



# **Data Science in Life Science**

**SS20**

**Quentin Quarantino**



Copyright © Quentin Quarantino

PUBLISHED BY PUBLISHER

BOOK-WEBSITE.COM

Licensed under the Creative Commons Attribution-NonCommercial 3.0 Unported License (the “License”). You may not use this file except in compliance with the License. You may obtain a copy of the License at <http://creativecommons.org/licenses/by-nc/3.0>. Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an “AS IS” BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

*First printing, March 2019*

# Contents

I	Part 1	
1	Introduction .....	7
2	Topic 1: Spreading Models .....	9
2.1	Background	9
2.2	Data and Methods	9
2.3	Discussion & Results	10
3	Topic 2: Data-based Time Series Prediction .....	13
3.1	Background	13
3.2	Data and Methods	13
3.3	Results	14
3.4	Discussion	14
4	Topic 3: Risk Factor Analysis .....	15
4.1	Background	15
4.2	Data and Methods	15
4.3	Results	15
4.4	Discussion	15
5	Topic 4: Diagnostic .....	17
5.1	Background	17

5.2	Data and Methods	17
5.3	Results	17
5.4	Discussion	23
<b>6</b>	<b>Topic 5: Origin Analysis</b> .....	<b>25</b>
6.1	Background	25
6.2	Data and Methods	25
6.3	Results	25
6.4	Discussion	26

## II

## Part 2

<b>7</b>	<b>Introduction</b> .....	<b>31</b>
7.1	Goal of the Project	31
7.2	Outcome	31
<b>8</b>	<b>Tasks</b> .....	<b>33</b>
8.1	Part1	33
8.2	Part2	36
8.3	Part3	40
8.3.1	Loading in the Student Dataset .....	40
8.3.2	Preprocessing, Clustering and Results .....	40
8.3.3	Creating a word cloud .....	45
	<b>Bibliography</b> .....	<b>47</b>
	Articles	47
	Books	47
	<b>Index</b> .....	<b>49</b>

# Part 1

<b>1</b>	<b>Introduction .....</b>	<b>7</b>
<b>2</b>	<b>Topic 1: Spreading Models .....</b>	<b>9</b>
2.1	Background	
2.2	Data and Methods	
2.3	Discussion & Results	
<b>3</b>	<b>Topic 2: Data-based Time Series Prediction .....</b>	<b>13</b>
3.1	Background	
3.2	Data and Methods	
3.3	Results	
3.4	Discussion	
<b>4</b>	<b>Topic 3: Risk Factor Analysis .....</b>	<b>15</b>
4.1	Background	
4.2	Data and Methods	
4.3	Results	
4.4	Discussion	
<b>5</b>	<b>Topic 4: Diagnostic .....</b>	<b>17</b>
5.1	Background	
5.2	Data and Methods	
5.3	Results	
5.4	Discussion	
<b>6</b>	<b>Topic 5: Origin Analysis .....</b>	<b>25</b>
6.1	Background	
6.2	Data and Methods	
6.3	Results	
6.4	Discussion	





## 1. Introduction

In this weeks project each group member was assigned one overarching topic pertaining to the current Covid-19 epidemic. The current epidemic is globalized, with severe consequences to social, health and economic order. As of today 212 countries are affected, with a total of 4,215,274 confirmed cases and a death toll of 284,672 [9]. Within each topic a short introduction to the general concept is given. The understanding of these concepts is then deepened by real world code examples, showing a glimpse of what is possible in each topic in regards to Covid-19.







## 2. Topic 1: Spreading Models

### 2.1 Background

Modeling the spread of infectious diseases is not only an essential tool in understanding the transmission rates and the trajectory of future cases but also has a significant influence on the appropriate guidelines to control the course of an epidemic. The approaches towards modeling the spread can range from computational models (e.g. agent-based) to mathematical modeling (e.g. compartmentalized models). In this short introduction we will focus on a SIR-model which is part of the compartmentalized subgroup. These models follow a deterministic pattern where each subpopulation is divided into groups. In SIR-models each letter stands for one group:  $S$  = *susceptible*,  $I$  = *infectious* and  $R$  = *recovered/death*. Then it follows that for each time independent point  $t$  the rates for each subgroup can be calculated by:

$$\begin{aligned}dS/dt &= \nu N - \beta SI/N - \mu S \\dI/dt &= \beta SI/N - \gamma I - \mu I \\dR/dt &= \gamma I - \mu R\end{aligned}$$

with  $\gamma$  denoting the time rate of death/recovery,  $\beta$  denoting the number of new infections one case causes per time point  $t$ ,  $\mu$  denoting general death rate and  $\nu$  denoting being the birthrate. hey

### 2.2 Data and Methods

This code is based on the kaggle notebook from <https://www.kaggle.com/lisphilar/covid-19-data-with-sir-model>. It uses python and a basic framework of libraries e.g pandas, sklearn, datetime etc.. The main data used is from the World Health Organization showing novel corona infections by country. Furthermore supplementary data is used to include the age pyramid for each country. The WHO Data set is preprocessed to include the variables: Date, Country, Province, Confirmed, Infected, Deaths and Recovered. A first visualization shows the global rate of infected, deaths and recovered people (Figure 2.1). Next the growth factor is calculated, which is given by:  $G_n/G_{n-1}$  with  $G$  = *confirmedcases*. Countries with growth factor higher then one have an increasing number

of cases. In contrast growth factor lesser then one shows a declining number of cases. The actual analysis is done for 5 countries: Italy, Japan, India, USA and New Zealand. Giving one case as an

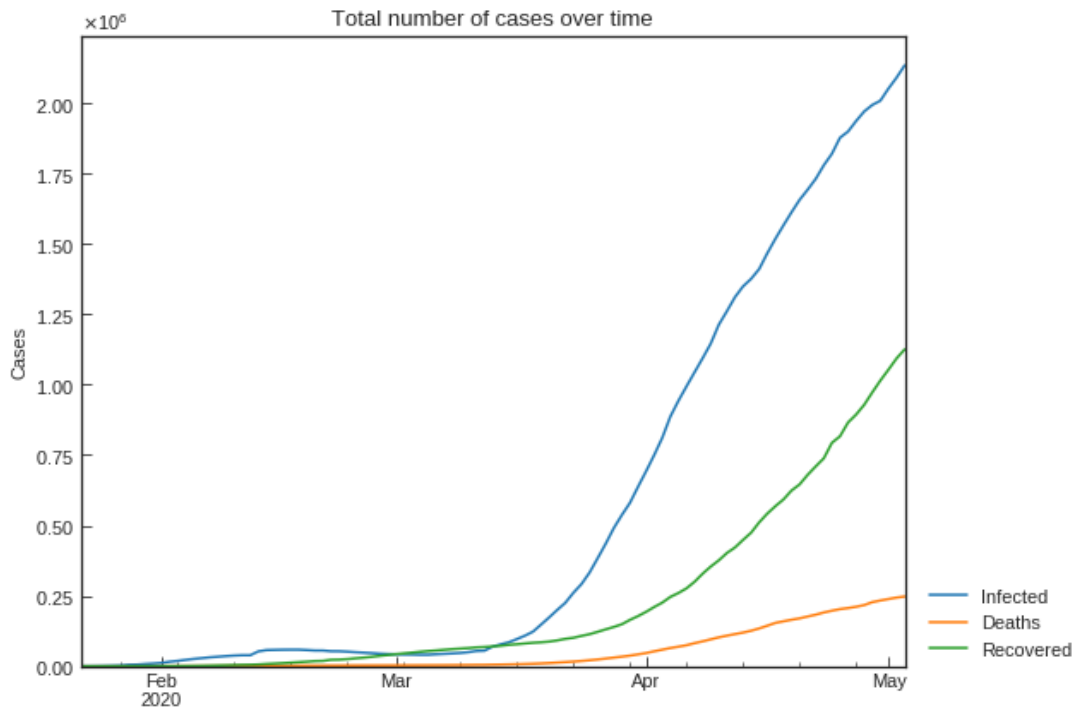


Figure 2.1: Global infections, deaths and recovered people

example as a first step a S-R trend is plotted (Figure 2.2). It shows the trend of susceptible against recovered people. 5 change points can be identified. Next the SIR-F model parameters are estimated for each change point. As a last step the changes in the  $p$  value are contrasted with measures taken by the country. While these results are interesting SIR modeling also has its limitations.

## 2.3 Discussion & Results

Even though the SIR-Model is one of the most basic infectious disease models available it can show promising results with careful consideration for parameter selection and data processing steps. In the case of Italy 3 measures could be shown to reduce the  $p$  value: quarantine of person contacted with positive patients, school closure and lock-down. It also has to be considered that the SIR model is based on very basic assumptions. For example the number of susceptible people is treated as fixed as well as the rates of change.

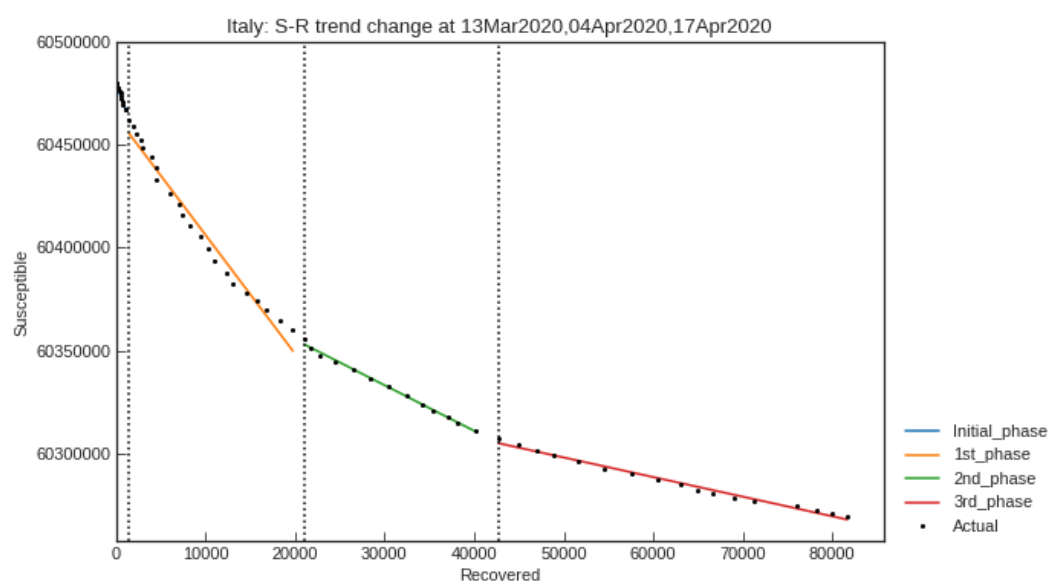


Figure 2.2: Trend of susceptible people versus recovered. 5 distinct change points can be identified.



