0: Framing the Problem

I do not have any projects that show I understand the data science process

I want to build a language model that you can ask questions to, and receive an answer in the syntactical style of Jiddu Krishnamurti

1: Collecting the data

Data is a numerical representation of a small part of reality we are trying to model

I want to collect data from transcripts of Jiddu Krishnamurti’s speeches and information

about where they were spoken

I want to store this data in a JSON file

2: Cleaning the data

The longest part of a data science process where you ensure your data is consistent and valid --getting rid of duplicates, missing, and null values. “Garbage in, Garbage out”

I want to format all the discussions and metadata in the same way

I could use regular expressions to pick out all the instances in which speakers are used

3: Exploring the data

This process is intended to find any remaining problems with your data, as well as generate insight that can be used during the rest of the data science process

I would like to have know high-level facts about the data we are using

(when were most of texts, where were they spoken)

4: Develop a model

Use the cleaned data and insights to build and deploy a language model

I want to train a language model on the texts of Krishnamurti

I want to explore all language models available

I want to understand the logic behind the model

What is it doing behind the scenes? What are its assumptions?

5: Interpret and Communicate

Interpret what your program is doing and how well it is doing it. Communicate those two things.

I want to interpret what my program is (a program that model the English syntax of

Krishnamurti) and what is not (a model reproducing the semantics of Krishanmurti’s philosophy)

Create an organized repository of each step in this data science process