An application of clustering methods to cluster COVID-19 scholarly articles

Quynh Anh Nguyen[*][†]

Abstract

The main objectives of this project are clustering academic articles into different groups. The data set is CORD-19, resource of over 300,000 scholarly articles about COVID19 from the White House and a coalition of research groups. A subset of 1000 articles will be spitted from the original data to use in this project. Initially, each article from the data set will be parsed, cleaned and vectorized. Principle Component Analysis (PCA) will be subsequently applied on vectorized data to reduce the dimensions of the data. By then, K-Means and Gaussian Mixture Method are applied to cluster the projected data set. Last but not least, we find the representative keywords of each cluster and visualize them on the word-cloud plot. The result reveals that there are approximately 17 clusters by K-Means Method and 6 cluster by Gaussian Mixture Method. The clusters determined by K-Means surprisingly performed better than Gaussian Mixture Methods. Each cluster represents a distinctly COVID19-related concern such as International Law and Public Health Policy, Coronavirus and Schooling issue, Symptoms and Treatment and so on.

Keywords: GMM, K-Means, Model-based Method, NLP

1 Introduction

This project aims to cluster 1000 academic articles into different similarity groups and extract its main literature topics by applying NLP techniques, reducing dimensions of the data and comparing the results of two clustering methods, K-Means and Gaussian mixture model (GMM). Author wishes to contribute the fundamental statistical understanding to assist the medical community and researchers retrieve article references more convenient and efficient.

[*] Data Science and Economics Program, University of Milan. Email: quynh.nguyen@studenti.unimi.it.

[†] Github: jyanqa

2 Dataset

CORD-19 is a resource of over 350,000 scholarly articles about COVID-19, SARS-CoV-2, and related corona viruses. The dataset is pre-processed by dropping duplicates articles, removing data points whose missing value and deleting articles that were published before 2019. This is a huge data set , thus I am going to create a subset of 1000 articles after having preprocessed the original data set. Since the data set provides a lot of features of each article, I am going to extract the set of 4 features (paper id, title, abstract, body text) of each article. By using library *langdetect*, a distribution of language using in 1000 articles is explored. All articles that are not in English are dropped, the data set now remains 966 articles. After that, the data will be converted text to lower case, removed punctuation and removed stop-words such as few, we, thru, etc. At the end of this data prepossessing step, we vectorized the body text of each article. The data frame now is converted into a matrix of 996 rows and 4096 predictors.

3 Method

Principle Component Analysis The number of predictors in the data is much larger than the number of data-points. Therefore, PCA will be applied to our vectorized data to reduce the dimensions of the data. Dataset will be applied PCA in two ways. First, data will be reduced by finding the first PCs explain 95% of variance. Thus, we avoid missing too much information from the original data. On the other hand, the dataset will be reduced into 2-dimensional data. This helps the visualization of the dataset be more convenient and intuitive.

3.1 Clustering analysis

Clustering is an unsupervised method helps assign data objects into similar groups. Members of the same cluster must be similar and a member of one cluster must be as dissimilar as the members of others. In this project, we will examine the results after having applied two different Clustering methods. In detail, we will consider the number of clusters suggested by each method, the similarity of each cluster and the meaningful information extracted from clusters of each method.

Tuning the number of cluster As K is increasing, the within-cluster variation W always decreases; while the between-cluster variation B always increases. We will apply Elbow and Gap Statistic method to define the optimal number of clusters.

Elbow The sum of squared distances from each point to its centroid is computed and plotted against k. There will be a k value after which decreases in distortion are minimal. The desired number of cluster is pointed out where the plot shows an ‘elbow’.

Gap Statistics We compare the observed within-cluster variation $W(K)$ versus the within-cluster variation uniformly distributed data $W_{uniform}(K)$.

$$Gap(K) = \log W(K) - \log W_{uniform}(K) \tag{1}$$

K-Means Method If centroids are $m_1, m_2, ...m_k$, and partitions are $c_1, c_2, ...c_k$, then one can show that K-Means converges to a *local* minimum of

$$\sum_{k=1}^{K} \sum_{i \in c_k} ||x_i - m_k||^2 \qquad \text{Euclidean distance}$$

Fitting K-Means algorithm:

- Randomly determining the K centroids
- Iterate until the cluster assignments stop changing

  - For each of K clusters, compute the centroid, which is the p-length vector of the means in that cluster.

  - Assign each observation to the cluster whose centroid is closest using in Euclidean distance.

