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-scrapping of FSM website.

Now Day-wise.

June 19 (Monday):

I successfully logged in into the website and get the web scraped data of dashboard page.

I had to extract out the loginToken value and pass into payload(four fields namely username, password, logintoken, remeberusrname). Now this desired-value(login-token), and other fields are sent and "post" method in order to login into the site. Now, the problem is to keep logged-in until process takes place so I used request.session for that.

June 20 (Tuesday):

I solved the problem mentioned in 19th June diary, first is to extract out all the "a" tags even after utf-8 and print it,

I found all the "a" tags are get them using for loop.

2nd is I extracted out all the "href" by using get () function and also searching only "http" links and scraped them also while scraping dashboard's link we get the information of each page and that pages' link are also extracted.

So, now I scrapped out whole "FSM" site

June 21(Wednesday):

Considering 20th June diary, aim was to convert /store information into json, for that I went through various videos and try the same but not able to do today I was getting errors. I decided to discuss with my mentor if we could convert to some other format and do the NLP work over that, if I was not able to do json format also, I had to discuss what information needs to be stored (just getting some idea).

June 23 (Friday):

Considering 21th diary, aim was to convert /store information into json, for that I go through various videos and try the same but not able to do today, I was getting errors. I have discussed with my mentor, and got some idea how to do the work. Or if I can not convert to json then I will look for some format on which other language models/chatbot had done some work. And, I will be trying for json work right now.

June 24 (Saturday):

I solved json problem i.e., the information from scraped web page data has been solved and stored in json file, also I found out what were the total tags in the given page and now I was to select which tags I have to convert to json or have to convert all tags to json format.

supervised data.

June 25 (Sunday):

Today, on discussion and checking of code with mentor, I found out instead of FSM skill website we have to scrape IAFSM website. In IAFSM website most of the link are starting from “/index/php” and some of them were http, so I first, found out all the links and search for pattern “/index/php” and concatenated with “https://iafsm.in/", and stored in list, then I searched for http links and again appended in the same list. On running the code, it’s become visible that most of the links were repeating, so I used set in order to store distinct links only. Now I have to store the information in json.