Corona Datenverarbeitung

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

**Studiengang:**  Angewandte Informatik

**Semester:** SS 2020

**Datum der Abgabe:** 07.05.2020

|  |  |  |
| --- | --- | --- |
| 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 Corona Virus 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 unterschiedlichen Verläufen kommt. 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, kurz ECDC, bezogen werden. Jedoch sollen Korrelationen zu verschiedenen anderen Faktoren festgestellt werden. Daher reichen die Daten des ECDCs allein nicht aus und es müssen zusätzlich 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 sich das Interesse der Bevölkerung im Verlauf der Pandemie verhält. Im folgenden Kapitel wird unsere Vorgehensweise knapp beschrieben.

## Lösungsansatz

Da es sich beim Corona Virus um eine Pandemie handelt, muss der gesamte Zeitverlauf möglichst Welweit betrachtet werden. Die Suche nach Geodaten ergab, dass die Corona Verbreitungsdaten vom ECDC bezogen werden. Das Google Interesse, welches sich aus den Suchanfragen der Suchmaschine in den jeweiligen Ländern ergeben, wird von der Google API bezogen. Außerdem werden die Gini-Koeffizienten (Index für Einkommens- und Vermögensverteilung) der verschiedenen Länder von der „World Bank Group“ als csv Tabelle angefragt. Sowohl die Datenverarbeitung, als auch die Visualisierung wird in Python realisiert. Es wurden aus Übersichtsgründen mehrere Skripte erstellt, die den Download, die Verarbeitung und die Darstellung vollautomatisiert durchführen. Die Skripte sind so konzipiert, dass sie durch das Ausführen einer Windows batch Datei gestartet werden. 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.

# Methodologie

## Datenbeschaffung

Vom ECDC konnten Daten im csv Format heruntergeladen werden. Dieser Datensatz beinhaltet den Anstieg der Corona Fälle in nahezu allen Ländern weltweit. Diese Daten bilden die Grundlage dieser Studienarbeit, da alle weiteren Analysen sich auf die Corona Fälle beziehen. 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 eine nicht öffentliche Google API. Um den Zugriff auf die Daten zu erleichtern wird die Python Bibliothek pytrends verwendet. Diese sollen das Interesse der Bevölkerung der einzelnen Länder wiederspiegeln.

Zusätzlich wurden Daten zum Gini-Koeffizienten einbezogen, um aufzuzeigen, wie sich COVID-19 auf Länder mit unterschiedlich starker Einkommensverteilung auswirkt. Diese Daten kamen von der World Bank ebenfalls im csv Format.

Als letztes wurden von der WHO Werte bezogen, welche die weltweiten Ausgaben für die Gesundheitsysteme in verschiedenen Ländern darstellen. Diese sollen später im Vergleich mit den Corona Fällen Aufschluss ergeben, ob die per capita Ausgaben für das Gesundheitssystem im Zusammenhang mit den COVID-19 Fällen stehen. Nachdem die richtigen Datensätze nun verfügbar sind, wird nun auf die Aufbereitung der Daten eingegangen.

## Datenverarbeitung

Die erhaltenen Daten werden automatisiert von den verschiedenen Skripten verarbeitet. Die Software kann über eine batch Datei gestartet werden, welche sich außerdem um die Installation benötigter Python Bibliotheken kümmert. (siehe Code 1)

@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

*Code 1: „run.bat“*

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. (siehe Code 2)

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/"

    ])

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

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 und als Dictionary zwischengespeichert. Die im Arbeitsspeicher vorliegenden Daten können nun visualisiert werden.

