



[skillenza-com/MishMash-India-2020] "Class"Apart - SARS CoV-2 / COVID-19 Prediction Using Corona Virus Genome Sequence - Theme - Social Impact (#114)

1 message

Anubhav Mishra <notifications@github.com>

Tue, Mar 31, 2020 at 11:27 PM

Reply-To: skillenza-com/MishMash-India-2020

<reply+AL43CXNIRFBEPQMTXINVLPV4R5SYVEVBNHHCGPUYIM@reply.github.com>

To: skillenza-com/MishMash-India-2020 <MishMash-India-2020@noreply.github.com>

Cc: Neeraj Joshi <neerajjoshi2308@gmail.com>, Mention <mention@noreply.github.com>

1. Theme - Social Impact
2. Project Name - SARS CoV-2 / COVID-19 Prediction Using Corona Virus Genome Sequence
3. Our project SARS CoV-2 / COVID-19 Prediction Using Corona Virus Genome Sequence , as the same says is used to predict COVID-19 in suspected patients using their genome sequence and comparing it to the sequence of a healthy , disease free person. For this we have made a machine learning model using various algorithms.
4. Team Name - "Class"Apart
5. Team Members :-
 - a) Anubhav Mishra Github - @AnubhavMishra22
 - b) Neeraj Joshi Github - @Neerajjoshi2308
 - c) Shruti Tyagi Github - @ShrutiTyagi27
5. Repository Link - <https://github.com/AnubhavMishra22/SARS-CoV-2-COVID-19-Prediction-using-Corona-Virus-Genome-Sequence>
6. Presentation Link - https://docs.google.com/presentation/d/1tmI5Ok6LvlQUUzcFEsYVVChJ6fZMI-BajB3vXtKRWVj0/edit#slide=id.g82b31db86e_0_206
7. Azure Services Used & that can be used in future Real Life applications - Microsoft Genomics , Machine Learning studio (Classic) workspace , VS code Extension , Logic apps, Azure Virtual Machine Azure backup, Software As A service (SaaS) , Machine Learning studio (Classic) web services , Microsoft Kubernetes, Azure API for FHIR & Jupyter Notebook Python SDK.
8. Pitch For Our Project-

Our project named "SARS-CoV-2-COVID-19-Prediction-using-Corona-Virus-Genome Sequence". DNA sequencing has become a key technology in many areas of biology and other sciences such as medicine, forensics, and anthropology. The process of determining the order of nucleotides adenine (A), thymine (T), cytosine (C), and guanine (G) along a DNA strand. We need to know the order of nucleotide bases in a strand of DNA for sequencing. All the information required for the growth and development of an organism is encoded in the DNA of its genome. So, DNA sequencing is fundamental to genome analysis and understanding the biological processes in general. DNA sequencing may be used to determine the sequence of individual genes, larger genetic regions (i.e. clusters of genes or operons), full chromosomes, or entire genomes of any organism. DNA sequencing is also the most efficient way to indirectly sequence RNA or proteins (via their open reading frames). So, we've used this application of DNA Sequencing in predicting the COVID by matching the DNA sequence via Machine Learning algorithms (Naïve Bayes) and natural language processing tools (Vectorization). We have collected a dataset of human DNA sequence, which is used for the purpose of analysing and matching the DNA sequence with COVID. As mentioned earlier, Naïve Bayes algorithm is used for predicting the possibility of COVID in a human by matching it by our model. The reason for using Naive Bayes classifiers is that the accuracy obtained in various classifier algorithm were least of Naïve Bayes, where every pair of features being classified is independent of each other. The "human dataset" which consist of DNA Sequence of 4808 humans with a class label of either infected by COVID or not. The dataset is divided into two parts, namely, feature matrix and the target vector. The feature vector (rows) of dataset in which each vector consists of the value of dependent features. In above dataset, features are 'Sequence' and target as 'Class' i.e. whether COVID infected or not. Before analysing with machine learning algorithm, we used the natural processing tool CountVectorizer, which Creates the Bag of Words of sequences, where each bag consists of k-word of the whole sequence in continuous fashion. Since a bag of sequence has created, now it can be fitted into the Naïve Bayes model and we get the result matrix of model as Accuracy = 0.973, Precision = 0.973, Recall = 0.973, f1 = 0.972. Now, we have used another dataset which contains the data of various countries regarding the cases of COVID. After applying various pre-processing techniques to clean-up, the data. Thus, various plots are created with respect to time showing the growth of virus in major countries. And

we've used the Linear Regression model to show the trajectory of growing of the virus. The various Azure Services that we have used & that can be used in future Real Life applications are Microsoft Genomics , Machine Learning studio (Classic) workspace , VS code Extension , Logic apps, Azure Virtual Machine Azure backup, Software As A service (SaaS) , Machine Learning studio (Classic) web services , Microsoft Kubernetes, Azure API for FHIR & Jupyter Notebook Python SDK. We have used them to run our code , deploy our code, we have made workspace in our azure account as Corona_Virus_Prediction and resource named as Kaggle. In the midst of this pandemic projects like this could be a life saver if used in association with various labs by collecting their mRNA genome samples , the can be used to predict COVID-19 in suspected patients using their genome sequence and comparing it to the sequence of a healthy, disease free person. For this we have made a machine learning model using various algorithms. And with the help of a proper business model this model could turn into a profitable business too for our startup in future too. At last I would like to quote "If you are alive you can do anything but if life isn't their nothing is their." Life of each human is the most precious thing on the entire Earth and using the correct form of technology such as ours it can be saved.

9. Raw Dataset Link-<https://www.kaggle.com/paultimothymooney/coronavirus-genome-sequence>
trained and tested data is present in the zip file.

You are receiving this because you were mentioned.
Reply to this email directly, [view it on GitHub](#), or [unsubscribe](#).