## 3. Topic 2: Data-based Time Series Prediction

### 3.1 Background

Many governments around the world are building their political decisions around the number of current confirmed cases of people infected by COVID-19. Nonetheless, not only the current number of confirmed cases is from greater interest, but also how the virus spreads in the future. One way of forecasting the spread of the virus is by using data-based time series prediction.

### 3.2 Data and Methods

Therefore machine learning models are calibrated using publicly available data sources like the WHO health report. Time series forecasting can be framed as a supervised learning problem. Other than agent-based spreading simulation such as the SIR model, the models used here do not simulate a population. The forecasting is performed using python's numpy and sklearn libraries. At first the data is downloaded. Since no data points are missing no preprocessing else than converting integers into date times and reorganizing dataframes is performed. The data contains for a wide range of countries the number of infected people per day starting January 22. A support vector machine model is implemented to forecast the number of infected people. The parameters that have been set can be seen in figure 3.1. The test and test training data sets are generated by splitting them without shuffling them, such that the time series is preserved.

```
[ ] 1 # svm_confirmed = svm_search.best_estimator_  
2 svm_confirmed = SVR(shrinking=True, kernel='poly', gamma=0.01, epsilon=1, degree=5, C=0.1)  
3 svm_confirmed.fit(X_train_confirmed, y_train_confirmed)  
4 svm_pred = svm_confirmed.predict(future_forecast)
```

Figure 3.1: Parameters set for SVM Model.

The model has been trained using the first 75 days since January 22. In figure 3.2 it can be seen that the model over estimates the number of confirmed infections by over 1.5 Million.

### 3.3 Results

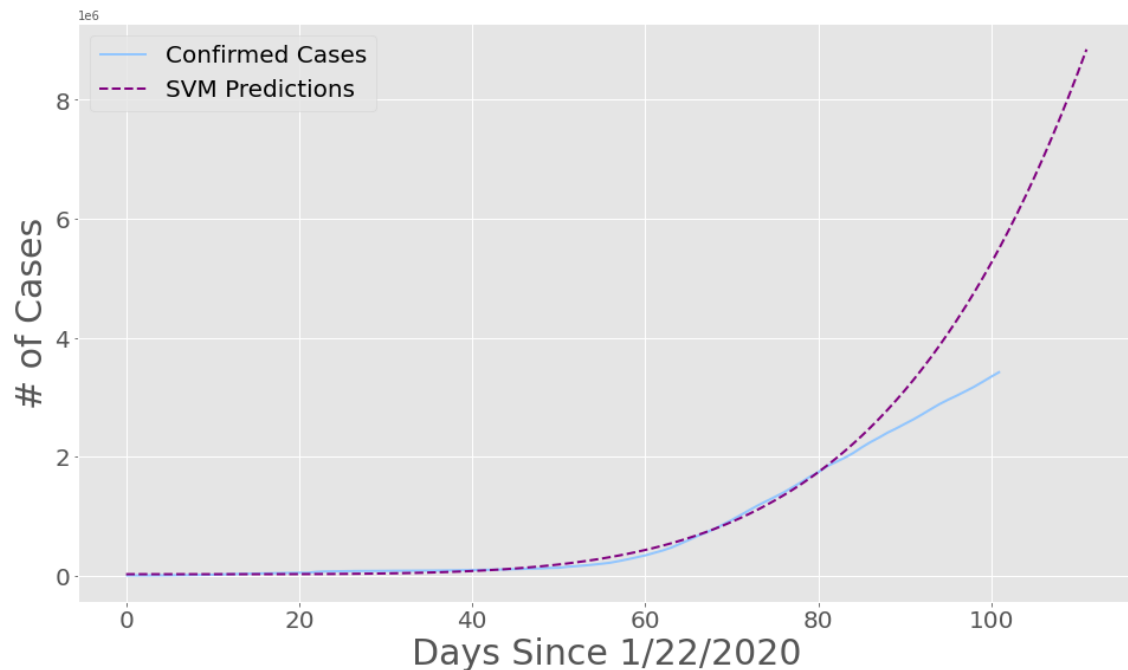


Figure 3.2: Comparison between the observed and the estimated number of infected people.

### 3.4 Discussion

The shown results are quite underwhelming. This fact has several reasons. One pandemic curves usually increase at the beginning exponentially but then flatten down e.g. because of restrictions in society to decrease the spread of the virus. The model is trained using data from the beginning of the crises where the number of cases rapidly grow. Based on this assumption the estimated number of infected people overshadows the confirmed number. Nonetheless due to different test capacities around the world the estimated might be closer to the real number than it seems to be the case shown in figure 3.2. Still the used model was quite simple and no testing was shown how the parameters were found. A more complex model may gives a better insight to the spread of the virus.





## 4. Topic 3: Risk Factor Analysis

### 4.1 Background

The potential dangers of 2019-nCoV have prompted a number of studies on its epidemiological characteristics. It is essential to estimate the number of infections (including those that have not been diagnosed), to be able to analyze the spread of the diseases. To better assess the epidemic risk of 2019-nCoV, among the key parameters to be approximated are the basic reproduction number  $R_0$  and the incubation period. Initially we estimate the cumulative number of cases in China outside Hubei province after 23 January, using a time-dependent compartmental model of the transmission dynamics and then we use that number as an input to the global transportation network to generate probability distributions of the number of infected travellers arriving at destinations outside China. Finally using a Galton–Watson branching process to model the initial spread of the virus.

### 4.2 Data and Methods

The analysis is performed using the python libraries namely numpy, matplotlib, kiwisolver, scipy and cycler. We computed the risk of the individual countries with the selected possible parameters like connectivity and  $R_{loc}$  where  $R_{loc}$  is the local reproduction number of the infection, Getting all the combination of the variables from the data surrounds the neighbour of the china to generate the Heat map.

### 4.3 Results

Heat map generated gives the information about the outbreak risks as functions of  $\Theta$  and  $R_{loc}$ , when  $C = 200,000$ . The arrows show the directions corresponding to the largest reductions in the risk, which is shown in the figure 4.1

### 4.4 Discussion

By combining three different modelling approaches helps to assess the risk of 2019-nCoV outbreaks in countries outside of China. This risk depends on three key parameters: the cumulative number of

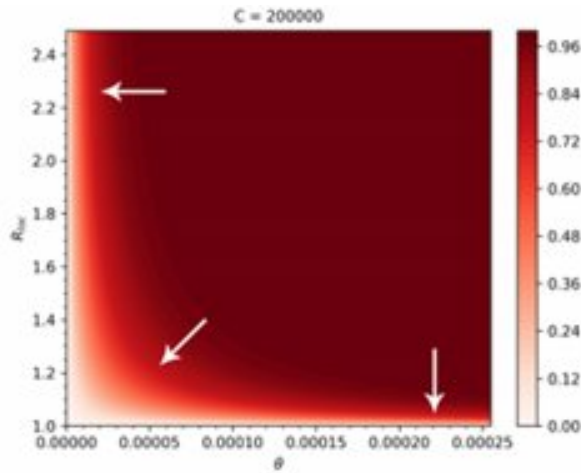
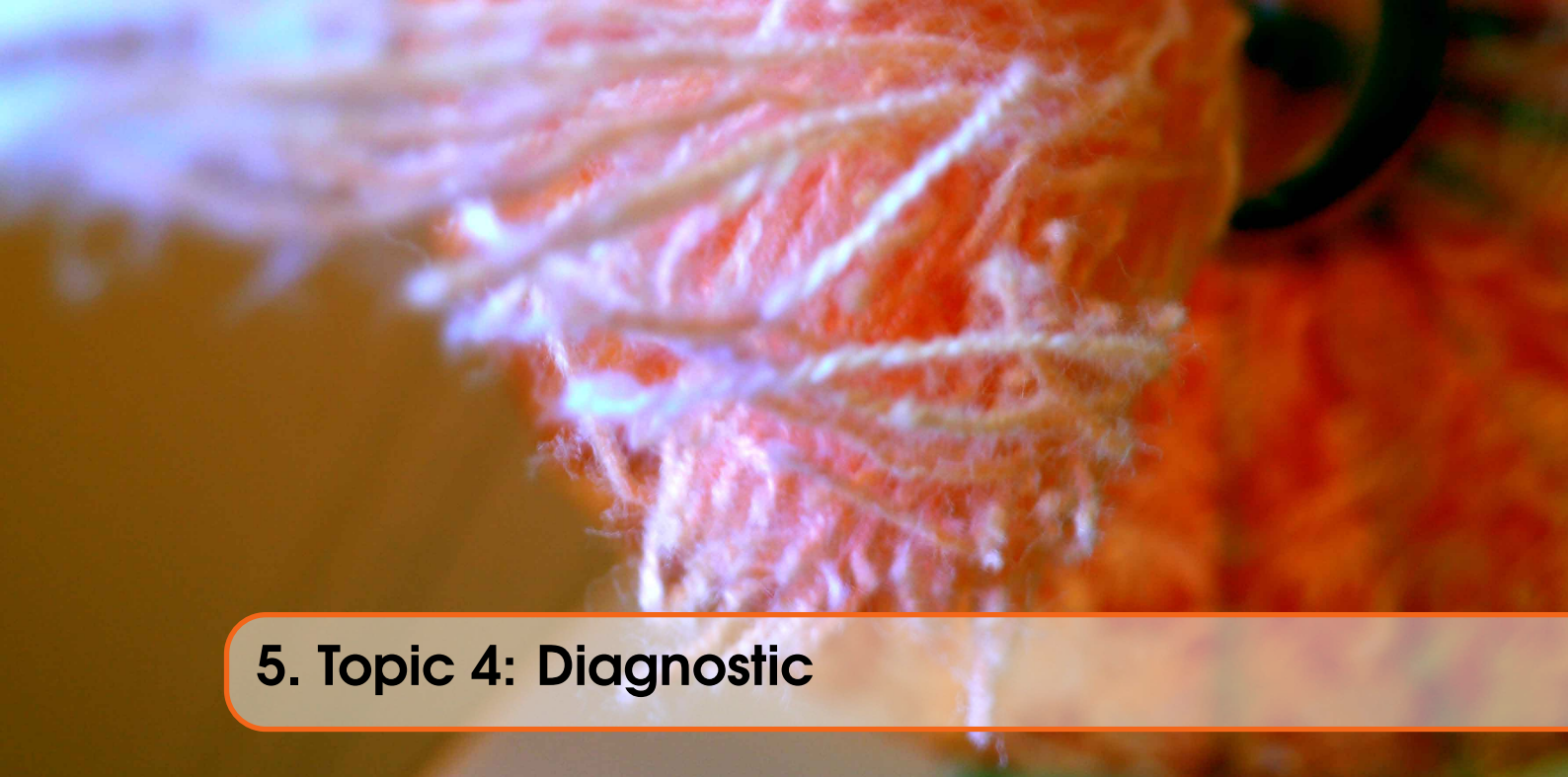