  This procedure is guaranteed to decrease the dissimilarity $W(K)$ at each step.

Gaussian Mixture Method This is one of the most popular model-based clustering approaches available. The key idea behind model-based clustering is that the data are considered as coming from a mixture of underlying probability distributions. The most popular approach is the Gaussian mixture model (GMM) (Banfield and Raftery, 1993) where each observation is assumed to be distributed as one of k multivariate-normal distributions, where k is the number of clusters, commonly referred to as components in model-based clustering.

Evaluating the performance K-Means or GMM clustered the articles but not extract its meaning. We will use LDA (Latent Dirichlet Allocation) to extract the most important terms of each cluster. In LDA, each document can be described by a distribution of topics and each topic can be described by a distribution of words. By checking the theme of extracted keywords, we can evaluate the good-fit result of clustering methods.

4 Results

4.1 Tuning the number of cluster

After reduced the dimension of our vectorized matrix by PCA method in which explain 95% of variance, we have a new data including 966 data-points and 708 predictors. Before

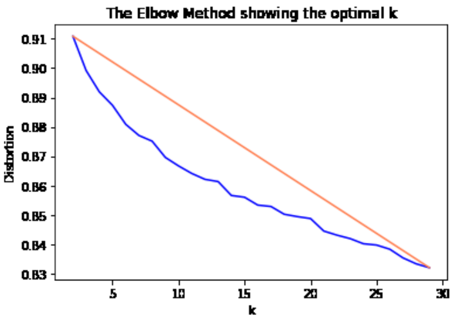


FIGURE 1: Elbow method.

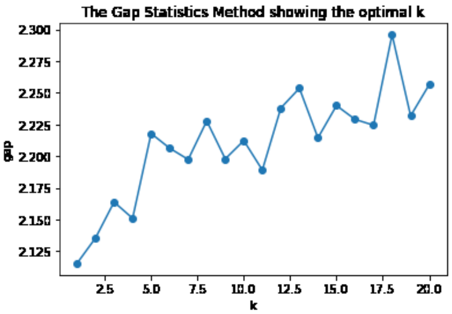


FIGURE 2: Gap Statistics method.

applying K-Means on this reduced data, we need to determine the optimal number of cluster.

According to statistical folklore, the best K is located at the 'elbow' of the clusters inertia, within-cluster sum-of-squares criterion, while K increases. However, the Figure 1 does not demonstrate apparently the 'elbow'. The optimal is determined in this case could be from 5 to 15 clusters. We can continue to find the optimal K by using Gap Statistics methods.

The Figure 2 shows the gap between the clusters inertia on the observed within-cluster variation $W(K)$ and the within-cluster variation of the reference data $W_{unif}(K)$. The optimal k is given by k for which the gap between the two results is maximum. According to Figure 2, the optimal k should be approximate 17 clusters.

4.2 K-Means

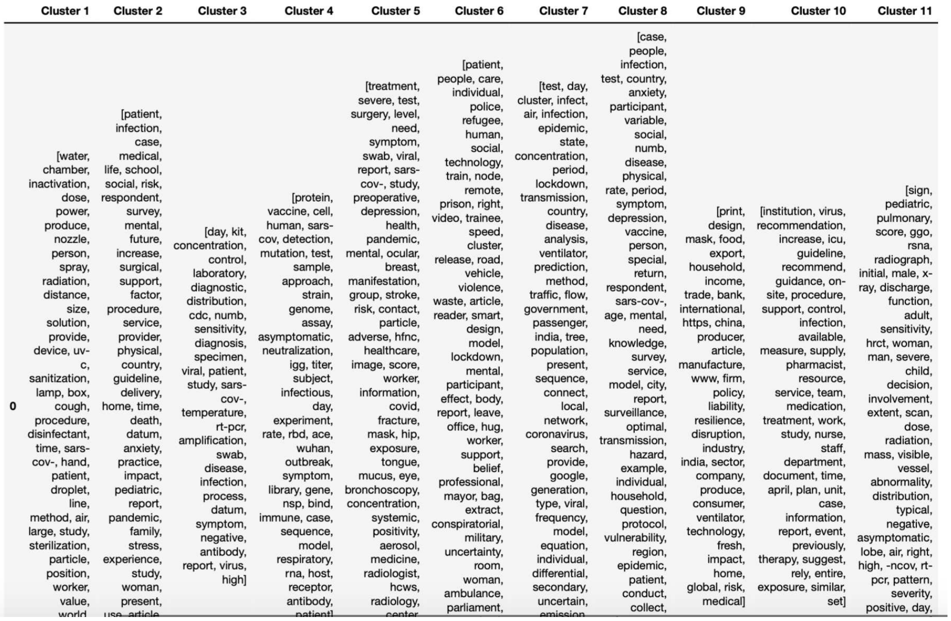
After determined the optimal number of cluster, we apply K-Means method.

