Project Selection Task:

Diverse, Explainable Multi-hop Question Answering

Deadline for the Task: 11th May 2021 11:59PM

In this task, you are supposed to build a simple text classification model

You would be working with the SST dataset.Train,Val, embedding files have been provided to make your work easier.

We will experiment with 2 different models , mainly CNN and RNN models to check which performs better on this dataset. The code for vanilla models is given feel free to make changes to the model layers/parameters and experiment. All your experiments will get logged automatically into a log.txt file(you are expected to submit this file also in the end)

dataloader.py : loads the SST datasets , please fill in the missing lines (do this FIRST)
model.py : contains the models and training loops , places where you have to add code has been cleared marked in comments, fill as much as you can , even if you are not able to get it running its fine, but do answer the theory questions at the end of this file.
(Do not change the directory structure)

You will have to install pytorch, refer to https://pytorch.org/ for installation.

Reference links are also provided at the corresponding locations.
The places where you need to fill in code have “YOUR CODE HERE” written in the scripts apart from that if you think you can improve something please feel free to.

Link to code:
https://drive.google.com/drive/u/0/folders/1pMTLGbevEc59lpayRVcs2BjPM-yPPGH_

Once again, we reiterate, there are **no prerequisites** for this project. All the learning happens after you get selected into the project team.
Note: We are **NOT** going to evaluate you based on the accuracy or performance of the model. Hence, it's completely fine if you are unable to train the model (due to compute reasons, or whatever it be). We are specifically looking to gauge your understanding and willingness to learn/explore.

Apart from the above classification task, we expect you to explore and **answer the following questions.** Once again there is no right or wrong answer to these questions and it's basically to understand your thought process.

1. Which model performed better on val set, why do you think it did so (Note: if you couldn't get the code running it's alright make a conjecture xD) Even state if you added any extra layers to the vanilla model and how that helped.
2. Given a particular sentence or document , how will you find similar sentences/documents from a large text corpus(Note: there are multiple correct answers to this,think of something **creative and simple**)
3. The progress in NLP , exponentially grew after the idea of "Attention" was introduced , if you haven't heard about this term have a look at these blog1 or blog2.
   Why do you think attention was so helpful ??
4. What are the drawbacks in Glove or Word2vec embeddings

In case of any doubts feel free to contact any of the Group Admins.

HAPPY SOLVING!!!