Figure 4.1: Heatmap of the outbreak risks as functions of Theta and Rloc

cases in areas of China which are not closed, the connectivity between China and the destination country, and the local transmission potential of the virus in countries with low connectivity to China but with relatively high  $R_{loc}$ , the most beneficial control measure to reduce the risk of outbreaks is a further reduction in their importation number either by entry screening or travel restrictions. Knowing  $R_{loc}$  and the generation interval are needed not only to have a better quantitative risk estimation, but also for guidance as to which types of control measures may reduce the outbreak risk the most effective.





## 5. Topic 4: Diagnostic

### 5.1 Background

The objective of diagnostics is to help effectively diagnose COVID-19 disease. Diagnostics based on RT-PCR-analysis is not very secure due to a high number of false positives. Diagnosis using X-Ray / CT scan images has objective to help effectively diagnose COVID-19 disease with the help of X-Ray/CT scan images in order to improve speed accuracy and scale of diagnosis.

### 5.2 Data and Methods

X-Ray Detection method reproduced here is done by training a deep learning model using x-ray images (see Figure 5.1 ) with TensorFlow and Keras in Python to predict whether a patient has COVID-19. The full list of required tools are here (see Figure 5.2)

### 5.3 Results

So at first X-Ray data were downloaded from the source and python scripts were downloaded. In the next step, anaconda was installed as it contains a lot of preinstalled packages. In separate environment all the packages listed (see Fig 5.2) were installed with needed help tools and also other needed packages needed (like cuda toolkit and cudnn) to run tensorflow were installed (see Figure 5.3) Then the step augmentation of given X-Ray images was performed for both classes covid positive and normal respectively (see Figure 5.4, 5.5) . In this step using 70 covid and 28 normal X-Ray data were 5088 covid and 2424 normal augmented data generated. In the next step the model was trained and tested using augmented data. The augmented data were divided in train and test data. 80% (6009 data) of augmented data were used as train data and were included in model and 20% (1503 data) of data were used as test data for predictions (see Figure 5.6, 5.7). The model was validated for 1503 test data 100 times with 46 repetitions . Using confusion matrix specificity, sensitivity and accuracy values were estimated and plotted.

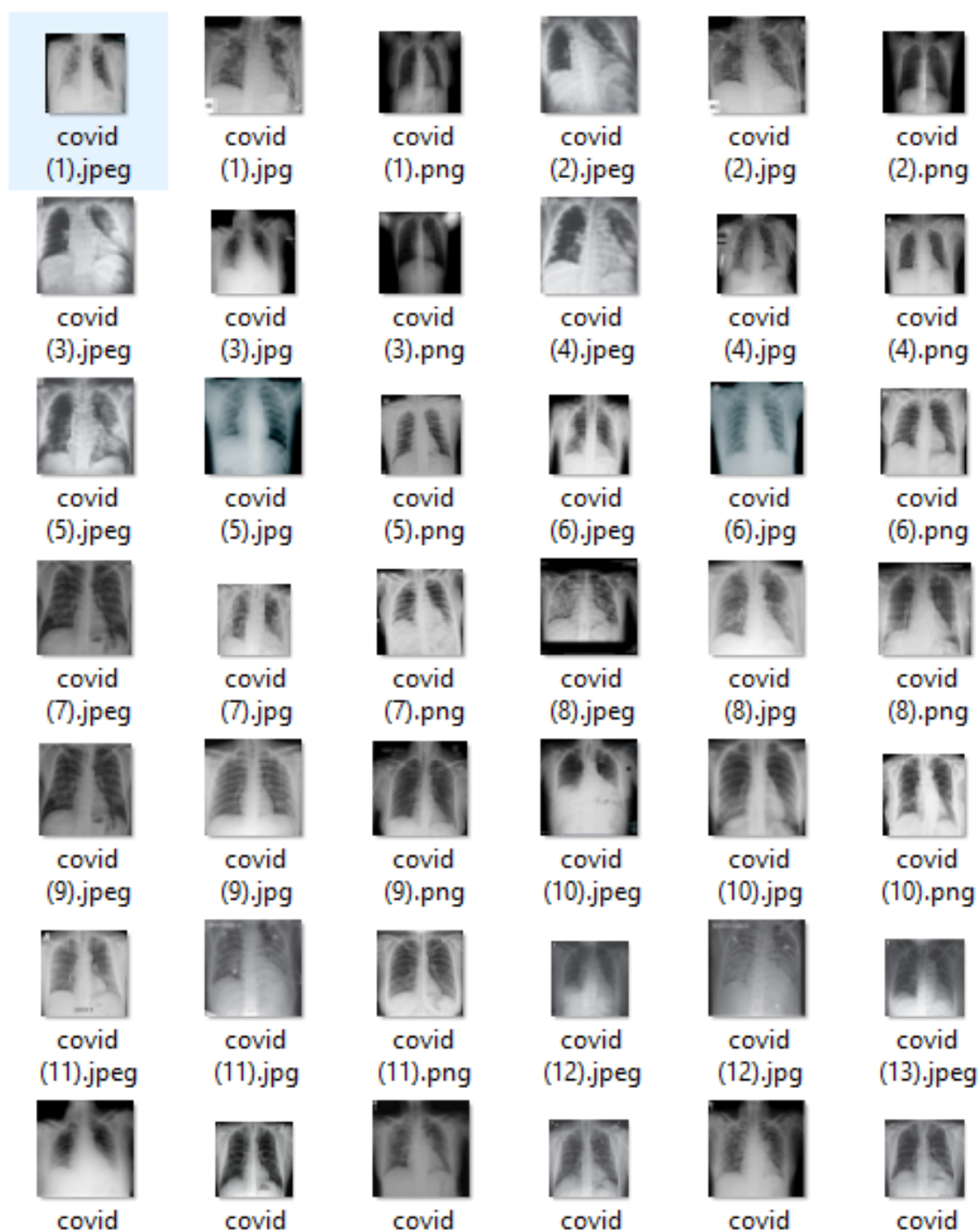


Figure 5.1: Xray data

```
abs1-py==0.9.0
astor==0.8.1
cachetools==4.0.0
certifi==2019.11.28
chardet==3.0.4
cycler==0.10.0
gast==0.2.2
google-auth==1.11.3
google-auth-oauthlib==0.4.1
google-pasta==0.2.0
grpcio==1.27.2
h5py==2.10.0
idna==2.9
imutils==0.5.3
joblib==0.14.1
Keras==2.3.1
Keras-Applications==1.0.8
Keras-Preprocessing==1.1.0
kiwisolver==1.1.0
Markdown==3.2.1
matplotlib==3.2.0
numpy==1.18.2
oauthlib==3.1.0
opencv-python==4.2.0.32
opt-einsum==3.2.0
pandas==1.0.2
Pillow==7.0.0
protobuf==3.11.3
pyasn1==0.4.8
pyasn1-modules==0.2.8
pyparsing==2.4.6
python-dateutil==2.8.1
pytz==2019.3
PyYAML==5.3
requests==2.23.0
requests-oauthlib==1.3.0
rsa==4.0
scikit-learn==0.22.2.post1
scipy==1.4.1
six==1.14.0
sklearn==0.0
tensorboard==2.1.0
tensorflow==2.1.0
tensorflow-estimator==2.1.0
tensorflow-gpu==2.1.0
tensorflow-gpu-estimator==2.1.0
termcolor==1.1.0
urllib3==1.25.8
Werkzeug==1.0.0
wrapt==1.12.1
```

Figure 5.2: List of required tools





```

45/46 [=====>.] - ETA: 29s - loss: 9.0710e-04 - accuracy: 1.00
46/46 [=====] - 1686s 37s/step - loss: 8.8856e-04 - accuracy
: 1.0000 - val_loss: 1.0893e-04 - val_accuracy: 1.0000
[INFO] evaluating network...

