Business understanding

The Tartu University Autonomous Driving Lab seeks to get a dataset that consist of cleaned input data obtained from the self-driving Lexus on Rally Estonia tracks in order to use it as input for machine learning models. The data itself is located in ROS bag file, thus meaning it needs to be extracted. The ROS bag file also contains data that is irrelevant to the project, which implies that only relevant data must be extracted.

Our team intends to produce a dataset, that would be cleaned of any unwanted data, so that the machine learning models that would be trained using this data, would produce the best results.

The project would be deemed successful if as a result our team will get a dataset, which would be cleaned of most of the unwanted data, as some of the data removal logic might be too advance for our team to implement

The resources available to us are: human resources, which consist of 2 second-year bachelors students, an ROS bag file that contains all of the data needed to complete this project. As for the software to be used: python+conda environment that will be used for implementing the data cleaning rules and Robostack, which will allow us to install ROS in conda environment.

The project must be completed by 16.12.2021. In order to deem the project successful, it must be cleaned of the parts of the recording where the car does not move, parts of recording where the turn signal is on, parts of recording where the car is reversing. The data must also be separated from straight sections of the road and curved regions of the road, as well as the data parts where the road has intersections. If all of these steps will be completed before the said completion date, we will attempt to clean the dataset of the data where there is another car on the road, going in opposite direction.

As for the possible delays in completion of the project, these may be caused by: insufficient knowledge in Robostack, which may be solved by thoroughly examining the documentation, as well as possible problems in implementing the logic, which may be solved by asking for advice the projects contact person and by searching for and fixing the errors that have broken the logic.

Terminology:

- Data cleaning - the process of preparing data for analysis by removing or modifying data that is incorrect, irrelevant
- ROS bag file - bags are the primary mechanism in ROS for data logging, which means that they have a variety of offline uses.

The benefits that our team will get from this project are that we will gain experience in working with big data and get to test our knowledge in quite complex project. There are no costs associated with this project.

The data mining goal is to produce a cleaned dataset and a poster for presenting our project.

As for the data-mining success criteria, we suppose that the cleaned dataset will be sent to the projects contact person and it will be their decision whether we have succeeded or not.

Data Understanding

Our project is to clean the input data of a self-driving car on the roads of Estonia so that the resulting dataset could be used in machine learning models. The purpose of this report is to give a clear indication of what type of data we are dealing with in this project. The report is an important step in a project so we can set specific boundaries on what we can achieve with the data we were initially given. Main goal of this project is to eliminate any unnecessary data when the car is either not driving, turning, reversing and so on. Hence we need to have enough data where we can for example specifically find out the velocity of the car, steering wheel or car wheel angle. Due to the data being in ROS bag format, the data needs to be extracted before we can make sense of it. We can be certain that there is also lots of data that we are not going to need for example RPM showings should be redundant for us because it we can determine whether the car is moving more accurately with the speedometer showings and it is not needed for detecting objects from camera view.

The data was gathered in Elva during a Rally Estonia weekend via all of the sensors on a self-driving Lexus. Acquiring additional data from the same test drive is not an option for us.

The data consists of numerical values associated with the drive including velocity, RPM, etc. Also camera footage, from what we are supposed to detect passing cars and also removing the frames where the distracting cars are visible. Same applies for humans and intersections although detecting cars seems to be a highly advanced task and might be difficult to execute.

Quality of the data is hard to determine as of yet due to the data being delayed a bit so we have not yet gotten an in depth look at it yet. We expect to cut out most of the footage from the start and end of the clips where the test drives were still being calibrated and the car was in a parked state.

Overall we expect to have good quality data and not many mishaps in the data as it was collected automatically through machines so human error and missing values are not expected in this data set, which makes cleaning it a lot easier and might even be unnecessary. In conclusion we expect to learn a lot from this project and handling this sort of a massive data set should be only beneficial and provide us with good experience moving onward.

Project plan

The project plan consists of 5 primary steps and 2 optional step. The first step is to extract the data needed from the ROS bag file. As the bag file contains unwanted data, we will need to decide which data is relevant for our project. This should take about 1-2 hours per person. The second step is to remove the parts of the recording where the car does not move. For this step it is necessary to determine which attributes determine whether the car was stationary or moving. As for the time estimate, this may take 2-3 hours per person. The third step is to remove parts of recording where the turn signal is on, this may take up the same amount of time as the previous step. The fourth step is to remove parts of recording where the car is reversing, estimated time for completion is the same as for the previous 2 steps. The fifth steep is to separate the data from straight sections of the road and curved regions of the road. This step appears to be more complex, so the estimated time to complete this step would be around 5 hours per person. The last 2 steps are to separate from the data parts where the road has intersections and to remove parts of the data where there is another car on the road. These step appear to be the most complex in this project, so this implies that the time for their completion would be 7-10 hours per student for each task.

Repo link: https://github.com/kaniser1/IDS_Project