In fact, a good cluster model is the model that could maximize the similarity within-cluster and minimize the similarity between-cluster. However, our data set is a 966×708 matrix, it is almost impossible to visualize to check whether the applied K-Means method well performed. Therefore, extracting keywords that represent for each cluster is an solution to examine the purity of these 17 clusters.

Figure 3 shows the keywords that are extracted from articles representing each cluster. The lexicon in the same group is pretty connected to each other and describes a certain topic. I have chosen some of 17 clusters and visualized them as word-cloud, figure 4. It is feasible to point out the Covid19-related concerns from these 6 clusters. They are treatments, business, guideline, authority policies, symptoms and education.

4.3 Gaussian Mixture Model Method

In order to visualize the data point in a scatter plot, I reduced the original dataset, 966×4096 matrix, into a 966×2 matrix by PCA method. By using Elbow and Gap Statistic method as the previous steps, the optimal k is pointed out at 10 clusters. Figure 5 shows the scatter



	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10	Cluster 11
0	[water, chamber, inactivation, dose, power, produce, nozzle, person, spray, radiation, distance, size, solution, provide, device, uv-c, sanitization, lamp, box, cough, procedure, disinfectant, time, sars-cov-, hand, patient, droplet, line, method, air, large, study, sterilization, particle, position, worker, value, world,	[patient, infection, case, medical, life, school, social, risk, respondent, survey, mental, future, increase, surgical, support, factor, procedure, service, provider, physical, country, guideline, delivery, home, time, death, datum, anxiety, practice, impact, pediatric, report, pandemic, family, stress, experience, study, woman, present, use, article,	[day, kit, concentration, control, laboratory, diagnostic, distribution, cdc, numb, sensitivity, diagnosis, specimen, viral, patient, study, sars-cov-, temperature, rt-pcr, amplification, swab, disease, infection, process, datum, symptom, negative, antibody, report, virus, high]	[protein, vaccine, cell, human, sars-cov, detection, mutation, test, sample, approach, strain, genome, assay, asymptomatic, neutralization, igg, titer, subject, infectious, day, experiment, rate, rbd, ace, wuhan, outbreak, symptom, library, gene, nsp, bind, immune, case, sequence, model, respiratory, rna, host, receptor, antibody, patient]	[treatment, severe, test, surgery, level, need, symptom, swab, viral, report, sars-cov-, study, preoperative, depression, health, pandemic, mental, ocular, breast, manifestation, group, stroke, risk, contact, particle, adverse, hfnc, healthcare, image, score, worker, information, covid, fracture, mask, hip, exposure, tongue, mucus, eye, bronchoscopy, concentration, systemic, positivity, aerosol, medicine, radiologist, hcws, radiology, center	[patient, people, care, individual, police, refugee, human, social, technology, train, node, remote, prison, right, video, trainee, speed, cluster, release, road, vehicle, violence, waste, article, reader, smart, design, model, lockdown, mental, participant, effect, body, report, leave, office, hug, worker, support, belief, professional, mayor, bag, extract, conspiratorial, military, uncertainty, room, woman, ambulance, parliament,	[test, day, cluster, infect, air, infection, epidemic, state, concentration, period, lockdown, transmission, country, disease, analysis, ventilator, prediction, method, traffic, flow, government, passenger, india, tree, population, present, sequence, connect, local, network, coronavirus, search, provide, google, generation, type, viral, frequency, model, equation, individual, differential, secondary, uncertain, emission	[case, people, infection, test, country, anxiety, participant, variable, social, numb, disease, physical, rate, period, symptom, depression, vaccine, person, special, return, respondent, sars-cov-, age, mental, need, knowledge, survey, service, model, city, report, surveillance, optimal, transmission, hazard, example, individual, household, question, protocol, vulnerability, region, epidemic, patient, conduct, collect,	[print, design, mask, food, export, household, income, trade, bank, international, https, china, producer, article, manufacture, www, firm, policy, liability, resilience, disruption, industry, india, sector, company, produce, consumer, ventilator, technology, fresh, impact, home, global, risk, medical]	[institution, virus, recommendation, increase, icu, guideline, recommend, guidance, on-site, procedure, support, control, infection, available, measure, supply, pharmacist, resource, service, team, medication, treatment, work, study, nurse, staff, department, document, time, april, plan, unit, case, information, report, event, previously, therapy, suggest, rely, entire, exposure, similar, set]	[sign, pediatric, pulmonary, score, ggo, rsna, radiograph, initial, male, x-ray, discharge, function, adult, sensitivity, hrct, woman, man, severe, child, decision, involvement, extent, scan, dose, radiation, mass, visible, vessel, abnormality, distribution, typical, negative, asymptomatic, lobe, air, right, high, -ncov, rt-pcr, pattern, severity, positive, day,