```

	precision	recall	f1-score	support
covid	1.00	1.00	1.00	1018
normal	1.00	1.00	1.00	485
accuracy			1.00	1503
macro avg	1.00	1.00	1.00	1503
weighted avg	1.00	1.00	1.00	1503

```

[[1017  1]
 [  0 485]]
acc: 0.9993
sensitivity: 0.9990
specificity: 1.0000
[INFO] saving COVID-19 detector model...
(myenv) PS E:\DSinLS20>

```

Figure 5.7: Screenshot of Validation process

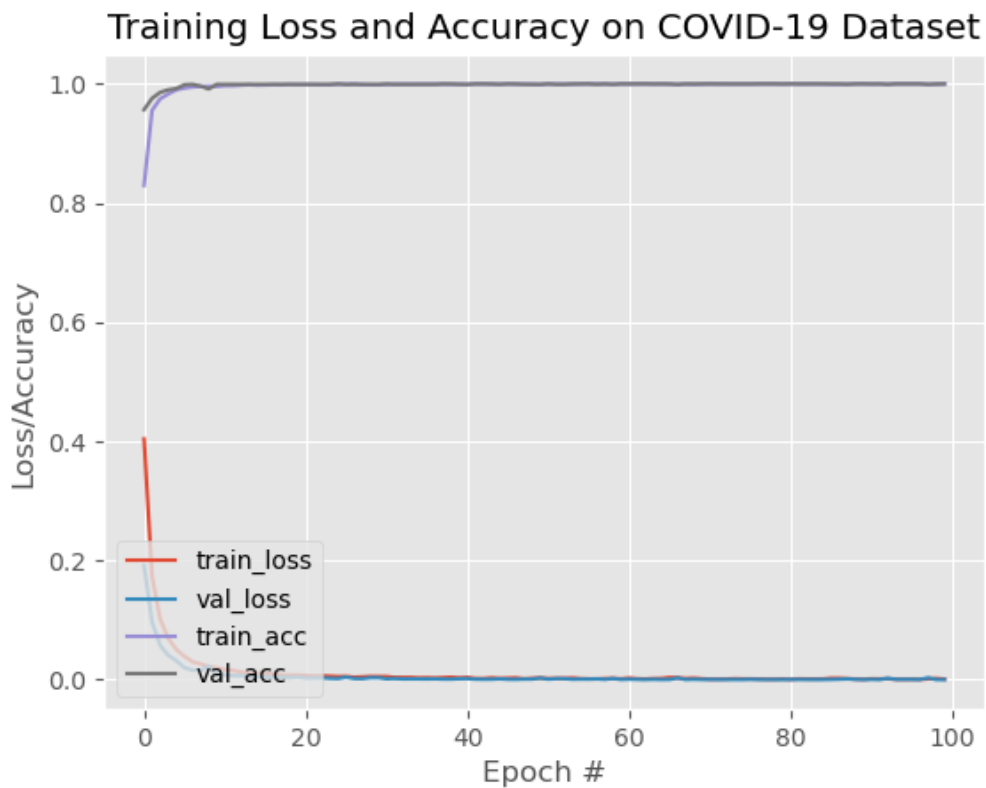


Figure 5.8: Plot of Validation

## 5.4 Discussion

Approach based on deep Learning described here is a very promising tool for Covid-19 detection in lungs. But on the other hand it is very time consuming. All the steps of this pipeline are very time consuming. Augmentation of images took 2.5 hours. Training and testing using model took about 40 hours. The accuracy, specificity and sensitivity are very high (see Figures 5.7, 5.8 ) and prove that this approach is very useful.







## 6. Topic 5: Origin Analysis

### 6.1 Background

Phylogenetic analysis aims to reconstruct phylogenies both for a group of species and also for the individuals within those species. For pathogens, we are interested in the evolution of the diseases at the genetic level, and what this can tell us about their past and present diversity [8]. An important fact about the Coronaviridae family is that its members tend to “jump” from one species to another. When the transmission occurs from a non-human host to a human host it is called zoonosis [3]. The determination of the most recent common ancestor of the human SARS-CoV-2 and the zoonotic transmission can provide important information about biological features, key mutations and properties of the virus. A detailed understanding of how an animal virus jumped species boundaries to infect humans will help in the prevention of future zoonotic events. [1].

### 6.2 Data and Methods

We will compare the genetic sequence of SARS-CoV-2 with other viruses of the Coronaviridae family in different hosts. The following analysis is based on a Github repository of Simon Burgermeister [2]. Six complete genomes were considered, whose names and hosts are listed in Table 6.1. The sequence data (fasta files) were downloaded from the NCBI Virus public library [4]. To compare the genetic sequences, a multiple sequence alignment needed to be performed. Clustal Omega is a software that uses seeded guide trees and HMM profile-profile techniques to generate alignments between multiple sequences. Unfortunately, my local computer was not able to compute the alignment due to RAM exceedance. Therefore, I submitted a request to the online version of Clustal Omega [7]. Based on the resulting alignment, a distance matrix was calculated with the *TreeConstruction* package from Biopython. Afterwards, the same package was used to create the phylogenetic tree based on the UPGMA algorithm.

### 6.3 Results

The resulting phylogenetic tree (Figure 6.1) shows that our human SARS-CoV-2 sequence is most similar to the SARS-like coronavirus sequence of the *Rhinolophus* (horseshoe bat) with a similarity

Accession number	Host	Description
MN996528	H. Sapiens	Human SARS-CoV-2
NC_019843	H.Sapiens	Human MERS-CoV
JQ065048	Anatidae	Ducks, geese and swans
MG772934	Rhinolophus	Horseshoe bats
NC_034972	Apodemus chevrieri	Rodent
KX38909	Gallus gallus	Chicken
MT084071	Manis javanica	Pangolin

Table 6.1: Considered Coronaviridae strains and hosts.

of 96%. The host with the next similar sequence is the *Manis javanica* (Pangolin) with a similarity of 0.89% between their genomes. The human MERS-Cov genome and the SARS-CoV-2 genome share only a sequence similarity of 0.74%.

## 6.4 Discussion

As many early cases of COVID-19 were linked to the Huanan market in Wuhan [10], it is possible that an animal source was present at this location. Given the similarity of SARS-CoV-2 to bat SARS-CoV-like coronaviruses, it is likely that bats serve as reservoir hosts for its progenitor. Although the similarity of 96% to the coronavirus sequence hosted by the *Rhinolophus*, Andersen et al. [1] identified that its spike protein diverges in the receptor binding domain (RBD), which suggests that it may not bind efficiently to the human ACE2 receptor. Furthermore, it is assumed in this and other studies [5, 6] that an intermediate host was probably involved.

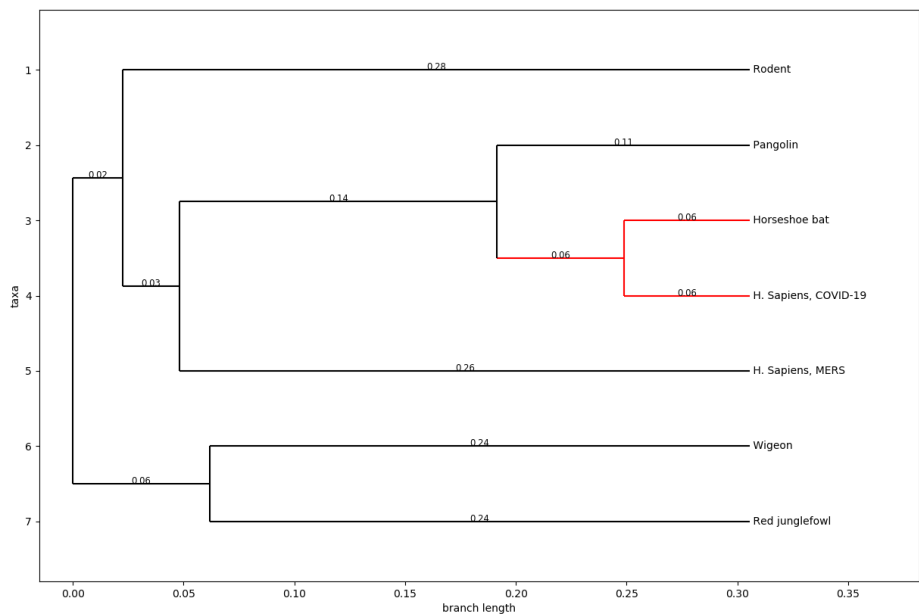
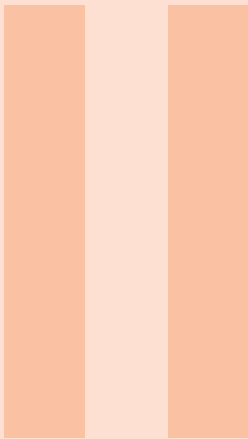


Figure 6.1: Phylogenetic tree of the origin detection analysis.





# Part 2

<b>7</b>	<b>Introduction .....</b>	<b>31</b>
7.1	Goal of the Project	
7.2	Outcome	
<b>8</b>	<b>Tasks .....</b>	<b>33</b>
8.1	Part1	
8.2	Part2	
8.3	Part3	
	<b>Bibliography .....</b>	<b>47</b>
	Articles	
	Books	
	<b>Index .....</b>	<b>49</b>





## 7. Introduction

### 7.1 Goal of the Project

The overwhelming amount of daily published papers correlated to the corona virus makes it difficult, even for health professionals, to keep up with new information about the virus. One way of managing the flood of information is by clustering them according to their topics to simplify the search. Therefore, we have performed a cluster analysis of the CORD-19 dataset, which contains roughly 60.000 articles.

After parsing the body of each article in the dataset, the extracted information is transformed into a feature vector. We afterwards apply dimensionality reduction using PCA and performed a k-means clustering. Subsequently, t-SNE is applied to project the original feature vector into two dimensions such that clusters become visible in the two dimensional space.

Each course participant selected five scientific papers that cluster in the same group as the article they introduced in *Part I*. The submissions have been used to create a new dataset. K-means was also applied to this dataset, to determine the cluster assignments and investigate patterns in the data. Finally, the selected articles of *Part I* were added to the CORD-19 dataset. The clustering was redone to see if the papers will be assigned to the expected clusters. In addition to reperforming the clustering, two methods for selecting the best k value and two distance metrics were compared: Silhouette Scoring vs Distortion and Euclidean vs Cosine Similarity.

### 7.2 Outcome

The five papers we selected in *Part I* were clustered into 3 different groups, whereas three of the papers have been assigned to the same cluster. Comparing both, the method of choosing k by elbow point or silhouette scoring and the distance metrics euclidean and cosine similarity, we determined that silhouette scoring and euclidean distance performed better. 10 clusters with unique topics were found (Table 8.2).







## 8. Tasks

### 8.1 Part 1

The literature clustering pipeline started with the data import of the CORD-19 dataset. Since we wanted to perform the calculations in a Google Colab notebook, we decided to create an API connection to the kaggle database. Using the API, we were able to download and unzip the dataset on our personal Google drive the fastest way possible. The resulting metadata dataframe listed 59.887 entries of coronavirus related publications.

