#########################################################################

DATA WRANGLING STEPS FOR ANOMALY DETECTION

#########################################################################

An unknown donor came to rescue when I was searching for the Anomaly Prediction dataset. I downloaded this dataset from the following link.

<https://drive.google.com/open?id=0B0u6UJI0ZnXXdHdac3pVcTZzZFU>

The dataset comprises of 28 files. Each file represents the reading from a sensor.

I discussed with my mentor about how to perform wrangling on the given dataset. All the input files were in the zip format. Therefore, the first logical thing I wanted to do was to unzip all those files and have a sense of how the data really looks like.

For unzipping all these files, I used the Linux Operating System. I pushed all the zip files to my Linux server. Once I copied all the files, I started unzipping those one by one.

I used the following command to perform the unzipping operation.

|  |
| --- |
| unzip AnomalyDetectionFullDataSet.zip |

I got all the text files after unzipping those files. I started going through all those files to have a feel of how those files look like.

To my luck, all the files had the same format. Therefore, I did not face any data quality or data inconsistency issues.

Before beginning with the data pre-processing steps, I did rename all the files for my convenience.

The following tables shows the old and new names of the text files.

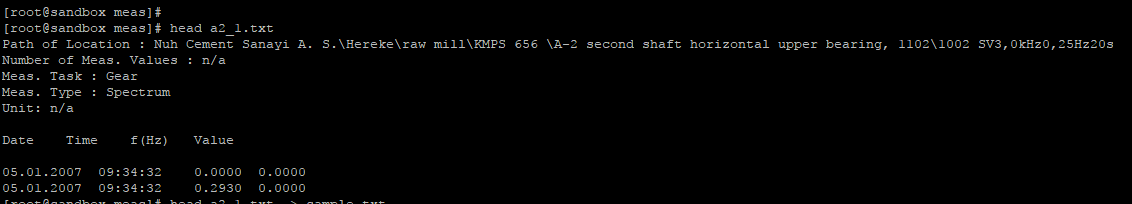
|  |  |
| --- | --- |
| **TEXT FILE OLD NAME** | **TEXT FILE NEW NAME** |
| A-1\_\_1002 SV3,0kHz0,25Hz20s.txt | a1\_1.txt |
| A-1\_\_1012 EA1,5kHz0,5-12kHz.txt | a1\_2.txt |
| A-1\_\_1020 SA12,0kHz1Hz10s.txt | a1\_3.txt |
| A-2\_\_1002 SV3,0kHz0,25Hz20s.txt | a2\_1.txt |
| A-2\_\_1012 EA1,5kHz0,5-12kHz.txt | a2\_2.txt |
| A-2\_\_1020 SA12,0kHz1Hz10s.txt | a2\_3.txt |
| A-3\_\_1002 SV3,0kHz0,25Hz20s.txt | a3\_1.txt |
| A-3\_\_1008 SV0,4kHz0,03Hz180s.txt | a3\_2.txt |
| A-3\_\_1012 EA1,5kHz0,5-12kHz.txt | a3\_3.txt |
| A-3\_\_1014 EA0,4kHz0,15-6kHz0,5Hz.txt | a3\_4.txt |
| A-3\_\_1020 SA12,0kHz1Hz10s.txt | a3\_5.txt |
| A-4\_\_1002 SV3,0kHz0,25Hz20s.txt | a4\_1.txt |
| A-4\_\_1008 SV0,4kHz0,03Hz180s.txt | a4\_2.txt |
| A-4\_\_1012 EA1,5kHz0,5-12kHz.txt | a4\_3.txt |
| A-4\_\_1014 EA0,4kHz0,15-6kHz0,5Hz.txt | a4\_4.txt |
| A-4\_\_1020 SA12,0kHz1Hz10s.txt | a4\_5.txt |
| A-5\_\_1002 SV3,0kHz0,25Hz20s.txt | a5\_1.txt |
| A-5\_\_1008 SV0,4kHz0,03Hz180s.txt | a5\_2.txt |
| A-5\_\_1012 EA1,5kHz0,5-12kHz.txt | a5\_3.txt |
| A-5\_\_1014 EA0,4kHz0,15-6kHz0,5Hz.txt | a5\_4.txt |
| A-5\_\_1020 SA12,0kHz1Hz10s.txt | a5\_5.txt |
| A-6\_\_1002 SV3,0kHz0,25Hz20s.txt | a6\_1.txt |
| A-6\_\_1012 EA1,5kHz0,5-12kHz.txt | a6\_2.txt |
| A-6\_\_1020 SA12,0kHz1Hz10s.txt | a6\_3.txt |
| A-7\_\_1002 SV3,0kHz0,25Hz20s.txt | a7\_1.txt |
| A-7\_\_1012 EA1,5kHz0,5-12kHz.txt | a7\_2.txt |
| A-7\_\_1020 SA12,0kHz1Hz10s.txt | a7\_3.txt |
| M-1\_\_1019 AC Spectrum Torque.txt | m1\_1.txt |

For performing the above rename operation, I am using the following commands.

