt weiterprozessieren zu können.

Als letztes wurden von der WHO Werte bezogen, welche die weltweiten Ausgaben der Gesundheitsysteme in verschiedenen Ländern behandeln. Diese sollen später im Vergleich mit den Corona Fällen darlegen, ob es Ländern mit guten Gesundheitssystemen in dieser Krise besser ergeht, als welchen mit schlechteren medizinischen Versorgungen.

Nachdem die richtigen Datensätze nun verfügbar sind, wird nun auf die Aufbereitung der Daten eingegangen.

## Datenverarbeitung

Die erhaltenen Daten werden automatisiert von unseren verschiedenen Skripten verarbeitet. Die Software kann über eine batch Datei gestartet werden, welches sich außerdem um die Installation benötigter Python Bibliotheken kümmert.

Code 1: „run.bat“

@echo off

cd scripts

echo Installing requirements...

pip install -r requirements.txt >> nul

if %errorlevel%==1 (

    echo Failed to install requirements!

    pause

    exit /B 1

)

echo Requirements installed successfully!

echo Running script...

python main.py

pause

Danach wird das main.py Skript gestartet welches übergeifend mit den anderen Skripten zusammenarbeitet. Bevor dieses andere Methoden startet, wird die Ordnerstruktur des Projekts erstellt.

Code 2: Auszug aus „main.py“. Erstellt die Ordnerstruktur.

def createDir(dirname):

    Path(dirname).mkdir(parents = True, exist\_ok = True)

def createDirs(dirnames):

    for dirname in dirnames:

        createDir(dirname)

def createAllDir():

    createDirs([

        "../dat/temp/",

        "../dat/temp/googleTrends/",

        "../dat/temp/countryBorders/",

        "../dat/temp/giniData/",

        "../out/",

        "../out/caseNumberHistoryPerCountry/",

        "../out/maps/",

        "../out/maps/giniCaseCoef/",

        "../out/maps/giniDeathCoef/",

        "../out/healthSpending/",

        "../out/giniCoefficient/"

    ])

download.py kümmert sich dabei um den Download der unterschiedlichen Datensätze und überschreibt zusätzlich möglicherweise veraltete, schon vorhandene Daten. Im nächsten Schritt werden die Informationen alle durch load.py in den Arbeitspeicher geladen, um sie verarbeiten zu können. Dabei werden einige Daten außerdem gruppiert bzw. als Dictionary zwischengespeichert. Problem dabei war, dass die .csv Tabellen, der Portale teilweise eine UTF-8 BOM Zeichenkodierung hatte. Diese Kodierung führt in den Skripten zu unerwünschten Fehlern, darum wird die Byte Abfolge, die am Anfang der Datei steht, entfernt. Dadurch erhält man dieselbe Datei im normalen UTF-8 Format. Die UTF-8 BOM Kodierung ist mittlerweile in den meisten Fällen überflüssig. Sie dient lediglich dazu die Interpretation verschiedener Zeichenkodierter Daten zu erleichtern.

Code 3: Auszug aus „download.py“. Hier werden die ersten 4 Zeilen inkl. der BOM Byte Folge entfernt.

dataFrame = pd.read\_csv(path + "WorldBankGiniIndex.csv", skiprows=4)

dataFrame.to\_csv(path + "WorldBankGiniIndex.csv", index=False)

Die im Arbeitsspeicher vorliegenden Daten können nun visualisiert werden.

Die beiden Skripte draw.py und plot.py übernehmen diese Aufgabe. Durch plot.py werden Graphen erstellt, die einen besseren Überblick über die genauen Zahlen geben. Aktuell werden Diagramme zum zeitlichen Ablauf der täglichen und der totalen Menge der Corona Fälle weltweit erstellt. Auf diesem Graph werden außerdem die Anzahl der Google Anfragen dargestellt, um das Interesse der Länder bzgl. Des Virus aufzuzeigen. Durch draw.py lassen sich ganze Karten erstellen, die einen Farbverlauf beinhalten. In diesem Projekt werden Karten gezeichnet, die die Corona Fälle visualisieren, aber auch das Verhältnis der Ausgaben eines Gesundheitsystems pro Einwohner und das Verhältnis des Gini-Koeffizienten und der COVID-19 Fälle. Um nun den zeitlichen Verlauf kenntlich zu machen wurden sogenannte GIFs erstellt. Ein GIF verhält sich wie eine Daumenkino und zeigt mehrere Bilder hintereinander. Daher bietet sich diese Dateiformat an, um den Ablauf der Corona Krise als Weltbild darzustellen. Als nächstes werden die gröberen Probleme, die während der Bearbeitung des Projekts aufgetreten sind, beschrieben.