The metadata information are subsequently merged with the body text of the papers that are stored in separated json files. Due to partially missing information only 43.331 entries of the metadata could be merged with the json files. To get an overview of the average text length of the abstracts and the body text information (on which the clustering will be performed) the overall and unique number of words were calculated. The result was an average abstract length of 157 words and an average body text length of 4.528 word (1376 unique). Since the data was uploaded by many different sources, duplicates were present in the dataset. These need to be filtered out such that 30.960 publication remained in the set. The subsequent calculation steps of the pipeline will require very high computing resources. Therefore, we randomly subsamples (seed=42) the dataset to a maximum of 10.000 instances. Unfortunately, we noticed afterwards that both, entries containing null values (1073) and non-english publications (242) were still present in the data. Since these would massively reduce the interpretability of the clustering result, they were also dropped. The final dataset consisted then of 8685 entries.

Another applied preprocessing step was to detect and remove stop words. These are common words in the written text, that do not contribute to the content and act as noise in the clustering procedure. The *spacy* package was used to determine the stopwords. Additionally, a predefined list of stopwords was appended to the list, that contained frequently used words of scientific publication in general. The last step of the preprocessing was to vectorize the cleaned data. Hereby, the string formatted data is converted into a vector-based measure of how important each word is to the instance out of the literature as a whole using the *tf-idf* package. This method creates a very high feature space and since a clustering by k-means need to be performed, a Principle Component Analysis (PCA) was applied to reduce the amount of feature by simultaneously keeping 95% of the

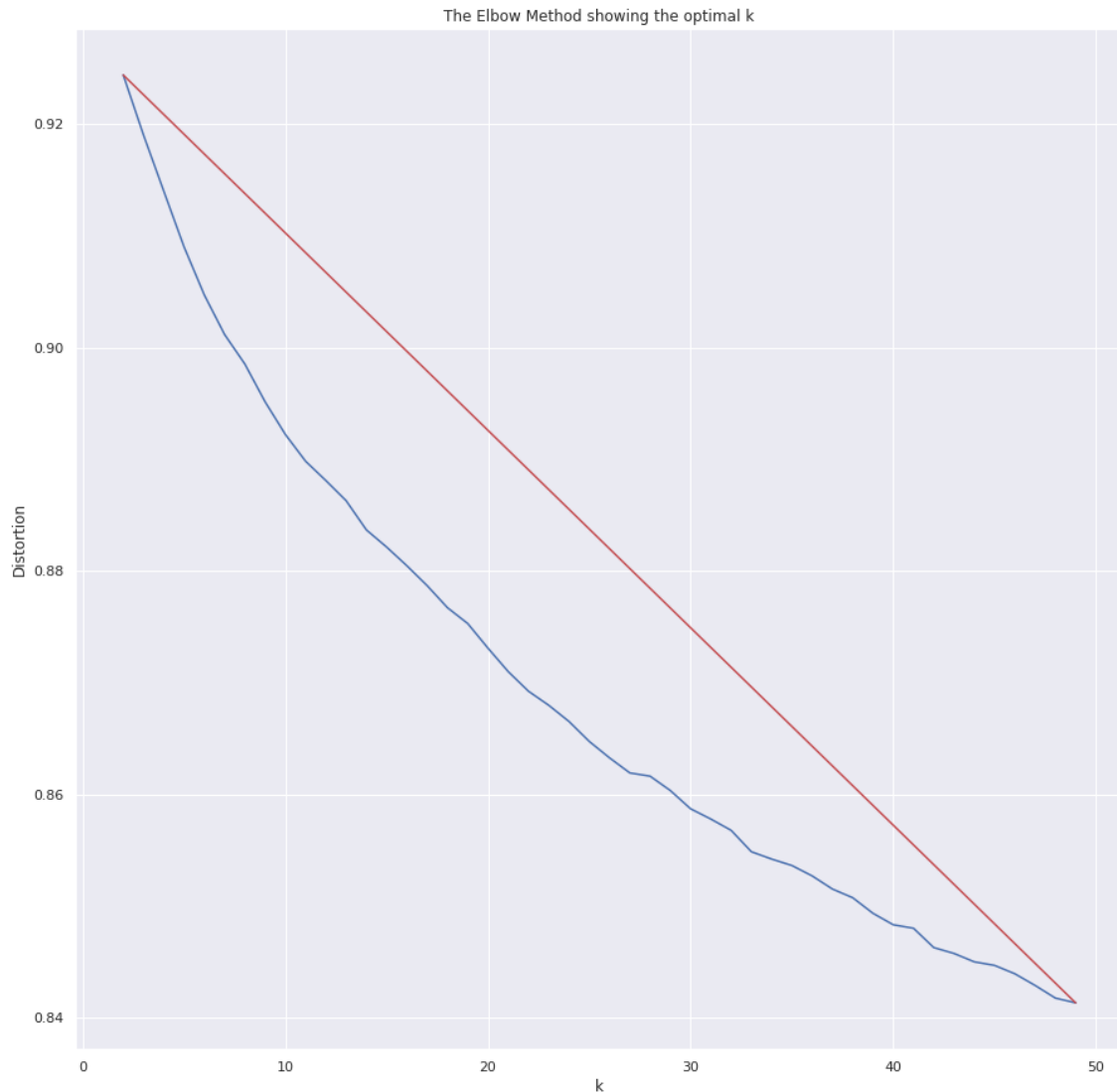


Figure 8.1: The figure shows an elbow plot of the k-means clustering the the distortion on the y-axis and the number of clusters on the x-axis.

data's variance and immensely reducing the algorithm's runtime. The best k number of clusters was determined by iterating through different values of k from two to 50. The resulting elbow plot (Figure 8.1) shows the elbow point at  $k=27$ , which is subsequently used as the best number of clusters.

A t-Distributed Stochastic Neighbor Embedding (t-SNE) was used to reduce the high dimensional features vector to two dimensions. This step provides to possibility to represent the clustering result in a plain coordinate system. The aim of the entire pipeline was to create an interactive bokeh plot. To create the plot, the results of all previous calculations are brought together. The location of each paper on the plot is determined by t-SNE while the label (color) is determined by k-means. Interestingly, the assignments match each other very well, even though they were calculated separately (Figure 8.2). Now, the clusters are calculated, but the information about the kind of papers, that are matched together is still missing. To solve this task, a Latent Dirichment Analysis (LDA) was performed to model the most important topics for each cluster. This information is also included in the final bokeh plot.

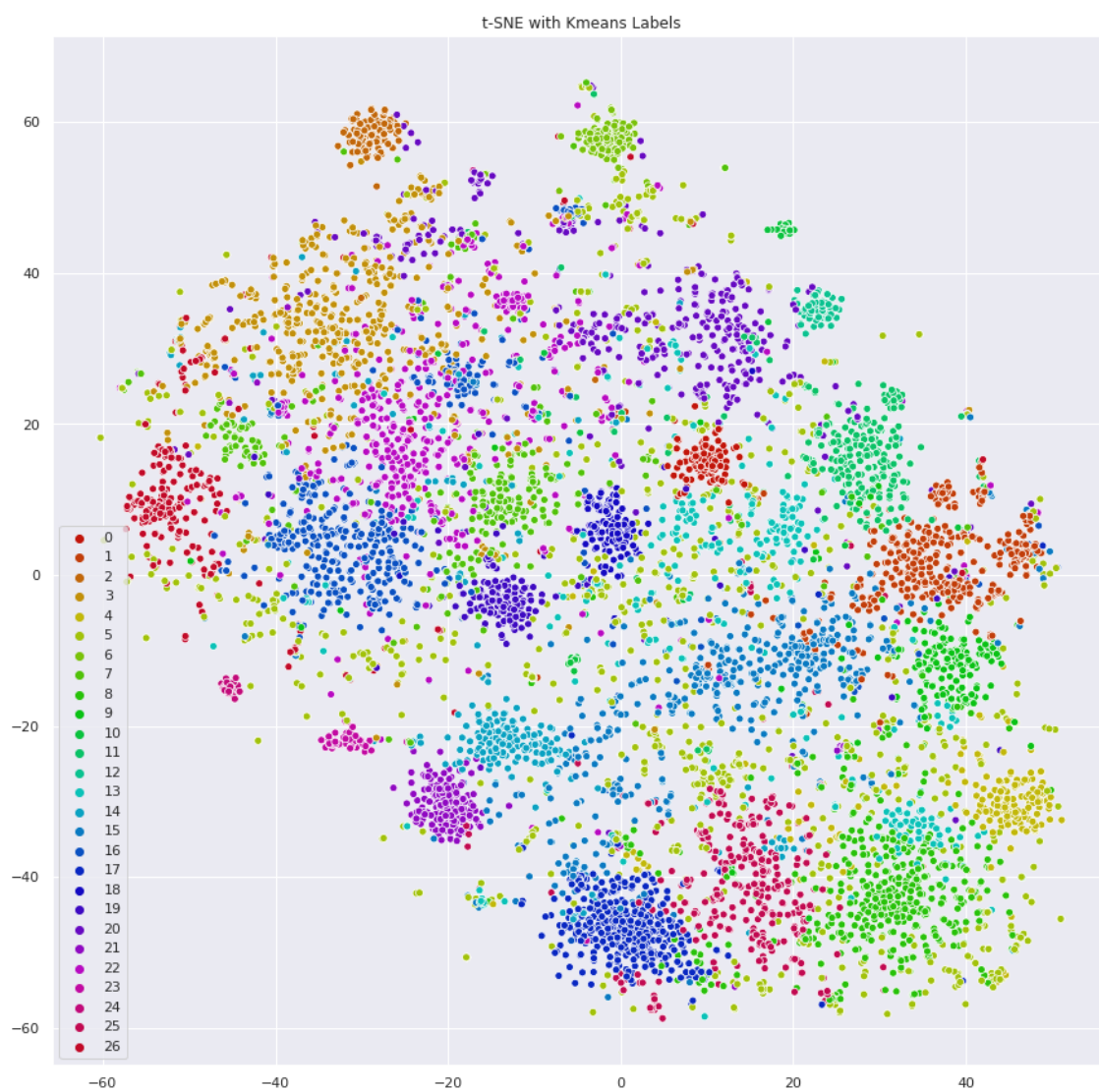


Figure 8.2: The plot shows the clustering result with the t-SNE positioning and the k-means labels.

## 8.2 Part2

In this task we were supposed to add the metadata information of the selected papers from *Part I* to the CORD-19 dataset and perform the clustering again. We wanted to find out if this approach improves and facilitates the search of papers for a specific topic field. In order to solve this task, we tried to generate a dataframe using the *pyPDF2* package to extract the metadata from the pdf files of the articles. Unfortunately, it did not work for all pdfs, because they did not contain uniform metadata fields. For this reason we decided to create a csv file and added all relevant metadata fields manually (Table 8.3). The manually created table was then added to the CORD-19 dataset (Figure 8.4).

