

To, IITD-AIA Foundation of Smart Manufacturing

Subject: Weekly Progress Report for Week 3

Dear sir, Following is the required progress report to the best of my knowledge considering relevant topics to be covered.

What's happening this week:

- EDA-Pre-Processing the data i.e., scrapped out from site.
- Tokenization and stemming of messages/questions.

My understanding of INTP23-ML-01: CHATBOT for FSM

Scope: In the Chatbot for Site, I will try to scrape out all important info from website and according to the questions of users, it will provide best possible and accurate answers and it should follow the upcoming check points analyze user queries and understand user messages, provide answers to the messages of the user accurately, provide all information about the updated activities and save their time from searching on website and google search the information, provide answers accurately as much as possible, i.e., instead of providing irrelevant information, its' better to admit that bot is not able to understand the question.

Solution: In this, concept of deep learning either Pytorch or TensorFlow will be used along with some other libraries like Beautiful Soup, and one for scrapping the data on the website or using some Rule based approach.

Approach: Rule-based approach will be used for most common questions like hi, hello, bye. And web-scraping the data in order to give website related answers. For, tokenization NLTK or spaCy will be used and for main process of mapping most-accurate answers TensorFlow will be used.

## **WEEKLY PROGRESS:**

As promised last week, I completed with web-scraping of IAFSM website.

**Now Day-wise.**

### **June 26(Monday)**

I first reduced the list of links i.e. I used set () and now most of the links are non-repetitive and then processed them for web-scraping. After that, I stored the parsed information in json format. Most of things are done but some contents are repeating in the doc; that is my first priority work. Also, I started working on research paper, and done with Abstract and some other contents.

### **June 27(Tuesday)**

I first found pattern in which data is stored in “p” tags or directly written in “div” tag. Whenever content is directly written in “div” tag their class name is “sppb-addon-content”. So, I searched for this class named div and printed the file in json. Here, yesterdays’ error of repeating content is removed. But now error is, value is stored in different curly brackets. So, in order to remove error, I first used “read and write” method on file. However, it didn’t work as most of the data is removed using this method. Now, I have to change the code, where I have to write in json only once so in order to that I used dictionaries and list. Under I dictionary I used four lists, in which values like link, paragraph, heading, div content are appended and in the last, this dictionary is stored in json.

### **June 28(Wednesday)**

I learnt about architecture of transformer, in order to apply transformer in my chatbot. But I found out just transformer will not be able to chat as transformer is just capable of text generation, semantic analysis, text suggestion etc. So, in order to make chatbot talk I need to use supervised data, to guide chatbot how to respond.

### **June 29(Thursday)**

I learnt about large language models like BERT, GPT 2.0, llama also the usage of pipeline in the implementation of pre trained models and also learnt architecture of transformer, in order to apply transformer in my chatbot. But I found out just transformer will not be able to chat as transformer is just capable of text generation, semantic analysis, text suggestion etc. So, in order to make chatbot talk I need to use supervised data, to guide chatbot how to respond.

### **June 30(Friday)**

I learnt about large language models like BERT, GPT 2.0, llama also the usage of pipeline in the implementation of pre trained models and also learnt architecture of transformer, in order to apply transformer in my chatbot. But I found out just transformer will not be able to chat as transformer is just capable of text generation, semantic analysis, text suggestion etc. So, in order to make chatbot talk I need to use supervised data, to guide chatbot how to respond. Also, tried for updating my json file format I order to update the context of paragraph or answer as heading with respect to paragraph. Also, I discussed the same thing with my mentor, where I got to know that we have to generate the context some what manually, that would work in our case.

**July 1 (Saturday)**

I learnt about attention mechanism, while watching tutorial I learnt that I missed various topics of Deep Learning like activation function, loss function, encoder-decoder, etc. So, I decided to start these lagging points from Tomorrow. Also, I learnt the series that initially RNN was developed but it can have Only one context, the new used LSTM -RNN, after getting problem in this we encoder, decoder and now, in term of number of contexts we use transformer which uses attention mechanism. This, series of development is very important to understand in order to work on transformer ad to decide on what factors we used transformers.

**July 2(Sunday)**

I started with learning Deep Learning concepts where I learnt about Neural Networks, activation Function, Loss Function, Cost Function, Forward, Back Propagation, etc. Also, I started generating context about IAFSM to start with the project model implication on the dataset.