**Some observations**

**Statistics on input length (This is variable depending on sampling)**

Mean length: 3.49345

Std length: 2.3171657466612094

Max length 36

Max length for model: 10 (mean + 3\*std)

Max length within 3 std is generally 10, but there are quite a few titles which even reach 50 in length

**Base GPT2 model**

Output within max seq length: 1.0

Output within req seq length: 0.946

Output equal req seq length: 0.829

MSE req vs gen seq length: 1.386

**Distilled GPT2 model**

Output within max seq length: 1.0

Output within req seq length: 0.945

Output equal req seq length: 0.789

MSE req vs gen seq length: 0.974

Base model generally produces more sequences which are of exact length required by the input

However, distilled model has lower MSE between lengths of generated sequences vs required length

**Rationale for using Distilled GPT2**

I have used Distilled GPT2 which has 82 million parameters. The main reason for using this model is the dependency of tokenizer. The Distilled GPT2 pretrained using the same tokenizer as GPT2. Other pretrained models have their own tokenizers which are different. Since we need to calculate loss between student and teacher model in distillation training, having different tokenizers prevent model from converging.