Link	Title
<a href="https://www.ncbi.nlm.nih.gov/pubmed/32276116">https://www.ncbi.nlm.nih.gov/pubmed/32276116</a>	Rapid and visual detection of 2019 novel coronavirus (SARS-CoV-2) by a reverse transcription loop-mediated isothermal amplification assay
<a href="https://www.nature.com/articles/s41598-018-37483-w">https://www.nature.com/articles/s41598-018-37483-w</a>	A method to identify respiratory virus infections in clinical samples using next-generation sequencing
<a href="https://www.tandfonline.com/doi/full/10.1586/14737159.2014.888313">https://www.tandfonline.com/doi/full/10.1586/14737159.2014.888313</a>	Advances and challenges in biosensor-based diagnosis of infectious diseases
<a href="https://ann-clinmicrob.biomedcentral.com/articles/10.1186/1476-0711-7-18">https://ann-clinmicrob.biomedcentral.com/articles/10.1186/1476-0711-7-18</a>	Predicting the sensitivity and specificity of published real-time PCR assays
<a href="https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3522074/">https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3522074/</a>	Application of Molecular Diagnostic Techniques for Viral Testing

Table 8.1: Papers to diagnostics

A	B	C	D	E	F	G	H
paper_id	doi	abstract	body_text	authors	title	journal	abstract_summary
week_2_spreading_models	10.3390/ijerph16234683	infectious diseases are an important cause of human death. The study of the pathogenesis, spread, regularity, and development trend of infectious diseases not only provides a theoretical basis for future research on infectious diseases, but also has practical guiding significance for the prevention and control of their spread. In this paper, a controlled differential equation and an We developed a computational tool to assess the risks of novel coronavirus outbreaks outside of China. We estimate the dependence of the risk of a major outbreak in a country from imported cases on key parameters such as: (i) the evolution of the cumulative number of cases in mainland China based on the official data modeling; this paper studies the transmission process of the Corona Virus Disease 2019 (COVID-19). The error between the model and the official data curve is quite small. At the same time, it realized forward prediction and backward inference of the epidemic situation and the relevant analysis helps relevant countries to Testine for COVID-19 has been unable to keep up	infectious diseases are diseases that can be transmitted from person to person, from person to animal, or from animal to animal after proto-microorganisms and parasites infect human beings or animals [1–3]. Infectivity, epidemic, and uncertainty are the three main characteristics of infectious diseases. A thorough study of the spread. A cluster of pneumonia cases in Wuhan, China, was reported to the World Health Organization (WHO) on 31 December 2019. The cause of the pneumonia cases was identified as a novel betacoronavirus: the 2019 novel coronavirus (2019-nCoV, recently renamed as SARS-CoV-2). At the end of 2019, the new coronavirus (COVID-19) spread widely in China, and a large number of people became infected. At present, the domestic outbreak has been effectively controlled, while the new coronavirus is spreading rapidly in other areas. Currently, Europe has become the center of identifying who has the COVID-19 virus is	Bin Sheng Sun Gengxin Chen Chih-Cheng	Spread or Infectious Disease Modeling and Analysis of Different Factors on Risk	International Journal of Environmental Research and Public Health	something
week_2_risk	10.3390/jcm9020571	We developed a computational tool to assess the risks of novel coronavirus outbreaks outside of China. We estimate the dependence of the risk of a major outbreak in a country from imported cases on key parameters such as: (i) the evolution of the cumulative number of cases in mainland China based on the official data modeling; this paper studies the transmission process of the Corona Virus Disease 2019 (COVID-19). The error between the model and the official data curve is quite small. At the same time, it realized forward prediction and backward inference of the epidemic situation and the relevant analysis helps relevant countries to Testine for COVID-19 has been unable to keep up	A cluster of pneumonia cases in Wuhan, China, was reported to the World Health Organization (WHO) on 31 December 2019. The cause of the pneumonia cases was identified as a novel betacoronavirus: the 2019 novel coronavirus (2019-nCoV, recently renamed as SARS-CoV-2). At the end of 2019, the new coronavirus (COVID-19) spread widely in China, and a large number of people became infected. At present, the domestic outbreak has been effectively controlled, while the new coronavirus is spreading rapidly in other areas. Currently, Europe has become the center of identifying who has the COVID-19 virus is	Boldog Péter Tekeli Tamás Vizi Zsolt Dénés Lixiang	Assessment of Novel Coronavirus COVID-19 Outbreaks	Journal of Clinical Medicine	something
week2_forecasting	<a href="https://doi.org/10.1016/j.idm.2020.03.002">https://doi.org/10.1016/j.idm.2020.03.002</a>	studies the transmission process of the Corona Virus Disease 2019 (COVID-19). The error between the model and the official data curve is quite small. At the same time, it realized forward prediction and backward inference of the epidemic situation and the relevant analysis helps relevant countries to Testine for COVID-19 has been unable to keep up	(COVID-19) spread widely in China, and a large number of people became infected. At present, the domestic outbreak has been effectively controlled, while the new coronavirus is spreading rapidly in other areas. Currently, Europe has become the center of identifying who has the COVID-19 virus is	Yang Zihang Dang Zhongkai Meng Cui Hall	Propagation analysis and prediction of the COVID-19	KeAi	something

Figure 8.3: csv-table

We preprocessed the dataframe and applied the previously describes clustering pipeline. Finally, the cluster membership of each paper from *Part I* could be identified. The titles of papers from the cluster were saved as .csv-file respectively (see Figure 8.5).

It could be figured out that three papers (spreading models, databased time-series prediction and risk factor analysis) from *Part I* belong to the same cluster 6 (Figure 8.6). The paper related to diagnostics was assigned to cluster 15 (Figure 8.7) and the origin analysis article was a member of cluster 9 (Figure 8.8).

### 2.3 Appending metadata of papers from week2

Loading metadata of papers from week2 from csv file as dataframe, preprocessing (stripping whitespace, removing char, adding word counts of abstract, body text and unique words) and appending metadata to a general dataframe containing all the papers), dropping duplicates, data summary

```
In [39]: import pandas as pd
df_papers_week2 = pd.read_csv('C:/Users/Natalja/shared_folder/DSinLS20/week3/Week2_Papers.csv', sep=';', dtype={'paper_id' : str})
for column in df_papers_week2:
    df_papers_week2[column] = df_papers_week2[column].apply(lambda x: x.strip())
df_papers_week2
df_papers_week2['title'] = df_papers_week2['title'].str.replace('<br>', ' ')
df_papers_week2.drop_duplicates(subset='title', keep='first', inplace=True)
df_papers_week2['abstract_word_count'] = df_papers_week2['abstract'].apply(lambda x: len(x.strip().split())) # word count in abstract
df_papers_week2['body_word_count'] = df_papers_week2['body_text'].apply(lambda x: len(x.strip().split())) # word count in body text
df_papers_week2['body_unique_words'] = df_papers_week2['body_text'].apply(lambda x: len(set(str(x).split()))) # number of unique words in body text
df = df_papers_week2.append(df)
df.drop_duplicates(['abstract', 'body_text'], inplace=True)
df['abstract'].describe(include='all')
```

```
Out[39]: count    10005
         unique    7205
         top
         freq     2794
         Name: abstract, dtype: object
```

Figure 8.4: Appending metadata of week2 papers to a data frame containing metadata to CORD-19-research-challenge

### Identifying of cluster for each week2 paper

using helper function save\_cluster\_week2\_titles clusters of each paper from week2 identified and paper titles of this cluster will be saved in extra .csv-file

```
In [113]: def save_cluster_week2_titles(title, table):
cluster = table.loc[table['title'] == title, 'y'].values
#print(cluster)
cluster_value = cluster[0]
print('Cluster {}'.format(cluster_value))
cluster = table.loc[table['y'] == cluster_value, 'title']
nameId_week2 = table.loc[table['title'] == title, 'paper_id'].values
nameId_week2 = nameId_week2[0]
print(nameId_week2)
savePath = 'C:/Users/Natalja/shared_folder/DSinLS20/week3/df_covid_cluster_topic{}.csv'.format(nameId_week2)
cluster.to_csv(savePath, index = False)
for index, row in df_papers_week2.iterrows():
    print('Title of week 2 paper is {}'.format(row['title']))
    save_cluster_week2_titles(row['title'], df)
```

```
Title of week 2 paper is Spread of Infectious Disease Modeling and Analysis of Different Factors on Spread of Infectious Diseases Based on Cellular Automata
Cluster 6
week_2_spreading_models
Title of week 2 paper is Risk Assessment of Novel Coronavirus COVID-19 Outbreaks Outside China
Cluster 6
week_2_risk
Title of week 2 paper is Propagation analysis and prediction of the COVID-19
Cluster 6
week2_forecasting
Title of week 2 paper is Finding COVID-19 from Chest X-rays using Deep Learning on a Small Dataset
Cluster 15
week2_diagnostics
Title of week 2 paper is Identification of a new coronavirus
Cluster 9
week2_phylogenetic_analysis
```

