SEASONS OF CODE 2024
Sentiment Analysis and Text Generation using Many-to-One LSTMs

Mentors: Shreyas Katdare & Dion Reji

Date: June 14, 2024

Week 3 - Resources

Hurray! We are finally starting the real project now! Part 1 of the project is on Sentiment Analysis and we will complete this through 2 assignments - one this week and the other next week.

1. In order to analyse sentiments, we will work with Movie Reviews - i.e, at the end of the first part, you would have developed a bot which can classify Movie reviews (as positive or negative), and infact give a rating out of 10. For training our model we will be using the IMDB dataset, which can be found inside `Week 3` folder of the repo.

2. Word embeddings like GloVe are essential in sentiment analysis because they capture semantic relationships between words by representing them as high-dimensional vectors. This allows the model to understand context and nuances in language, improving its ability to accurately detect sentiment. Embeddings help in transforming textual data into a format that machine learning models can effectively process, leading to better performance in identifying positive, negative, or neutral sentiments. To learn more about Word Embedding, see this YouTube video. We have already provided you with GloVe embeddings. (See `Week 3` folder)

3. Regular expressions (regexes) are used in sentiment analysis to efficiently preprocess text by identifying and manipulating patterns within the data. They can be employed to clean and normalize text by removing unwanted characters, extracting specific phrases, or identifying sentiment-indicative words and emoticons. This preprocessing step helps in standardizing the input, reducing noise, and improving the accuracy and performance of sentiment analysis models. You can learn more about regexes here

Assignments

Complete the following tasks this week.

1. (Not to be submitted.) Refer to the resources above and familiarize yourself with:
   - Word embeddings, specifically Glove emebedding
   - Why we use regexes?

2. (To be submitted.) This week we shall be pre-processing the data. This involves preparing textual data for analysis by cleaning and transforming it. This typically includes tasks such as removing stop words, punctuation, and special characters, converting text to lowercase, stemming or lemmatizing words, and tokenizing the text into meaningful units. These steps help to standardize the text, reduce noise,

and enhance the performance of sentiment analysis models by ensuring that they focus on the most relevant and informative parts of the data.

In order to complete the task of data preproccessing, complete the Jupyter notebook file named `Assignment_3.ipynb`, which can be found inside `Week 3` folder. Detailed instructions to be completed can be found inside the file

Submission. You need to submit the Jupyter Notebook file after completing the assignmnet by 19.06.2024 (Next Wednesday) EOD.