In preprocessing.py befinden sich Methoden, welche die geladenen Daten verarbeiten und für die visuelle Aufbereitung vorbereiten. Beispielsweise wird in Code 3 die Bevölkerung eines Landes für ein bestimmtes Jahr aus den Bevölkerungsdaten der UN bezogen.

def extractCountryPopulationForYear(populationRaw, year):

    population = {}

    for countryPopulation in populationRaw[year]:

        if countryPopulation["VarID"] == '2':

            country = pycountry.countries.get(

                numeric=countryPopulation["LocID"].zfill(3))

            if country != None:

                population[country.alpha\_2] = float(

                    countryPopulation["PopTotal"])

            else:

                log.logWarning(

                    "Countrycode of " + countryPopulation["Location"] + "couldnt be parsed")

    return population

*Code 3: Auszug aus „preprocessing.py“. Extrahiert die Population des Jahres für alle Länder.*

Die beiden Skripte draw.py und plot.py übernehmen die Visualisierung der Daten. 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 gesamten Menge der Corona Fälle weltweit erstellt. Auf diesem Graph wird außerdem das Interesse an Corona anhand der Google Suchanfragen dargestellt. 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 dieses Darstellungsform an, um den Ablauf der Corona Krise als Weltbild abzubilden. Als nächstes werden die gröberen Probleme, die während der Entwicklung des Projekts aufgetreten sind, beschrieben.

## Probleme

Die erste große Frage, die sich am Anfang des Projekts stellte, ist die Visualisierung der Ergebnisse. Mit Graphen beziehungsweise 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 und GIFs durch Graphen mit genaueren Wertangaben unterstützt. Ein weiteres Problem war, dass manche csv Dateien im UTF-8 BOM Format ankommen und folglich konvertiert werden müssen. 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.

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

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

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

Da sich das Problem durch Löschen der ersten Zeichen der Datei 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. Hierfür wird eine Schnittstelle verwendet, welche nicht mehr öffentlich dokumentiert ist. Somit musste die Funktionsweise selbst ergründet werden, wobei es zu verschiedenen Authentifikationsschwierigkeiten kam. Als letztendliche Lösung wurde die python Bibliothek pytrends gewählt, welche einen verhältnissmäßig leichten Zugang zu dieser Google trends API ermöglicht. Eine weitere Komplikation stellte der ErrorCode 429 (To many requests) dar. Dieser Fehler tritt auf, da die Google API nur eine limitierte Anzahl an Zugriffen zulässt und sich so vor Überlastung schützt. Die Lösung bestand darin einen einminütige Anfragepause einzusetzen. Weiterhin gab der Datensatz des ECDCs zwei Länder nicht im richtigen Ländercode zurück und hatte eine Unterscheidung zwischen Japan und dem japanischen Schiff (diamond princess JPG11668). Diese müssen als Sonderfälle betrachtet werden und der identifizierende Ländercode muss dementsprechend angepasst werden. Hierbei wurde die ISO 3166 (alpha 2) gewählt. Das ist wichtig, da die Skripte die unterschiedlichen Werte über den Zeitverlauf durch diesen Ländercode zusammenfügen können. 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. Zunächst war eine detaillierte Ansicht für Taiwan geplant. Diese scheiterte aus politischen Gründen, da durch Chinas Einfluss keine vertraulichen Daten für Taiwan auffindbar waren. Das liegt daran, dass China Tawian nicht als autonom Staat ansieht, sondern als Teil der Republik. Im folgenden Kapitel werden nun die Ergebnisse der Arbeit beschrieben.

# Ergebnisse

# 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 möglich alle Daten in einer Karte oder einem Graphen darzustellen. Mehrere visuelle Aufbereitungen werden aufgrund der Menge der Unterthematiken benötigt. Die Farbcodierung der Karten könnte noch verbessert werden. Dies gestaltet sich schwierig, da kleine Differenzen in einer globalen Farbskala nicht ersichtlich werden würden. Die Analyse der Corona Fälle beziehungsweise der Tode im Vergleich zu den pro Kopf Ausgaben der Länder für das Gesundheitssystem ist strittig, weil unter anderem ein Land mit schlechtem Gesundheitssystem Neuinfektionen kaum bis gar nicht feststellen und melden kann. Für die Analyse des Gesundheitssystems müssten somit andere Kennzahlen und Daten einbezogen werden. Da es sich um ein sehr aktuelles Thema handelt ist es wahrscheinlich, dass die Datenquellen Ihre API-Endpunkte verändern könnten, was ein Problem für das Skript darstellt.

# Anhang

* Größere Abbildungen
* Code