FIGURE 3: Keywords of 11 clusters out of 17 clusters.

plot of reduced dataset applied K-Means method and the centroid of each cluster. K-Means

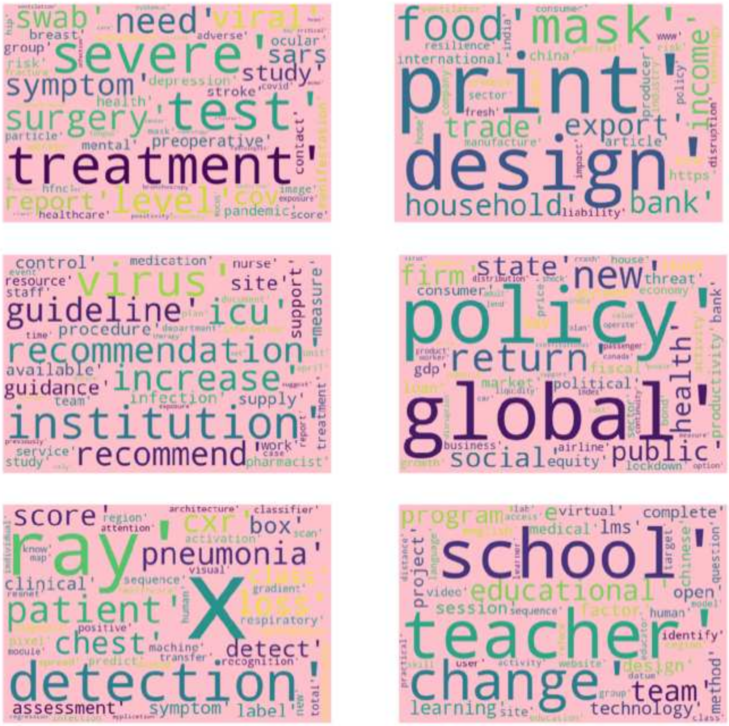


FIGURE 4: WordCloud of 6 clusters by K-Means method.

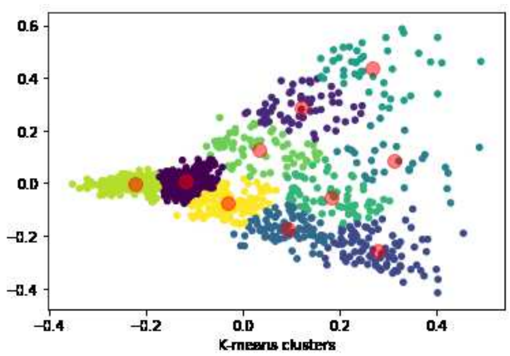


FIGURE 5: K-Means Method.

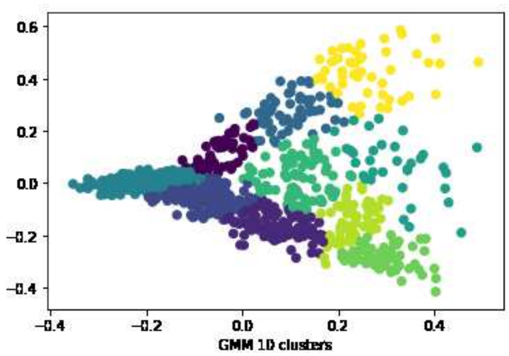


FIGURE 6: Gaussian Mixture Method.

finds suitable clustering result for well-separated data. For example, if we have simple binary large object of data, the K-Means algorithm can easily label those observations. As we can see in the figure 5, there appears to be a slightly overlap between clusters, such that we might not have complete confidence in the cluster assignment of points between them.

