Hi,

I'm Chip Huyen -- a master's student at Stanford University where I'm teaching the course "CS 20SI: TensorFlow for Deep Learning Research".

As the third assignment for my class, I made my students build a chatbot. In the process of research for a suitable dataset, I realized that the only publicly available datasets that we can use are from movie dialogues, Twitter conversations, or Reddit comments. These aren't realistic conversations. Since chatbots can only speak as well as the datasets we train them on, to build chatbots that can talk realistically, we need a realistic dataset. For more information about the chatbot assignment as well as the demo of the chatbot that I build, please see the assignment handout.

I hope to build a realistic conversation dataset by collecting real chatlog donated by people like you. Anyone can donate any of their chatlog. I will make the dataset publicly available after I've anonymized the data and done some basic data pre-processing.

To donate your chatlog, please follow these 3 quick steps:

Step 1
Copy - paste any chatlog you'd like to share into a .txt file. Each chatlog should be in a separate file.

Step 2
Make relevant changes to the names of the participants to anonymize your chatlog. If possible, I'd ask you to kindly specify the gender of the speakers. For example, you can replace your name with "male 1" (if you identify with males) and your friend's name to "female 1" (if your friend identifies with females).

Step 3
Head over to this Google Script form to submit your chatlog files. You can enter your real email if you want me to get in touch with you at some point in the future (e.g. to share the publicly available dataset with you), or you can enter None if you'd like to stay anonymous. In the "Primary language" field, please enter the primary language used in the chatlog.

If you have any question or recommendation, please send to huyenn@stanford.edu. Thank you very much!