Figure 8.5: Clustering and identifying cluster of week2papers



```

title
Spread of Infectious Disease Modeling and Analysis of Different Factors on Spread of Infectious Disease Based on Cellular Auto
Risk Assessment of Novel Coronavirus COVID-19 Outbreaks Outside China
Propagation analysis and prediction of the COVID-19
Anthropological Perspectives on the Health<br>Transition
How HIV patients construct liveable<br>identities in a shame based culture: the case of Singapore
Estimating the economic impact of pandemic<br>influenza: An application of the computable general<br>equilibrium model to the
" Travelling to scientific meetings is a<br>mission, not a vacation"
Perspectives of public health laboratories in<br>emerging infectious diseases
D(2)EA: Depict the Epidemic Picture of<br>COVID-19
Suicide news reporting accuracy and<br>stereotyping in Hong Kong
Pandemic Risk Modelling
Reflections on travel-associated infections<br>in Europe
Chapter 27 Disaster Mitigation
Chapter 3 Emerging Infectious Diseases and the<br>International Traveler
Learning from recent outbreaks to strengthen<br>risk communication capacity for the next influenza<br>pandemic in the Western
Impact of the topology of metapopulations on<br>the resurgence of epidemics rendered by a new<br>multiscale hybrid modeling a
" After Malaria Is Controlled, What's Next?"
A High-Resolution Human Contact Network for<br>Infectious Disease Transmission
Generality of the Final Size Formula for an<br>Epidemic of a Newly Invading Infectious Disease
Committed to Health: Key Factors to Improve<br>Users' Online Engagement through Facebook
Temporal patterns and geographic<br>heterogeneity of Zika virus (ZIKV) outbreaks in French<br>Polynesia and Central America
The challenges of implementing an integrated<br>One Health surveillance system in Australia
The legal determinants of health: harnessing<br>the power of law for global health and sustainable<br>development
Beyond the 'nanny state': Stewardship and<br>public health
Pandethics
International Organizations and Their<br>Approaches to Fostering Development
Using core competencies to build an evaluative<br>framework: outcome assessment of the University of Guelph<br>Master of Publ
China's distinctive engagement in global<br>health
A planetary vision for one health

```

Figure 8.6: Cluster 6: titles of related papers to the papers from week2 about risk factor analysis, spreading and forecasting

```

title
Finding COVID-19 from Chest X-rays using Deep Learning on a Small Dataset
COVID-19 and Dialysis Units: What Do We Know Now<br>and What Should We Do?
G6PD deficiency in COVID-19 pandemic: "a ghost<br>in the ghost"
COVID-19 pneumonia with hemoptysis: Acute<br>segmental pulmonary emboli associated with novel<br>coronavirus infection
" Maintenance Hemodialysis and Coronavirus<br>Disease 2019 (COVID-19): Saving Lives With Caution,<br>Care, and Courage"
Continuing education in oral cancer during<br>coronavirus disease 2019 (covid-19) outbreak
Inuit communities can beat COVID-19 and<br>tuberculosis
Tackling the COVID-19 Pandemic
Fellowship Training in Adult Cardiothoracic<br>Anesthesiology - navigating the new educational landscape due<br>to the corona
Pediatric Airway Management in Coronavirus<br>Disease 2019 Patients: Consensus Guidelines From the<br>Society for Pediatric A
" COVID-19, A Clinical Syndrome Manifesting as<br>Hypersensitivity Pneumonitis"
Editorial. Endonasal neurosurgery during the<br>COVID-19 pandemic: the Singapore perspective
Increased risk of ocular injury seen during<br>lockdown due to COVID-19
COVID-19 in pregnancy: early lessons
Clinical course and mortality risk of severe<br>COVID-19
Reply to "The use of traditional Chinese<br>medicines to treat SARS-CoV-2 may cause more harm than<br>good"
Knowledge and attitudes of medical staff in<br>Chinese psychiatric hospitals regarding COVID-19
The preventive strategies of GI physicians<br>during the COVID-19 pandemic
" Coronavirus disease (COVID-19) in a<br>paucisymptomatic patient: epidemiological and clinical<br>challenge in settings with
Perspectives from the Cancer and Aging<br>Research Group: Caring for the vulnerable older patient<br>with cancer and their ca
SARS-CoV-2 infection in a patient on chronic<br>hydroxychloroquine therapy: Implications for prophylaxis
COVID-19 Diagnostic and Management Protocol<br>for Pediatric Patients
" Spinal anaesthesia for patients with<br>coronavirus disease 2019 and possible transmission rates<br>in anaesthetists: retros
" Epidemiology, causes, clinical<br>manifestation and diagnosis, prevention and control of<br>coronavirus disease (COVID-19) d
Concerns for activated breathing control<br>(ABC) with breast cancer in the era of COVID-19:<br>Maximizing infection control
Heart Failure Editorial Emergencies in the<br>COVID-19 Era
Ayurveda and COVID-19: where<br>psychoneuroimmunology and the meaning response meet
Pulmonary Pathology of Early-Phase 2019 Novel<br>Coronavirus (COVID-19) Pneumonia in Two Patients With Lung<br>Cancer
WFUMB Position Statement: How to perform a safe<br>ultrasound examination and clean equipment in the context<br>of COVID-19

```

Figure 8.7: Cluster 15: titles of related papers to the paper from week 2 about diagnostics

```

title
Identification of a new coronavirus
High Resolution Analysis of Respiratory Syncytial Virus Infection In Vivo
Detection of Novel SARS-like and Other Coronaviruses in Bats from Kenya
Recombinant infectious bronchitis coronavirus H120 with the spike protein S1 gene of the nephropathogenic IBV strain
Nucleotide Sequence of the Inter-Structural Gene Region of Feline Infectious Peritonitis Virus
Molecular characterization of bovine noroviruses and neboviruses in Turkey: detection of recombinant strains
" Detection and characterisation of canine astrovirus, canine parvovirus and canine papillomavirus in puppies using next
Identification and Characterization of Severe Acute Respiratory Syndrome Coronavirus Subgenomic RNAs
Coevolution of activating and inhibitory receptors within mammalian carcinoembryonic antigen families
CHAPTER 1 Remarks on the Classification of Viruses
Canine kobuvirus infections in Korean dogs
Coronavirus Transcription: A Perspective
Identification and Analysis of Frameshift Sites
" Codon usage in Alphabaculovirus and Betabaculovirus hosted by the same insect species is weak, selection dominated and
Genic amplification of the entire coding region of the HEF RNA segment of influenza C virus
Comprehensive codon usage analysis of porcine deltacoronavirus
The First Detection of Equine Coronavirus in Adult Horses and Foals in Ireland
Sequences Promoting Recoding Are Singular Genomic Elements
Recombination and Coronavirus Defective Interfering RNAs
A recombinant infectious bronchitis virus from a chicken with a spike gene closely related to that of a turkey coronavirus
Single Stranded DNA Viruses Associated with Capybara Faeces Sampled in Brazil
Genetic diversification of penaeid shrimp infectious myonecrosis virus between Indonesia and Brazil
" Discovery of novel virus sequences in an isolated and threatened bat species, the New Zealand lesser short-tailed bat
" Spliced Leader RNAs, Mitochondrial Gene Frameshifts and Multi-Protein Phylogeny Expand Support for the Genus Perkinsus
" Genomic Organization, Biology, and Diagnosis of Taura Syndrome Virus and Yellowhead Virus of Penaeid Shrimp
" Polymorphisms and Tissue Expression of the Feline Leukocyte Antigen Class I Loci FLAI-E, -H and -K
WHO says coronavirus causes SARS
Conserved tertiary structure elements in the 5' untranslated region of human enteroviruses and rhinoviruses
Standards for Sequencing Viral Genomes in the Era of High-Throughput Sequencing

```

Figure 8.8: Cluster 9: titles of related papers to the paper from week 2 about origin analysis

As it can be seen the titles of the articles in the identified clusters are related to the papers from *Part I*. Therefore, it can be a helpful approach to find related papers.

### 8.3 Part3

#### 8.3.1 Loading in the Student Dataset

After adding only the information about the five papers that we have chosen in *Part I* to the CORD-19 dataset, we created a new dataset containing all submitted papers by the course participants. It turned out that opening each of the 60.000 json files in the CORD-19 dataset to filter those jsons that do not match one of the submitted articles is too time-consuming. Title names have been stripped since traveling whitespaces result in mismatches. The runtime was dramatically decreased by first joining the course dataset with the metadata. The included paper id ("sha") was then used to search in the file names of the jsons for the papers of interest. This leads to a runtime reduction from several hours to less than 3 minutes. We matched 177 of the 195 submitted articles. And after removing duplicated entries our new dataset based on the submissions of the course participants contains 146 papers.

#### 8.3.2 Preprocessing, Clustering and Results

Next, we proceeded with the preprocessing and final clustering step. For the preprocessing, we followed the previously described pipeline of the kaggle notebook closely. We removed common stopwords as they act as noise. Next, a vectorizer with a noise filter of  $2^{12}$  is applied, counting words and scoring less frequent words higher. Afterwards, the dimension of the dataset is reduced by PCA from over 4069 to 119. For the clustering we deviated from the kaggle notebook. First we tested a second method to determine the best k value. This was done by computing the silhouette score (Figure 8.9). Second we added another preprocessing step, to transform the data to a unit vector of 1, thus using the equivalent of cosine similarity as distance metric. After carefully comparing both methods with and without cosine similarity, one of elbow distortions and the other of silhouette scoring, we determined k by silhouette scoring and euclidean to be the best method (Figure 8.10). Thus we proceeded with k of 18. We choose 18, because it showed a noticeable bump in scoring, but is still a reasonable estimate based on the chosen paper data set by students. After running a t-SNE visualisation (Figure 8.11) and a Keyword extraction per cluster using LDA, with a minimum number constraint of 10 topics per cluster, we found 10 distinct clusters with an overarching topic. 8 clusters were removed due to insufficient number of data points. The results can be found in the Table 8.2.

The assignment of topics to each unique cluster was surprisingly easy, indicating a meaningful result. Some problems could be identified, for one some clusters had not enough articles assigned to them. Thus we could stipulate that the chosen k value of 18 is too high, rerunning the clustering at a lower value might give better results. On the contrary some clusters with high assignment could be identified (cluster 11). This could mean two things: Not much variation in the chosen cluster topics from the students or a more granular clustering is needed.



