Corona Datenverarbeitung

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

**Studiengang:**  Angewandte Informatik

**Semester:** SS 2020

**Datum der Abgabe:** 30.04.2020

|  |  |  |
| --- | --- | --- |
| Name | Matrikelnummer | Email |
| 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](#_Toc39141843)

[1.1 Motivation 3](#_Toc39141844)

[1.2 Problemstellung 3](#_Toc39141845)

[1.3 Lösungsansatz 4](#_Toc39141846)

[2 Methodologie 4](#_Toc39141847)

[2.1 Datenbeschaffung 4](#_Toc39141848)

[2.2 Datenverarbeitung 6](#_Toc39141849)

[2.3 Kapitel 3 [Name d. Autors] 6](#_Toc39141850)

[3 Ergebnisse [Name d. Autors] 6](#_Toc39141851)

[4 Schlussfolgerungen [Name d. Autors] 7](#_Toc39141852)

[5 Anhang 7](#_Toc39141853)

# 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 soll im nächsten Unterkapitel genauer veranschaulicht werden.

## 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 das 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. Es wurden lediglich Deutschland, USA und Taiwan in Detailbetrachtung bearbeitet. 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 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 das Ausführen einer Windows .bat Datei gestartet wird. Die Python Bibliothek matplotlib unterstützt das Zeichnen von Graphen, mit der es möglich ist einfache 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.

Versuchen Sie, die Themenfelder klar in Unterkapitel abzugrenzen. Das können Bearbeitungsschritte wie etwa Datenakquise, -verarbeitung und Analysen sein.

Zudem wird empfohlen, die Textabsätze mit „Topic-Sentences“ zu beginnen. Diese definieren geben eine Hauptaussage und die restlichen Sätze eine passende Argumentation.

Kapitel können mit einer schlanken Zusammenfassung beginnen. (Bspw. Dieses Kapitel beinhaltet alle Schritte, die zur Sammlung der Ausgangsdaten sowie ihrer Speicherung beitragen.)

## 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. Danach wird das main.py Skript gestartet welches übergeifend mit den anderen Programmen zusammenarbeitet.

Abbildungen und Tabellen müssen im Text referenziert werden, wie z. B. siehe Abb 1. Hierzu sind Querverweise ein hilfreiches Werkzeug, um die Nummern während der Bearbeitung nicht zu „verlieren“.

Tabelle : Exemplarische Ziffern

|  |  |  |  |
| --- | --- | --- | --- |
| Spalte | Spalte | Spalte | Spalte |
| 1 | 2 | 3 | 4 |
| 5 | 6 | 7 | 8 |

## Kapitel 3 [Name d. Autors]

# Ergebnisse

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

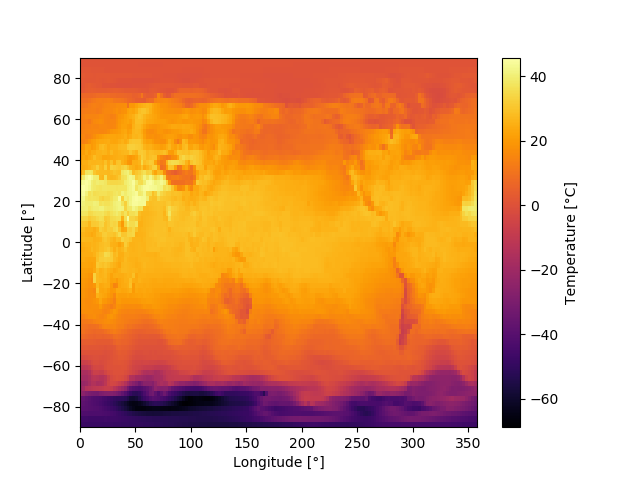


Abb. : Globale Temperaturmesswerte

# Schlussfolgerungen [Name d. Autors]

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

# Anhang

* Größere Abbildungen
* Code