Designing a Chatbot using Cornell Movie-Dialog Corpus Dataset

Team 3

- Cesar Lucero Ortiz
- Akram Mahmoud
- Jeremy Cryer
- October 23,2023

Team Contributions

Cesar Lucero

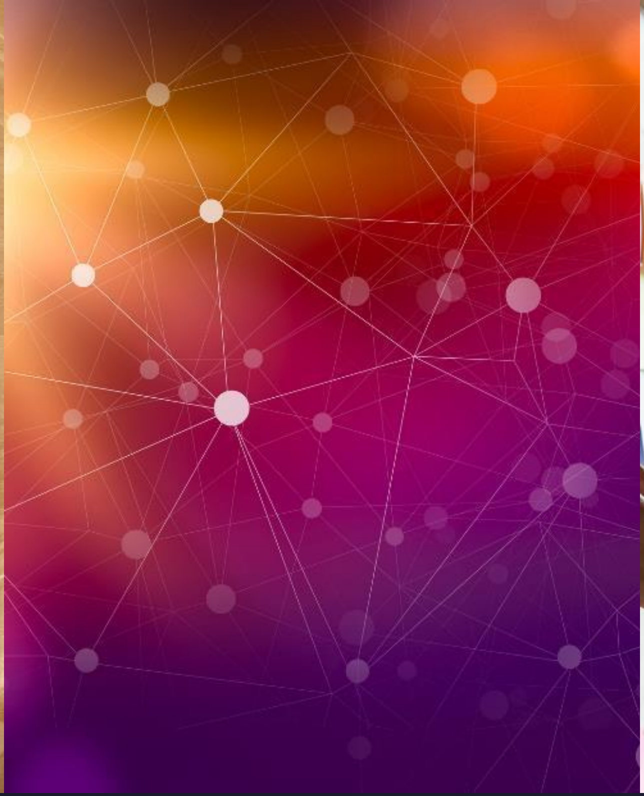
- Bulk of software was generated by Cesar.
- Including text preprocessing, model training, and model testing.
- Report development/support

Akram Mahmoud

- Developing presentation
- Consolidating report information to present.

Jeremy Cryer

- Coding and brainstorming support.
- Critiquing and code clean up.
- Report development