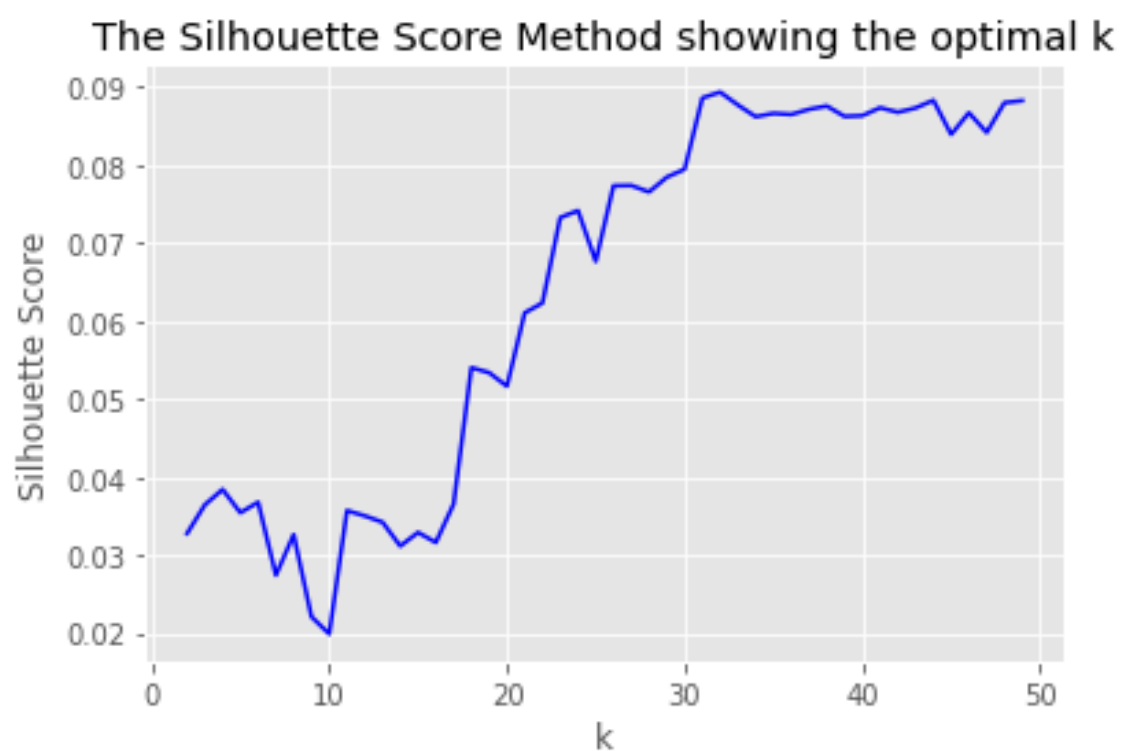


Figure 8.9: Silhouette scoring method for k-means. At the cluster point 17 a significant bump in scoring can be seen.

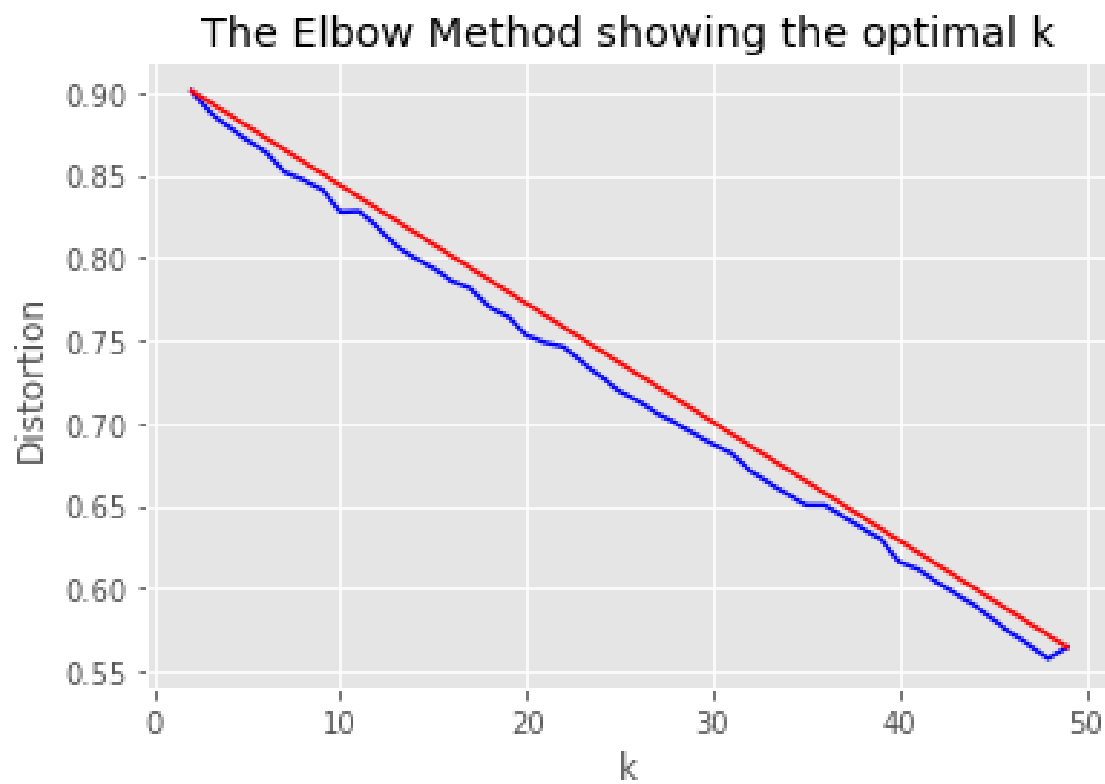


Figure 8.10: Distortion scoring method for k-means. Due to the low difference in scoring an almost linear line is produced. Thus in this case the method is unsuited to determine the optimal k for clustering



Figure 8.11: 2 dimensional tsne visualisation of the data set. Due to the low presence of articles the projection seems to be sparse. Still some clusters can be determined like: Top middle cluster 14: virus detection or middle left cluster 15: compartmentalized models. Big clusters like 11: modeling of spread are not clumped together.

Cluster	No. Articles	Assigned Topic	Keywords
0	11	phylogenetics	'sars', 'set', 'gene', 'rate', 'orf', 'recombination', 'datum', 'frequency', 'region', 'distance'
1	13	study of initial outbreak	'symptom', 'hospital', 'sars-cov-', 'china', 'country', 'mortality', 'use', 'respiratory', 'risk', 'evidence'
2	8	network-modelling of spread	'community', 'degree', 'threshold', 'mix', 'heterogeneity', 'group', 'node', 'use', 'transmission', 'size'
5	6	network-modelling of spread	'mix', 'wave', 'estimate', 'outbreak', 'total', 'community', 'human', 'delay', 'overall', 'additional'
9	11	disease forecasting	'case', 'disease', 'outbreak', 'estimate', 'influenza', 'process', 'interval', 'forecast', 'day', 'peak'
11	31	modelling of spread	'parameter', 'sequence', 'disease', 'method', 'change', 'city', 'spread', 'network', 'value', 'interaction'
12	18	origin detection	'case', 'camel', 'sars-cov-', 'human', 'protein', 'isolate', 'healthcare', 'bat', 'viral', 'sars-cov'
14	12	virus detection	'sars', 'sample', 'lung', 'serum', 'finding', 'detection', 'virus', 'care', 'study', 'swab'
15	8	compartmentalized modelling	'rate', 'risk', 'model', 'outbreak', 'death', 'datum', 'day', 'virus', 'state', 'patient'
16	6	diagnostics	'pcr', 'rsv', 'sequence', 'pneumonia', 'age', 'child', 'rhinovirus', 'associate', 'young', 'presence'

Table 8.2: Results of the clustering with unique topic assignment based on the first 10 keywords. 8 clusters are omitted due to low assignment of articles.

### 8.3.3 Creating a word cloud

Finally, we generated a basic word cloud. Here, we took the extracted keywords from the final clustering and used the word cloud package. The package provides a basic understanding of the word cloud with the use of some simple python libraries like numpy, pandas, matplotlib and pillow. The figure 8.12 shows the visualized wordcloud. Words like sars, virus and cov are bold and large which shows that the frequency of usage of these words is high and denotes their importance of it during the recent times. While these wordclouds are easy to look at, not much useful information can be extracted.

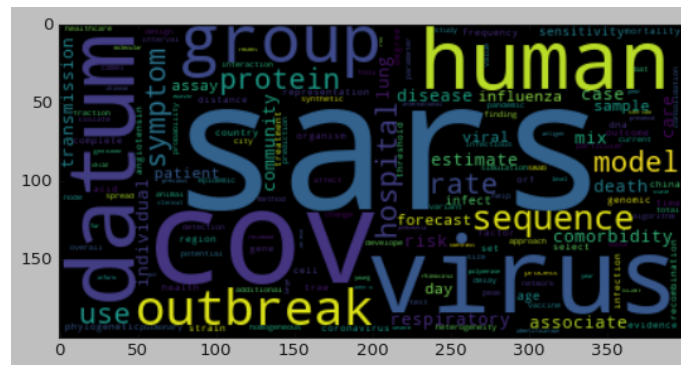


Figure 8.12: Word cloud of most frequent words.





## Bibliography

### Articles

- [1] Kristian G Andersen et al. “The proximal origin of SARS-CoV-2”. In: *Nature medicine* 26.4 (2020), pages 450–452 (cited on pages 25, 26).
- [3] Kuldeep Dhama et al. “SARS-CoV-2: Jumping the species barrier, lessons from SARS and MERS, its zoonotic spillover, transmission to humans, preventive and control measures and recent developments to counter this pandemic virus”. In: (2020) (cited on page 25).
- [4] Eneida L Hatcher et al. “Virus Variation Resource–improved response to emergent viral outbreaks”. In: *Nucleic acids research* 45.D1 (2017), pages D482–D490 (cited on page 25).
- [5] Tommy Tsan-Yuk Lam et al. “Identifying SARS-CoV-2 related coronaviruses in Malayan pangolins”. In: *Nature* (2020), pages 1–6 (cited on page 26).
- [6] Zhixin Liu et al. “Composition and divergence of coronavirus spike proteins and host ACE2 receptors predict potential intermediate hosts of SARS-CoV-2”. In: *Journal of medical virology* 92.6 (2020), pages 595–601 (cited on page 26).
- [7] Fábio Madeira et al. “The EMBL-EBI search and sequence analysis tools APIs in 2019”. In: *Nucleic acids research* 47.W1 (2019), W636–W641 (cited on page 25).
- [10] Fan Wu et al. “A new coronavirus associated with human respiratory disease in China”. In: *Nature* 579.7798 (2020), pages 265–269 (cited on page 26).

### Books

- [8] Michel Tibayrenc. *Genetics and evolution of infectious diseases*. Elsevier, 2017 (cited on page 25).







## Index

### B

Background . . . . . 9, 13, 15, 17, 25

### D

Data and Methods . . . . . 9, 13, 15, 17, 25

Discussion . . . . . 14, 15, 23, 26

Discussion & Results . . . . . 10

### R

Results . . . . . 14, 15, 17, 25