K-Means method lacks of flexibility in cluster shape and lack of probabilistic cluster assignment. This weakness is improved in Gaussian mixture model, one of the most popular model-based method. Figure 6 shows the scatter plot of the reduced 2-dimensional data set by choosing $k = 10$ as the having applied on K-Means method in figure 5. The 10 clusters in the figure 6 are more robust than 10 clusters in the previous step. However, there is still the overlap parts between those clusters. In order to determine the optimal number of components for our model, we can consider AIC and BIC indicators.

Figure 7 reveals AIC and BIC index over the number of cluster. The optimal number of clusters is the value that minimizes the AIC or BIC. The AIC tells us that our test of 10 components above was not the best: either 7 or 8 components would have been a better option. Meanwhile, the BIC recommends 5 or 6 components for the model. Thus, we choose 6 components as a compromise. By assigning the optimal number of components to GMM model, the plot in Figure 8 shows us 6 clusters. These cluster are robust with respect to noise but there are less intersection compared to Figure 6, the plot when we applied GMM to 10 clusters.

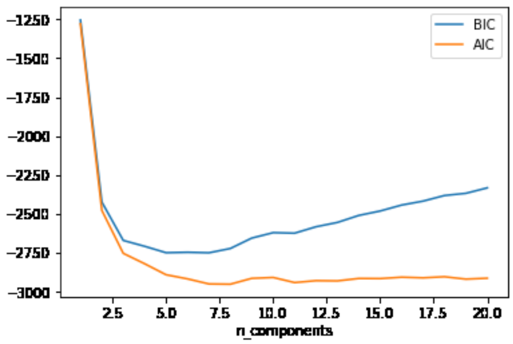


FIGURE 7: AIC and BIC index.

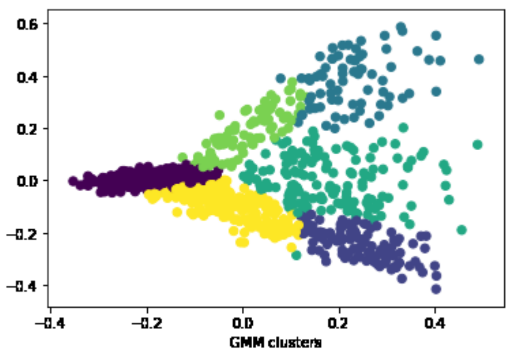
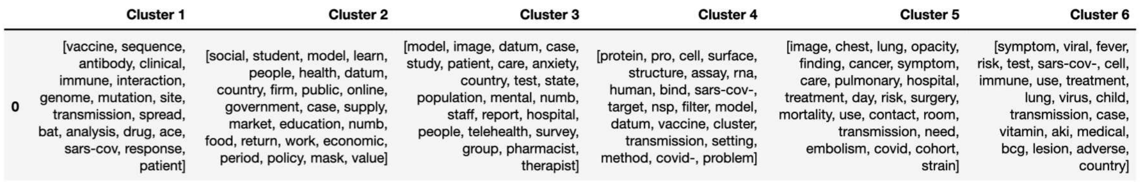


FIGURE 8: n_component is 6.



	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
0	[vaccine, sequence, antibody, clinical, immune, interaction, genome, mutation, site, transmission, spread, bat, analysis, drug, ace, sars-cov, response, patient]	[social, student, model, learn, people, health, datum, country, firm, public, online, government, case, supply, market, education, numb, food, return, work, economic, period, policy, mask, value]	[model, image, datum, case, study, patient, care, anxiety, country, test, state, population, mental, numb, staff, report, hospital, people, telehealth, survey, group, pharmacist, therapist]	[protein, pro, cell, surface, structure, assay, rna, human, bind, sars-cov-, target, nsp, filter, model, datum, vaccine, cluster, transmission, setting, method, covid-, problem]	[image, chest, lung, opacity, finding, cancer, symptom, care, pulmonary, hospital, treatment, day, risk, surgery, mortality, use, contact, room, transmission, need, embolism, covid, cohort, strain]	[symptom, viral, fever, risk, test, sars-cov-, cell, immune, use, treatment, lung, virus, child, transmission, case, vitamin, aki, medical, bcg, lesion, adverse, country]

FIGURE 9: Keywords of 6 clusters by GMM method.

