THE ML LOG FILE PROJECT

What I did my last few weeks

CONTENTS

- To build a model capable of looking into unseen log data sets and to pick out interesting or unusual parts in the logs automatically

STAGES OF THE PROJECT

- Data Discovery – Framing the problem statement, Looking into data sources
- Data Preparation – Examining the Data, Converting raw data to usable data, data conditioning
- Model Planning – Model Selection, Feature engineering
- Model Building – Hardcoding the model in python
- Inference – Using the model practically

DATA DISCOVERY

- Framing the Problem Statement in the first slide
- Looking into Data Sources – The log data was created by each of the nodes logging each step, and this data was provided by Steve

DATA PREPARATION

- Examining the Data
- Converting the text log data to the usable format of Pandas DataFrame
- Encoding the textual data into a numeric format as the model cannot work with text data

MODEL PLANNING

- Model Selection – determining based on the problem statement which kind of model would be most suitable
- Had first considered data clustering, then a regular neural network. Then finally settled on a Recurrent Neural Network
- This stage involved a lot of research

RECURRENT NEURAL NETWORK

This is a special type of Neural Network which has the ability to retain information from past inputs

In this way an RNN can decide differently for an input based upon different prior inputs

They possess an internal memory which retains prior inputs

These are used for speech recognition, grammar learning, time series prediction, anomaly detection

LONG SHORT TERM MEMORY (LSTM)

- LSTMs are a special variation of Recurrent Neural Networks
- They have the ability to retain prior input information over long time steps (1000s and more)
- LSTMs contain information outside the normal flow of the recurrent network in a gated cell
- Information can be stored in, written to, or read from a cell, much like data in a computer's memory
- The cell makes decisions about what to store, and when to allow reads, writes and erases, via gates that open and close.

USE OF RNN IN THE TASK

- The log data has a particular sequence of specific logs which take part in the normal regular functioning of the node
- Thus, an RNN model can learn this sequence of logs from an examined set of logs free of problems
- This RNN after trained can be used to make predictions on unseen log data and if there is any discrepancy between the predicted log and the actual log, this can be evaluated to detect the interesting parts of the log data

MODEL BUILDING

- The model was built using TensorFlow, a library in python
- An LSTM was created and used

[Show the code]

INFERENCE

- The model once trained can then be used to predict on unseen log data
- And so the according to the predictions made, the deviations between it and the actual log data can be recorded as a metric and measured
- Once this metric passes a certain threshold, it can be flagged up to draw attention to that particular part of the log

CHALLENGES FACED

The main challenge was about selecting the right model to do the job. This part took the most research and study

Another challenge was about converting the raw log file data to a numeric format understandable by the model

Implementation of the model for the task efficiently was another challenge

THANK YOU