## Probleme

Die erste große Frage, die sich am Anfang des Projekts stellte, ist die Visualisierung der Ergebnisse. Mit Graphen bzw. Diagrammen ist es beispielsweise einfacher einen zeitlichen Verlauf zu zeigen, als mit ganzen Weltkarten. Jedoch sollten möglichst alle Länder bildlich verständlich erklärt werden. Daher wurde sich in dieser Arbeit für Weltkarten entschieden, die den zeitlichen Verlauf der Krise als GIF ablaufen lassen. Zusätzlich werden die Karten bzw. GIFs durch Graphen mit genaueren Wertangaben unterstützt. Ein weiteres Problem war, dass manche .csv Tabellen, wie schon vorher, erwähnt im UTF-8 BOM Format ankommen und folglich konvertiert werden müssen. Da sich das Problem durch Löschen der ersten Zeichen der Tabelle lösen lässt, handelt es sich hier um eine kleinere Komplikation. Die Schwierigkeit bestand lediglich darin den Fehler zu erkennen. Weiterhin gab es bei der Google API Erschwernisse durch deren Sicherheitsmaßnahmen, da das Skript zu viele Anfragen an die Server stellt. Dies wurde durch eine Pausierung des Skripts bei einem „TooManyRequests“ Error gelöst. Außerdem gab die API zwei Länder nicht im richtigen Ländercode zurück. Diese müssen als Sonderfall betrachtet werden und der eindeutige Code muss umbenannt werden, damit diese im gesamten Projekt einheitlich sind. Das ist wichtig, da die Skripte die unterschiedlichen Werte über den Zeitverlauf durch diesen Ländercode zusammenfügen kann. Eine weitere Abweichung wurde bei dem Land Kosovo festgestellt. Dieses Land wurde amtlich noch nicht als solches anerkannt und deshalb wird vom Skript dieser Ländercode als Falsch erkannt, da dieser ja formell noch nicht besteht. Durch Einzelfallbehandlung kann aber auch dieses Problem gelöst werden. Im folgenden Kapitel werden nun die Ergebnisse der Arbeit beschrieben.

# Ergebnisse

Als Ergebnis präsentiert die Projektarbeit drei Weltkarten GIFs. Darunter die Ausbreitung der COVID-19 Fälle, die Ausgaben an Gesundheitssysteme (pro Kopf gerechnet) und den Gini-Koeffizienten im Verhältnis mit den Corona Fällen. Außerdem gibt es für jedes Land einzeln ein Diagramm, das die totale Anzahl der Patienten, die täglichen Fälle und durch die Google Anfragen das Interesse an Corona aufzeigt. Bezüglich der Google Anfragen ist zu beachten, dass einige Länder in großen Teilen wenige Internetzugänge haben und deshalb das „Interesse“ recht niedrig ist. ZU DEN CORONA MAPS KANN ICH LEIDER NOCH NICHTS SCHREIBEN.

* Was sind die wichtigsten Beiträge, die Ihre Arbeit liefert?
* Graphische Darstellung und Beschreibung

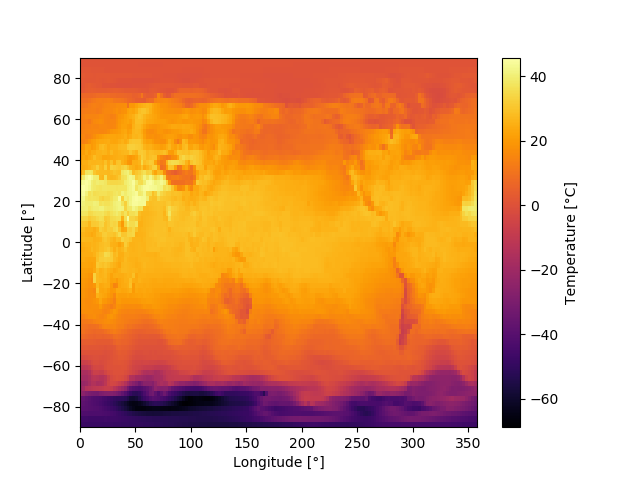


Abb. 1: Globale Temperaturmesswerte

# Schlussfolgerungen

Zusammenfassend lässt sich sagen, dass sich der Historienverlauf der Corona Ausbreitung sehr gut durch die GIF Bildabfolge darstellen lassen konnte. Auch die Diagramme veranschaulichen die Zahlen und Fakten, die bis jetzt zur COVID-19 Thematik gesammelt werden konnten sehr gut. Die größte Problematik bestand darin die Daten möglichst wahrheitsgemäß darzustellen. Außerdem ist es kaum vereinbar alle Daten in eine Karte bzw. einen Graph darzustellen. Mehrere visuelle Ableger werden aufgrund der Datenmenge auf jedenfall benötigt. Die Farbcodierung der Karten könnte noch verbessert werden. Dies gestaltet sich schwireig, da sehr kleine Länder mit beachtet werden und evtl. auf der Karte gar nicht sichtbar werden.

* War die Methodik erfolgreich und wurde die Problemstellung beantwortet?
* Wo steckten die größten Herausforderungen?
* Was könnte noch verbessert werden?

# Anhang

* Größere Abbildungen
* Code