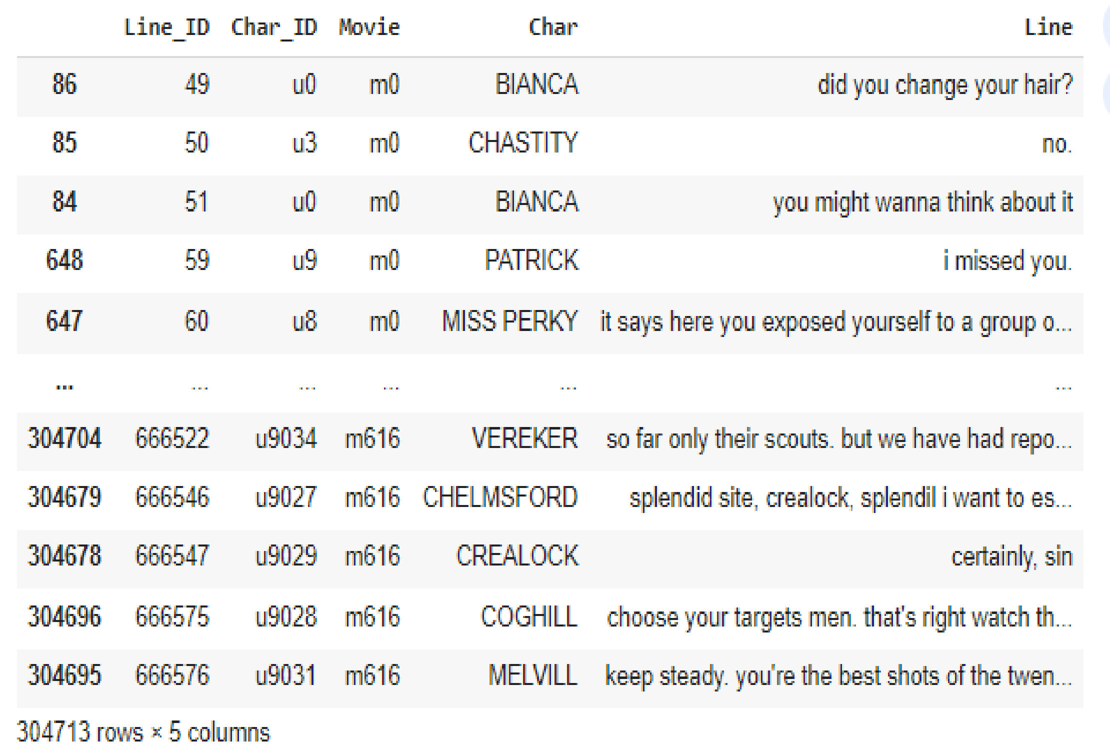
Introduction

Goal: Create and train a chatbot using a given data set.

Dataset Overview

Cornell Movie-Dialog Corpus dataset.