|  |
| --- |
| mv A-1\_\_1002\ SV3\,0kHz0\,25Hz20s.txt a1\_1.txt  mv A-1\_\_1012\ EA1\,5kHz0\,5-12kHz.txt a1\_2.txt  mv A-1\_\_1020\ SA12\,0kHz1Hz10s.txt a1\_3.txt  mv A-2\_\_1002\ SV3\,0kHz0\,25Hz20s.txt a2\_1.txt  mv A-2\_\_1012\ EA1\,5kHz0\,5-12kHz.txt a2\_2.txt  mv A-2\_\_1020\ SA12\,0kHz1Hz10s.txt a2\_3.txt  mv A-3\_\_1002\ SV3\,0kHz0\,25Hz20s.txt a3\_1.txt  mv A-3\_\_1008\ SV0\,4kHz0\,03Hz180s.txt a3\_2.txt  mv A-3\_\_1012\ EA1\,5kHz0\,5-12kHz.txt a3\_3.txt  mv A-3\_\_1014\ EA0\,4kHz0\,15-6kHz0\,5Hz.txt a3\_4.txt  mv A-3\_\_1020\ SA12\,0kHz1Hz10s.txt a3\_5.txt  mv A-4\_\_1002\ SV3\,0kHz0\,25Hz20s.txt a4\_1.txt  mv A-4\_\_1008\ SV0\,4kHz0\,03Hz180s.txt a4\_2.txt  mv A-4\_\_1012\ EA1\,5kHz0\,5-12kHz.txt a4\_3.txt  mv A-4\_\_1014\ EA0\,4kHz0\,15-6kHz0\,5Hz.txt a4\_4.txt  mv A-4\_\_1020\ SA12\,0kHz1Hz10s.txt a4\_5.txt  mv A-5\_\_1002\ SV3\,0kHz0\,25Hz20s.txt a5\_1.txt  mv A-5\_\_1008\ SV0\,4kHz0\,03Hz180s.txt a5\_2.txt  mv A-5\_\_1012\ EA1\,5kHz0\,5-12kHz.txt a5\_3.txt  mv A-5\_\_1014\ EA0\,4kHz0\,15-6kHz0\,5Hz.txt a5\_4.txt  mv A-5\_\_1020\ SA12\,0kHz1Hz10s.txt a5\_5.txt  mv A-6\_\_1002\ SV3\,0kHz0\,25Hz20s.txt a6\_1.txt  mv A-6\_\_1012\ EA1\,5kHz0\,5-12kHz.txt a6\_2.txt  mv A-6\_\_1020\ SA12\,0kHz1Hz10s.txt a6\_3.txt  mv A-7\_\_1002\ SV3\,0kHz0\,25Hz20s.txt a7\_1.txt  mv A-7\_\_1012\ EA1\,5kHz0\,5-12kHz.txt a7\_2.txt  mv A-7\_\_1020\ SA12\,0kHz1Hz10s.txt a7\_3.txt  mv M-1\_\_1019\ AC\ Spectrum\ Torque.txt m1\_1.txt |

Each file began with some housekeeping information such as the file name, file size, the data format. This housekeeping information spanned for eight lines, after which the actual data was visible. For the analytics purpose, I took the decision of removing the first eight lines from each file and then proceed ahead with the dataset.

For your reference, the following snippet shows the first eight lines in each file.



I used the following commands to remove the first eight lines from each of the files mentioned above.

|  |  |  |  |
| --- | --- | --- | --- |
| sed -i 1,8d a1\_1.txt  sed -i 1,8d a1\_2.txt  sed -i 1,8d a1\_3.txt  sed -i 1,8d a2\_1.txt  sed -i 1,8d a2\_2.txt  sed -i 1,8d a2\_3.txt  sed -i 1,8d a3\_1.txt | sed -i 1,8d a3\_2.txt  sed -i 1,8d a3\_3.txt  sed -i 1,8d a3\_4.txt  sed -i 1,8d a3\_5.txt  sed -i 1,8d a4\_1.txt  sed -i 1,8d a4\_2.txt  sed -i 1,8d a4\_3.txt | sed -i 1,8d a4\_4.txt  sed -i 1,8d a4\_5.txt  sed -i 1,8d a5\_1.txt  sed -i 1,8d a5\_2.txt  sed -i 1,8d a5\_3.txt  sed -i 1,8d a5\_4.txt  sed -i 1,8d a5\_5.txt | sed -i 1,8d a6\_1.txt  sed -i 1,8d a6\_2.txt  sed -i 1,8d a6\_3.txt  sed -i 1,8d a7\_1.txt  sed -i 1,8d a7\_2.txt  sed -i 1,8d a7\_3.txt  sed -i 1,8d m1\_1.txt |

Removing those eight lines made all the files look identical to each other.

The next step was to look for any missing data.

The strategy for looking at the missing data for this dataset is a little bit different.

I found out that there are no missing values for the existing dates for all the 28 files, but there are no entries for some of the dates, which are reasoned to be related to the machine failure or network issue.

Given the above information, I investigated the reasoning behind the missing dates and explored it in the following ways.

1. **WEEKEND**: I tried to find out whether the machine fails to receive the reading because of the weekend. But, it was not because of this factor, because all the 28 sensors were functioning on the weekends and the readings mentioned in the input files are proof of that.
2. **PATTERN OF DATE MISSING**: I wanted to find out whether there exists a pattern of the missing dates. Upon going through all the 28 data files, I found out that there is no pattern of the missing dates. The missing date pattern is not uniform across all the data files, therefore we cannot comment anything on it.

The next thing I want to do now is to decide the strategy for the following things.

1. Which dates to include in the dataset to build the model for anomaly detection?
2. Which model should be used for detecting the anomaly?
3. How the data should be divided into training, validation, and testing datasets?
4. How to measure the accuracy of the model generated?

The above mentioned things will be covered in the Milestone Report as those are not part of the data wrangling steps.

#########################################################################

END OF DOCUMENT

#########################################################################