To, IITD-AIA Foundation of Smart Manufacturing

Subject: Weekly Progress Report for Week 2

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:

- FSM Website scrapping for collecting info regarding the site through login username and password.
- 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 did learn concepts like tokenization, stemming along with their implementation on small random data.

Also, I got introduced to new and important factors of NLP i.e., RNN, LSTM-RNN, word2vec, transformer, Attention mechanism and keras library.

Along with that I got bit idea of Scrapy. And I tried for web-scrapping using Beautiful Soup there I am trying to retrieve the data through login username and password. This was the overall report.

Now Day-wise.

June 13 (Tuesday):

I tried for Data storing i.e., scraping out the information from FSM website where I got confused which Framework should be used (Beautiful Soup or Scrapy)

Also, which heading/keyword will be used for scrapping out or identifying the data for storage.

That day, I was mix-up trying to search the things but not able to do so. And, my mentor provided me direction to get more towards NLP through feedback on this day report (June13).

June 14 (Wednesday):

So, after the feedback from my mentor, I learnt NLTK module, where I studied various new things tokenization, stemming, Regular expressions use of ‘re” library and how to extract data using regular expression

Also learnt the lemmatization and difference between lemmatization and stemming.

After that meeting got arranged with mentor and there, I got to know some new concepts to focus on like RNN, attention mechanism, transformer for predicting/ generating answers and EDA (exploratory data analysis).

June 15 (Thursday):

And, after attending June 14 doubt session I got to know various new concepts of NLP, which can be applied in deep learning. So, in order to learn new topics like word2vec, attention mechanism, first we should have initial pre-processing knowledge. So, I applied tokenization, stemming, lemmatization on random data and also bag of words now knowing the shortcoming of tokenization, next day I planned to learn word2vec.

Also, I watched the beautiful soup and applied it side by side. So, in this, website is been scrapped out and if I was searching using tags, I was getting right answers. But if we were searching for string or pattern then nothing is been returned or printed. Also, we were getting basic parsed file, I think user information is not there in the parsed data.

June 16 (Friday):

I learnt NLP in more depth concepts like word embeddings, word2vec and its two types continuous bag of words and skip-gram. Also learnt the architecture of them and moving further I learnt RNN, in RNN I learnt it's architecture and types of RNN (one to one, many to one, one to many, many to many) and found out "many to many" will be used in chatbot. In many to many we used to take multiple inputs and provide multiple outputs. Also learnt forward propagation, backward propagation, the loss function and introduced with the "Vanishing Gradient problem" in which derivative of sigmoid value decreases and become very small that weight updating becomes constant in backward propagation. So, to solve this we use LSTM RNN i.e., long short- term memory recurrent neural network.

June 17 (Saturday):

I learnt NLP in more depth concepts which includes LSTM (long short-term memory) RNN and its four-parts: memory cell, forget, input, output gated cell.

I learnt the architecture of LSTM and also its parts like memory cell where information can be added or removed. Then forget cell which comes in handy when context of sentence is important if context of previous and current information is same value given is near to 1 else 0.

And, did a small implementation of tokenization, stemming, word2vec in spam detection on supervised data.

June 18 (Sunday):

I did some LSTM implementation on small random data, where I developed our own model for training not using any API like "Google News" and learnt pre-padding and keras. Also, back I'm trying for pagination through Beautiful Soup and parse the data.