- Movie_lines.txt

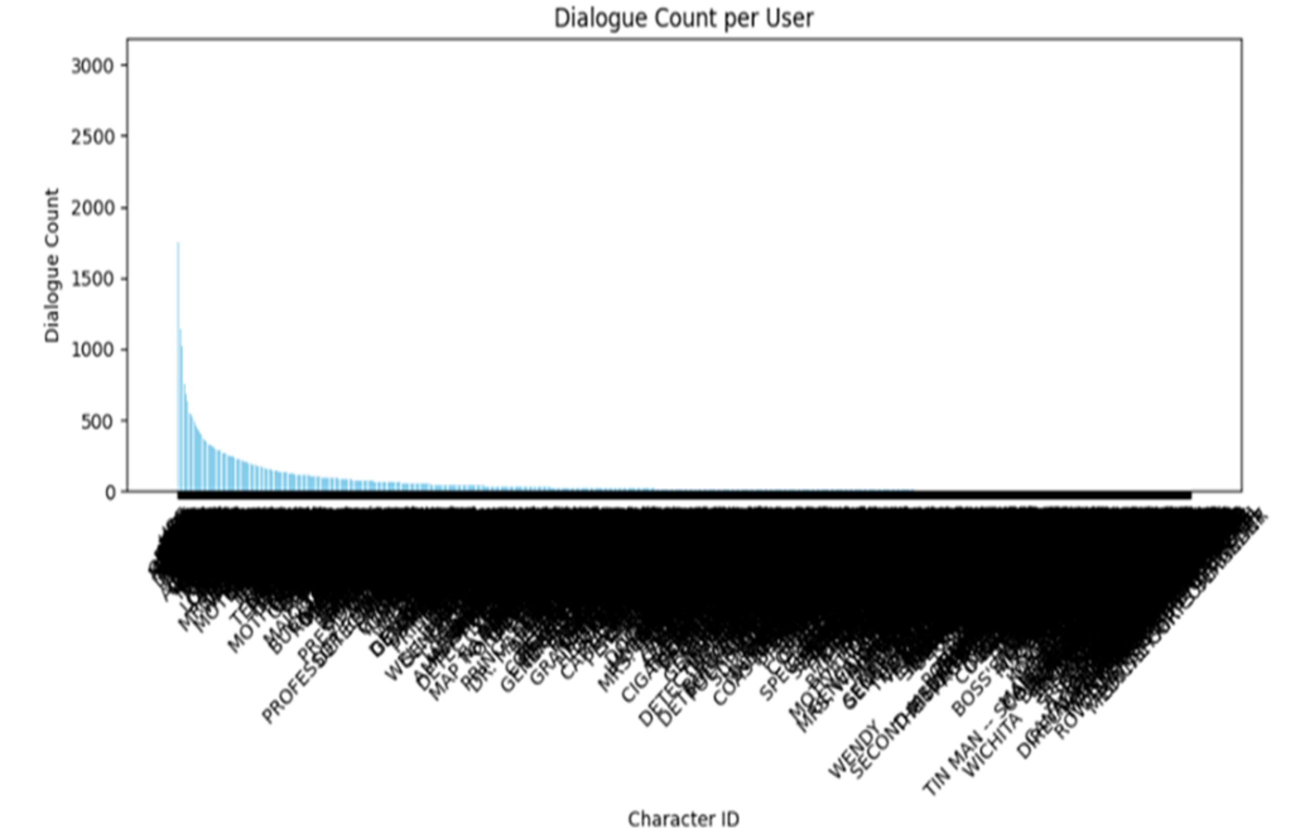


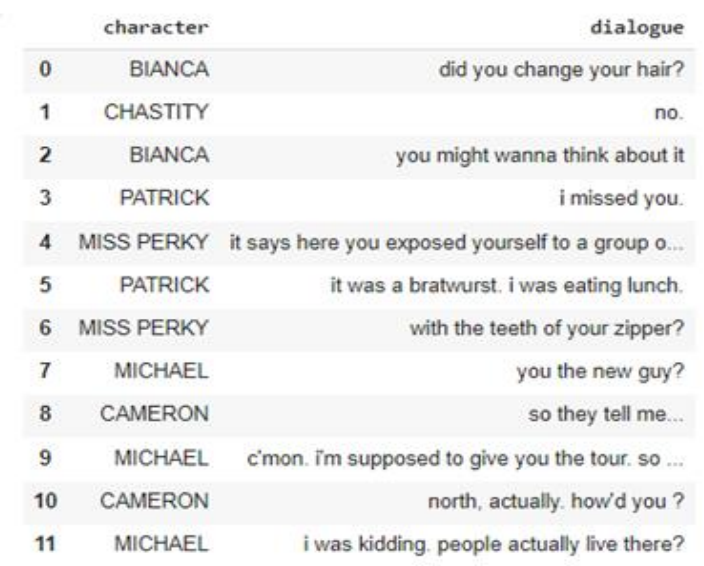
	Line_ID	Char_ID	Movie	Char	Line
86	49	u0	m0	BIANCA	did you change your hair?
85	50	u3	m0	CHASTITY	no.
84	51	u0	m0	BIANCA	you might wanna think about it
648	59	u9	m0	PATRICK	i missed you.
647	60	u8	m0	MISS PERKY	it says here you exposed yourself to a group o...
...	...	...	...	...	...
304704	666522	u9034	m616	VEREKER	so far only their scouts. but we have had repo...
304679	666546	u9027	m616	CHELMSFORD	splendid site, crealock, splendil i want to es...
304678	666547	u9029	m616	CREALOCK	certainly, sin
304696	666575	u9028	m616	COGHILL	choose your targets men. that's right watch th...
304695	666576	u9031	m616	MELVILL	keep steady. you're the best shots of the twen...

304713 rows × 5 columns

Data Preprocessing

- Changing text to lower case
- We kept stop words, and punctuation
- Structuring dialogue (text grouping)





	character	dialogue
0	BIANCA	did you change your hair?
1	CHASTITY	no.
2	BIANCA	you might wanna think about it
3	PATRICK	i missed you.
4	MISS PERKY	it says here you exposed yourself to a group o...
5	PATRICK	it was a bratwurst. i was eating lunch.
6	MISS PERKY	with the teeth of your zipper?
7	MICHAEL	you the new guy?
8	CAMERON	so they tell me...
9	MICHAEL	c'mon. i'm supposed to give you the tour. so ...
10	CAMERON	north, actually. how'd you ?
11	MICHAEL	i was kidding. people actually live there?

Model Selection, Training, and Fine-tuning

MODEL SELECTION

- GPT2LMHeadModel
  - Corresponding Tokenizer
  - Huggingface transformer library
  - Lighter processing load

TRAINING AND FINE-TUNING

- Trainer Object
- Data Collator
- Chatbot
  - Hyperparameters
  - Conversation Loop and Response Loop

Challenges Faced

- BERT Model difficulty processing
- Line sequencing

Future Improvements

- Line sequencing understanding
- More data (use of additional data parameters)
- Post-processing (checks and balances)
- Evaluation Metrics (BLEU and ROGUE)
- Physical crosschecks

Thank You