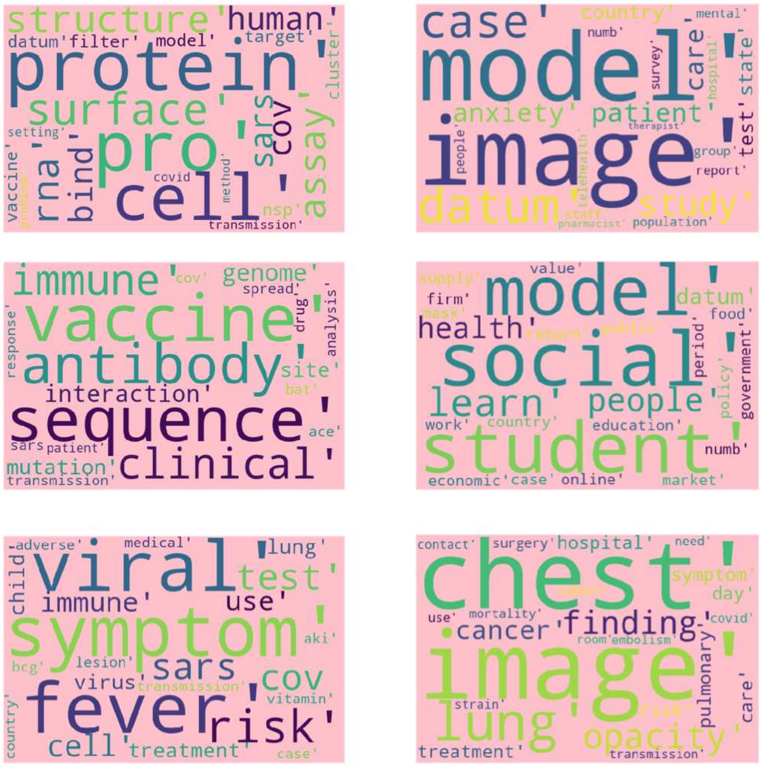


FIGURE 10: WordCloud of 6 clusters by GMM method

According to the scatter plots of 2-dimensional projected data in Figure 5 and Figure 8, the GMM prevails. But we will check whether this model-based method performed truly well. Based on GMM model, we assigned the label for each article, the label is one of 6 clusters. By then, the group of lexicon representing each group is drawn out. Figure 9 demonstrates the keywords retrieved from 6 clusters. Those keywords are plotted as word-cloud plot in Figure 10. It is undeniable that the keywords from GMM model are not as good as from K-Means model in term of describing a theme, which is our main objective of this project.

5 Conclusion

- *The optimal number of clusters of K-Means method are 10; whereas it is 6 for GMM model.* We can estimate the optimal k of K-Means method by Elbow method or Gap Statistics. Meanwhile, AIC and BIC are two indicators used to estimates the optimal components of the model-based method GMM.
- *The GMM model shows the better result on two-dimensional data compared to K-Means method.* On this data set, after converted the original data to two-dimensional data, the clusters determined by GMM are more separated, less overlap compared to K-Means.
- *In reality, the extracted keywords proved that K-Means is the better method.* Keywords of each clusters by K-Means model are describing a concrete content. Meanwhile, the keywords representing 6 clusters of GMM model are pretty confusing. This is one of the disadvantages of the Normal mixture models."The likelihood could be unbounded for the general normal mixture model and spurious solutions might emerge. Because, several components form a single mode and observations are not assigned to the closest cluster due to different component-specific co-variance matrices."
- *The different results of two methods can be also explained by the dimension of the dataset, by the method determine the optimal cluster or components as well.* Applying PCA methods helps reducing dimensional of the data but we might lost a lot of information.

References

Manzi, G. (2020), Lecture 20: Course Recap, lecture notes. *Advanced Multivariate Statistics B7416*, University of Milan, delivered December 2020.

Boehmke, B. & Greenwell, B. (2020). Chapter 22 Model-based Clustering. *Hands-On Machine Learning with R*, https://bradleyboehmke.github.io/HOML/model-clustering.html.

VanderPlas, J. (2016). In Depth: Gaussian Mixture Models. *Python Data Science Handbook*, https://jakevdp.github.io/PythonDataScienceHandbook/05.12-gaussian-mixtures.html.

MaksimEkin (2020). Loading data. *COVID-19 Literature Clustering*, https://www.kaggle.com/maksimeren/covid-19-literature-clustering/notebook.

Palafox, L. (2019). A visual introduction to the Gap Statistics. *The Glowing Python*, https://glowingpython.blogspot.com/2019/01/a-visual-introduction-to-gap-statistics.html.

Data Source

The White House (2020). Data set. *CORD-19*, https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge.