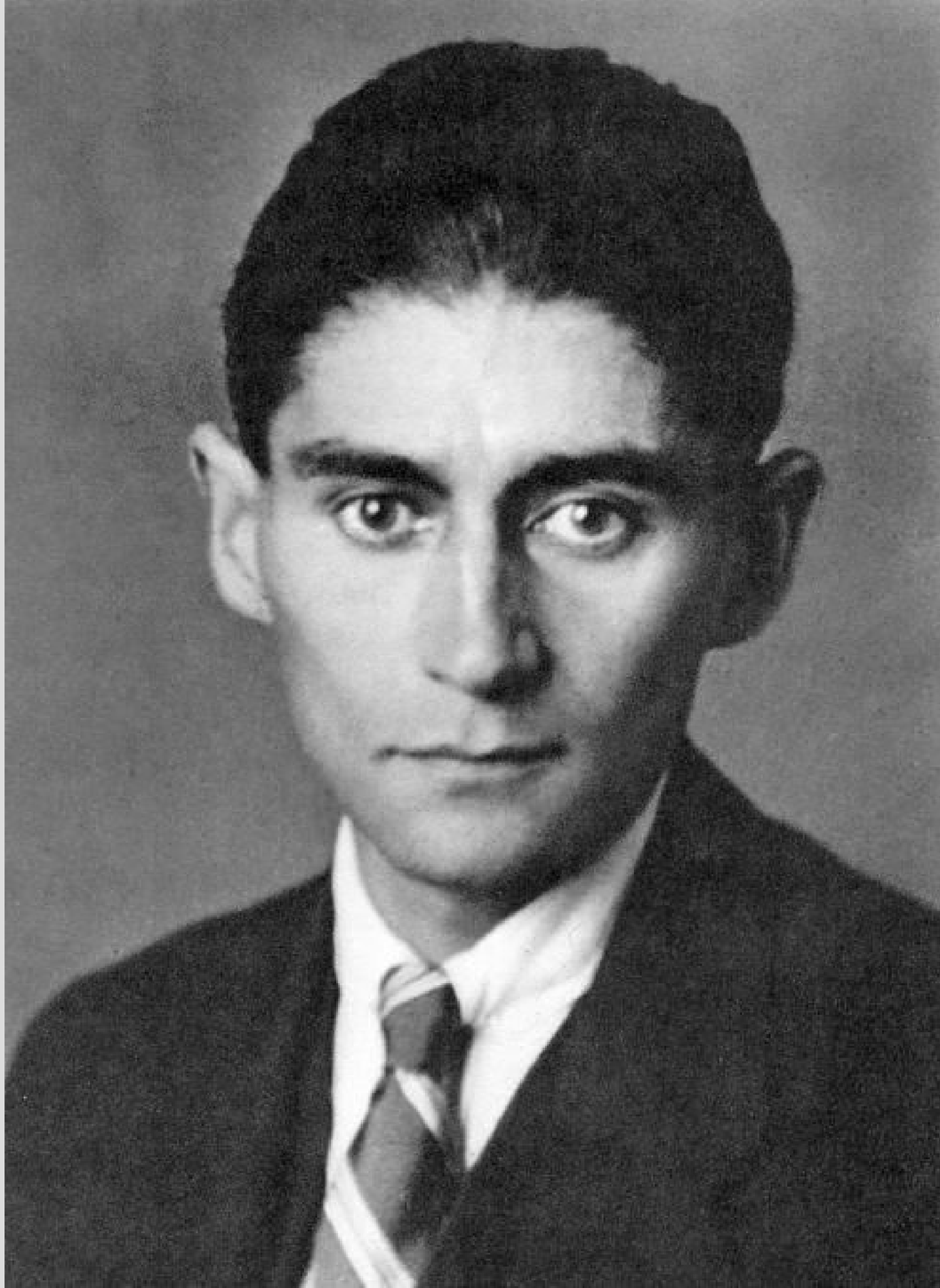
Generating Kafka

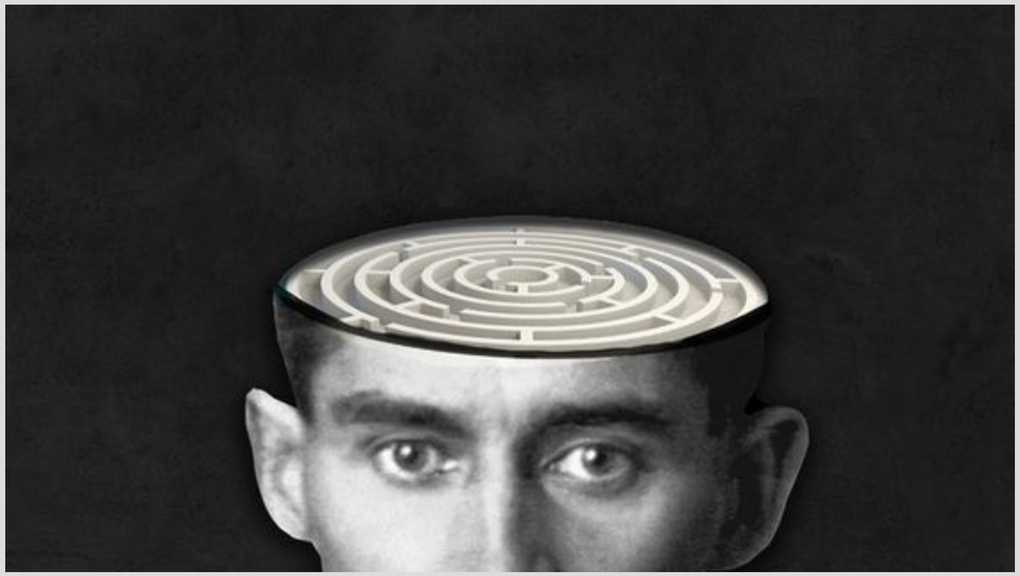
TEXT GENERATION USING
NATURAL LANGUAGE MODELLING

Who is Frank Kafka?

Frank Kafka was a German-speaking writer born on 3 July 1883 in Prague, Czech Republic. He is widely regarded as one of the major figures of 20th-century literature. He is known as the creator of a unique genre called Kafka-esque.



What is Kafka-esque?

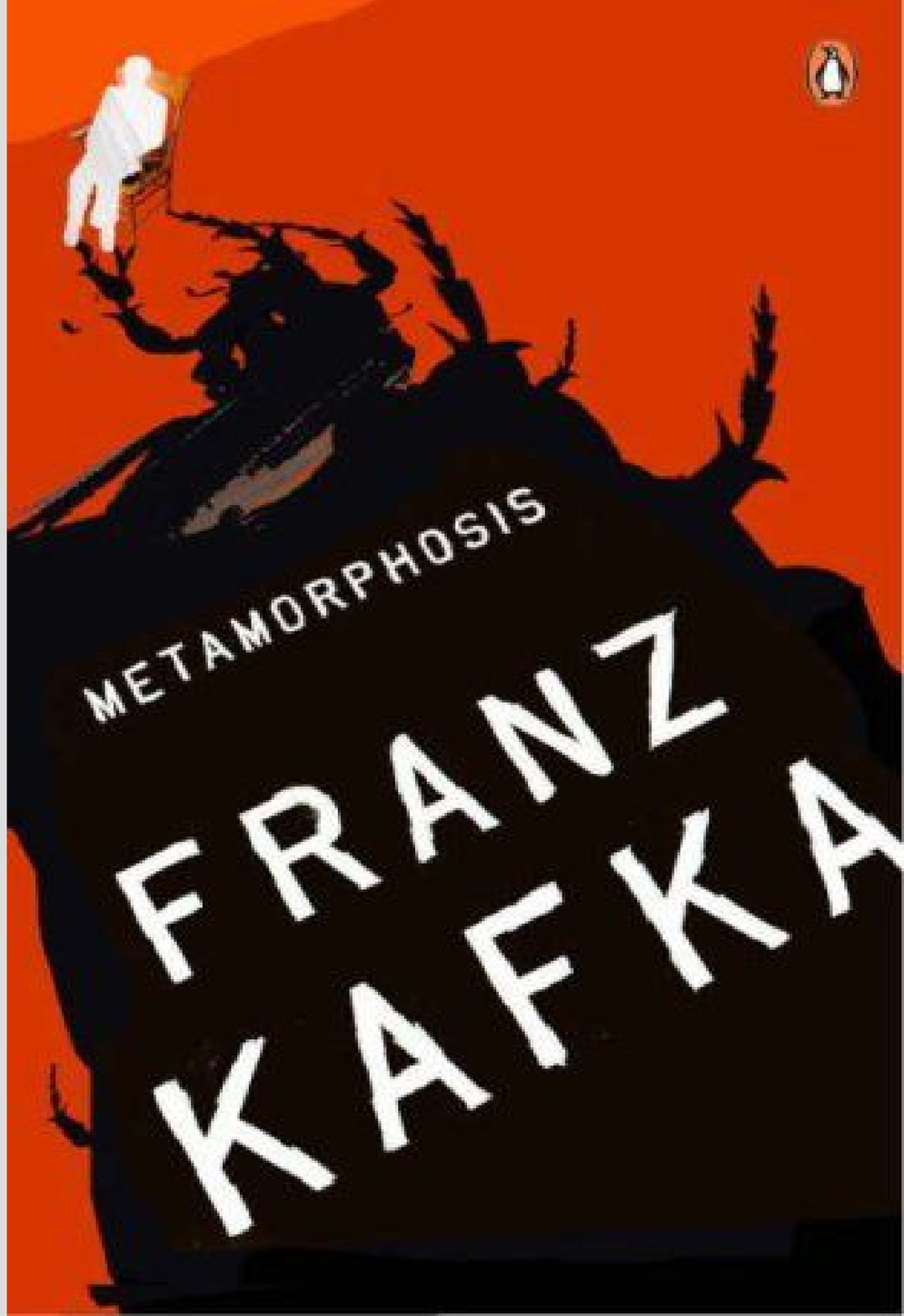


Kafka-esque is a genre of story-telling that fuses the elements of realism and fantasy. It has been widely adapted in culture, often used in movies, music and theatre. The stories usually possess a nightmarish, complex, bizarre, or illogical quality.

Metamorphosis

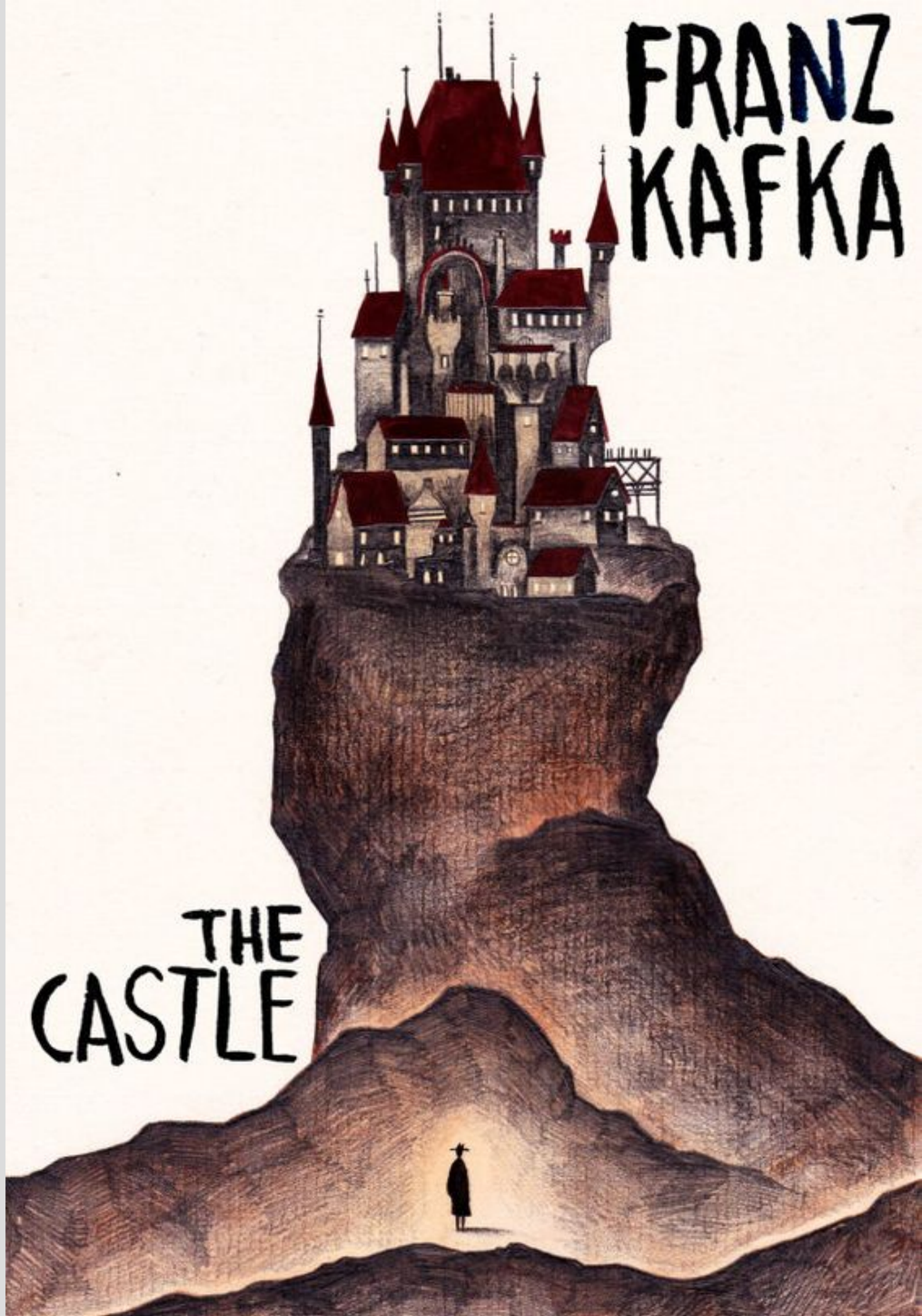
For this study, we will be using the complete text from the english translation of "Metamorphosis" to build a text generation model using natural language processing.

Metamorphosis tells the story of salesman Gregor Samsa, who wakes one morning to find himself unexpectedly transformed into a huge beetle and subsequently struggles to adjust to his new condition.



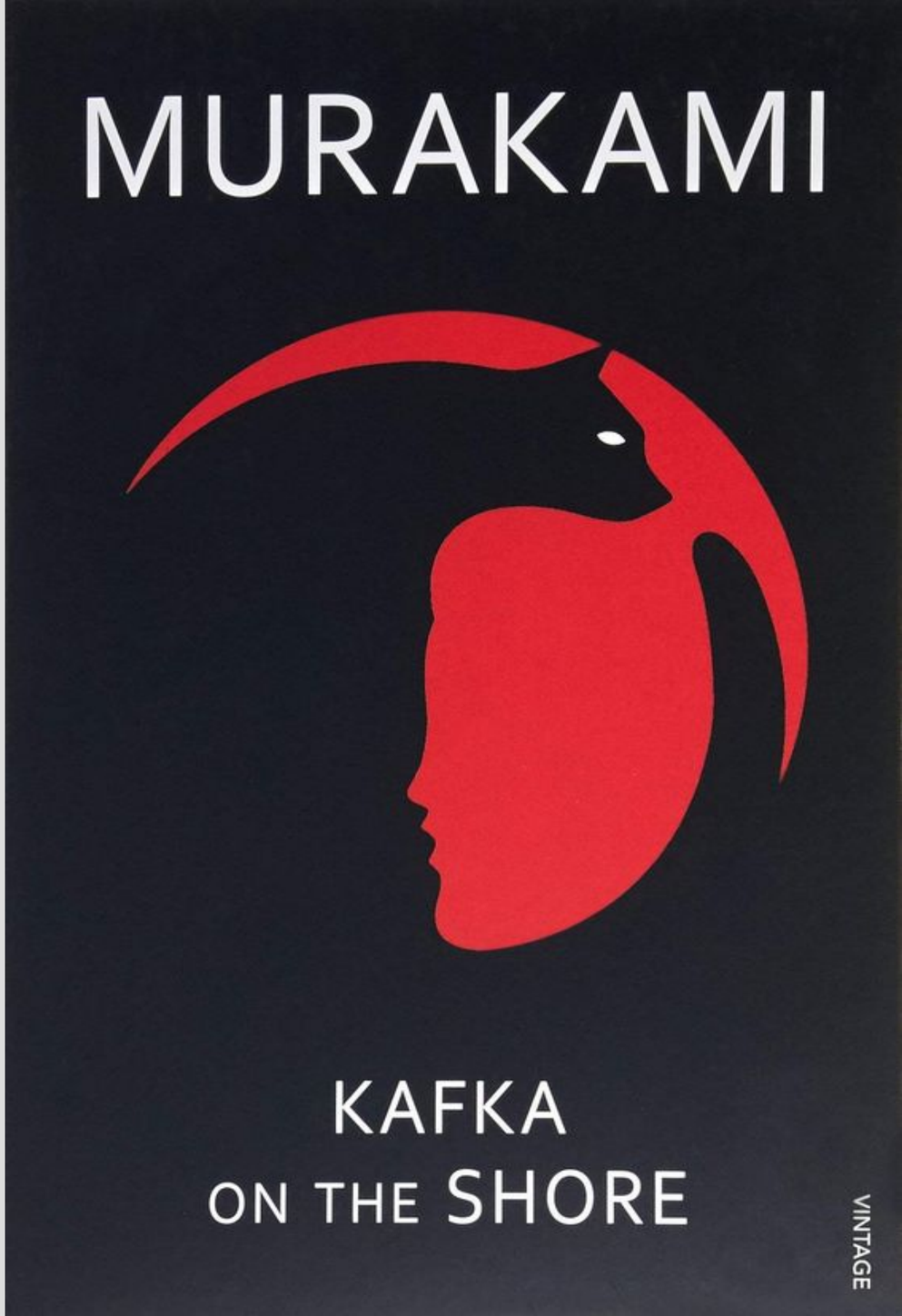
The Castle

To test the text generation model further, we will input a random excerpt from "The Castle" by Frank Kafka to observe how the model performs on unseen text from the same author.



Kafka on the Shore

In addition, we will input another excerpt from "Kafka on the Shore" by Haruki Murakami to test the model on unseen text from a separate author.



What is Neural Language Modelling?

Natural language processing is the area of study dedicated to the automatic manipulation of speech and text by software. A language model can predict the probability of the next word in the sequence, based on the words already observed in the sequence. Neural network models are a preferred method for developing statistical language models because they can use a distributed representation where different words with similar meanings have similar representation.

Problem Statement

Literature is an important part of our education, it can inspire empathy and give people a new perspective on their lives and that of others. However, creating stories requires a great amount of effort. What if we could use a language model to understand writing patterns and create new stories. This could help us generate new ideas and aid in the creative process for a new piece of work. With the text generation model, a body of text could be produced by inputting a few lines of text.

Evaluation Metrics

BLEU

(Bilingual Evaluation Understudy)
is a precision focused metric that calculates n-gram overlap of the reference and generated texts. In this study, Sentence BLEU is used to evaluate a candidate sentence against one or more reference sentences.

ROUGE

(Recall-Oriented Understudy for Gisting Evaluation)
is a set of metrics that measures the number of matching 'n-grams' between the model-generated text and a 'reference'. This indicates the extent of similarity between the generated text and the reference text sequence.

Sequence Length

EDA

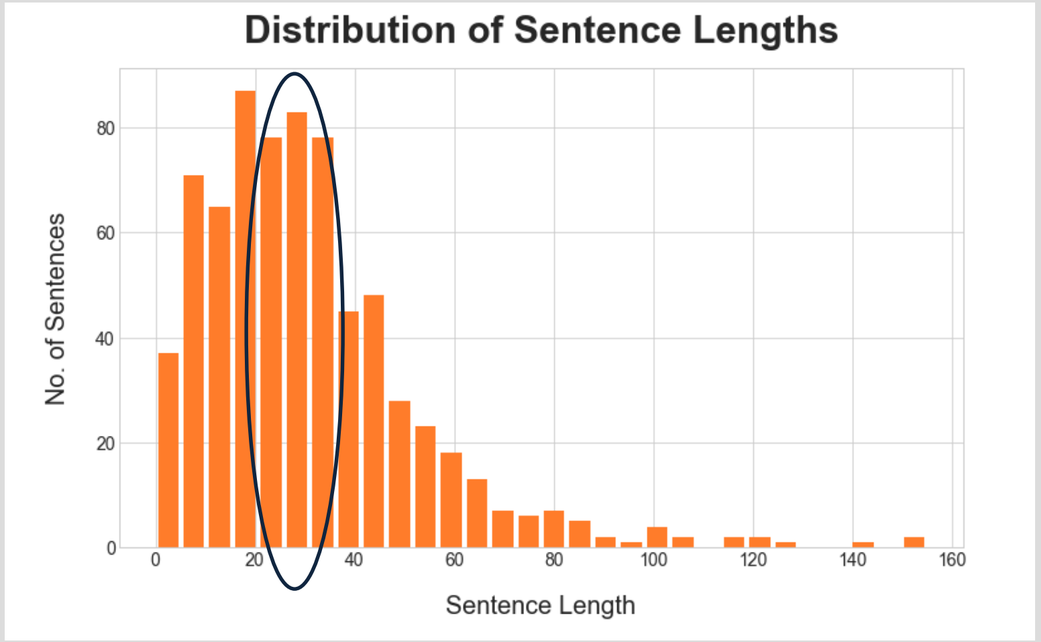
Average Sentence Length ~30 words

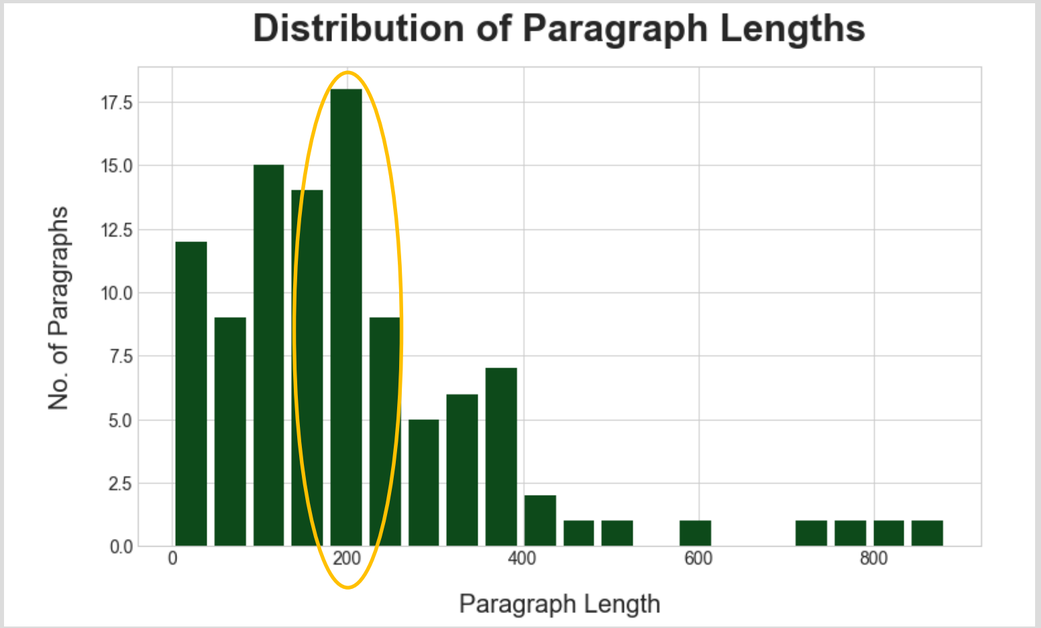
Average Paragraph Length ~200 words

MODELLING

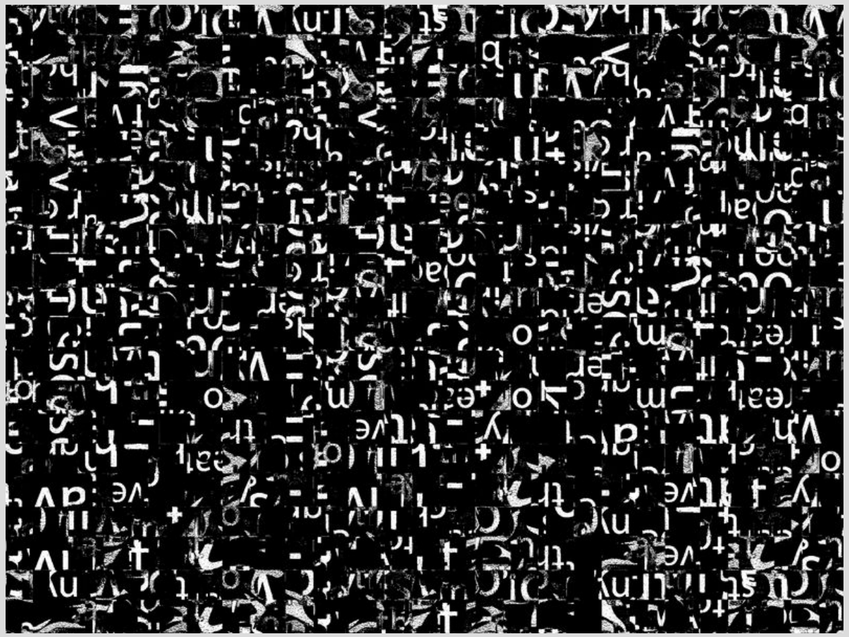
Reference Model Sequence Length – 50 words

Extended Sequence Length – 100 words





Text Processing



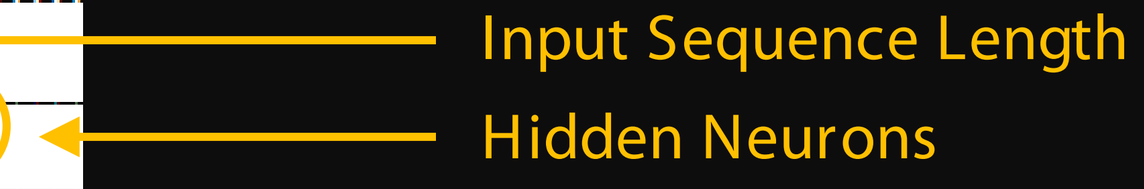
Several operations were performed to transform the raw text into a series of tokens to train the model.

- Replaced “ - “ with white space
- Split the words based on white space, “ “
- Removed all punctuation from words
- Removed all words that are non-alphabetic
- Normalized all words to lowercase

Model Architecture

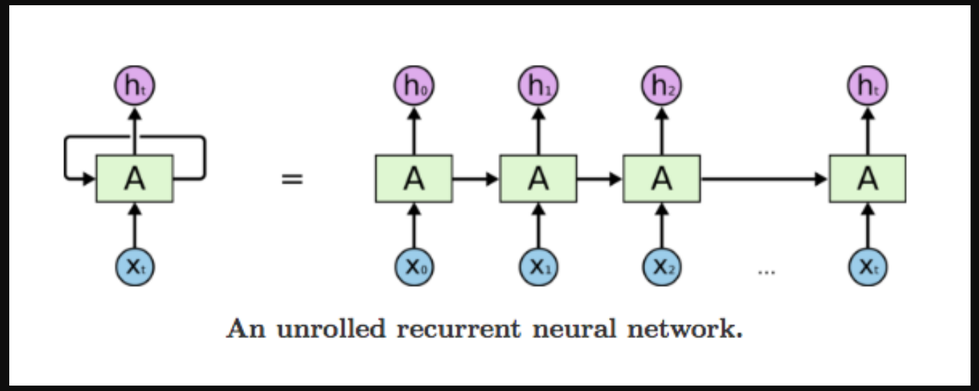
Reference Model

```
_________________________________________________________________
Layer (type)                 Output Shape
=================================================================
embedding (Embedding)        (None, 50, 50)
_________________________________________________________________
lstm (LSTM)                  (None, 50, 100)
_________________________________________________________________
lstm_1 (LSTM)                (None, 100)
_________________________________________________________________
dense (Dense)                (None, 100)
_________________________________________________________________
dense_1 (Dense)              (None, 2534)
=================================================================
Total params: 533,534
Trainable params: 533,534
Non-trainable params: 0
_________________________________________________________________
```

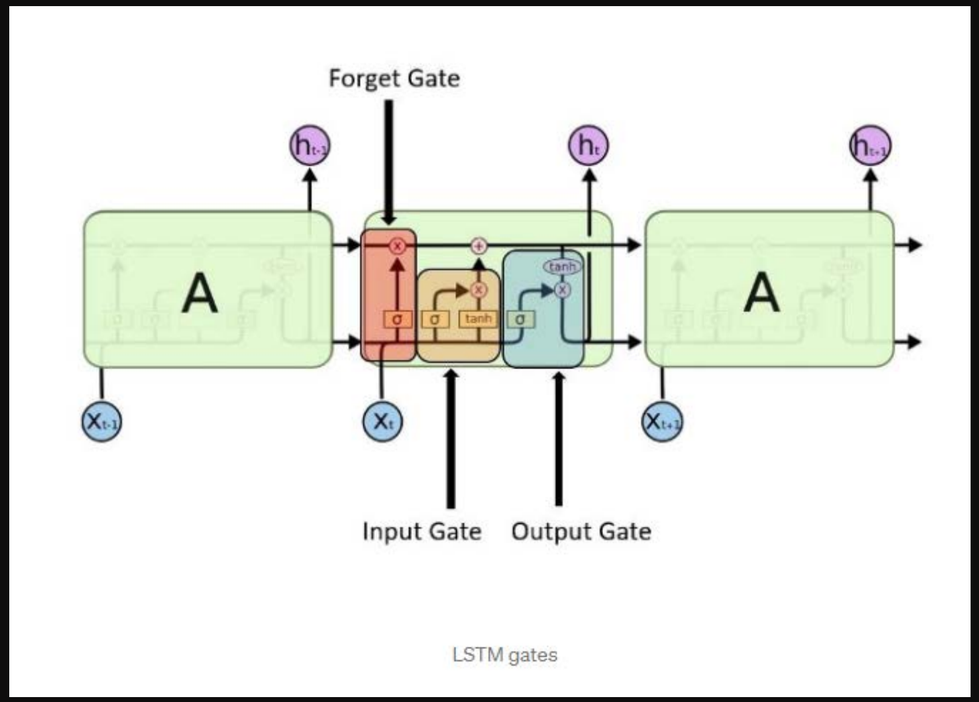


LSTM

- Long Short-Term Memory (LSTM) networks are a modified version of recurrent neural networks, which makes it easier to remember past data in memory.
- The vanishing gradient problem of RNN is resolved with LSTM.
- LSTM is well-suited to classify, process and predict text in a text generation model.



An unrolled recurrent neural network.



LSTM gates

Model Optimization

1. Number of hidden neurons was increased from 100 to 150.
2. Number of Epochs increased from 100 to 150
3. Dropout layers were added after each LSTM layer to reduce overfitting, using dropout rate of 0.2.
4. A unidirectional LSTM layer was replaced with a bidirectional LSTM layer to study the past and future inputs.
5. Decreased batch size from 128 to 64 to increase the number of iterations per epoch.

Optimized Model

Layer (type)	Output Shape
embedding (Embedding)	(None, 50, 50)
bidirectional (Bidirectional	(None, 50, 300)
dropout (Dropout)	(None, 50, 300)
lstm_1 (LSTM)	(None, 150)
dropout_1 (Dropout)	(None, 150)
dense (Dense)	(None, 150)
dense_1 (Dense)	(None, 2534)

Total params: 1,043,784
Trainable params: 1,043,784
Non-trainable params: 0

Model Performance

Input Sequence

“to avoid being seen at the window during the day the few square meters of the floor did not give him much room to crawl about it was hard to just lie quietly through the night his food soon stopped giving him any pleasure at all and so to entertain himself”

Reference Output

“he got into the habit of crawling up and down the walls and ceiling he was especially fond of hanging from the ceiling it was quite different from lying on the floor he could breathe more freely his body had a light swing to it and up there relaxed and almost”

Model Performance

Output from Reference Model

“he was entitled to even if he was not hungry sister no longer thought about painfully is something that they had electric motors an apple thrown without much force glanced against back and slid off without doing the door i beg of and seemed the door behind her to insist”

Output from Optimized Model

“he would have woken him to enable the family and was no longer able to take the chief clerk had been going on the living room he had already finished close to the door was opened to the task of swinging the clock struck went on the floor and needed”

Results & Analysis

Score Type	Ref Model	Ref Model LS	Opt Model
Cumulative 1-gram	0.2549	0.2871	0.2353
Cumulative 2-gram	0.1428	0.0758	0.1815
Cumulative 3-gram	0.0963	0.0399	0.1621
Cumulative 4-gram	0.0645	0.0155	0.1432

Table 1: BLEU Scores for "Metamorphosis" Text

Score Type	Metric	Ref Model	Ref Model LS	Opt Model
Rouge-1	Recall	0.2500	0.2740	0.3000
	Precision	0.2941	0.2778	0.2791
	F1	0.2703	0.2759	0.2892
Rouge-2	Recall	0.0800	0.0200	0.1400
	Precision	0.0975	0.0208	0.1429
	F1	0.0879	0.0204	0.1414
Rouge-l	Recall	0.2000	0.1233	0.2500
	Precision	0.2353	0.1250	0.2326
	F1	0.2162	0.1241	0.2410

Table 2: ROGUE Scores for "Metamorphosis" Text

Results & Analysis

Metamorphosis	Cumulative 1-gram	0.23525
	Cumulative 2-gram	0.18150
	Cumulative 3-gram	0.16214
	Cumulative 4-gram	0.14323
The Castle	Cumulative 1-gram	0.23525
	Cumulative 2-gram	0.09702
	Cumulative 3-gram	0.02777
	Cumulative 4-gram	0.01415
Kafka on the Shore	Cumulative 1-gram	0.03921
	Cumulative 2-gram	0.00886
	Cumulative 3-gram	0.00572
	Cumulative 4-gram	0.00428

Table 3: BLEU Scores for Optimized Model

- The BLEU and ROGUE scores for "The Castle" were lower than "Metamorphosis" since the word dictionary was derived from the latter.
- As a result, some of the words used in "The Castle" may not be found in the model's word dictionary.

- The BLEU and ROGUE scores for “Kafka on the Shore” were lower than "The Castle" text.
- This could be due to different gramatical structures between Frank Kafka and Haruki Murakami.
- Both authors produced their work in different time periods, Kafka in 1920 and Murakami in 1990.

- The output text from both unseen texts retained the main character, Gregor, from "The Metamorphosis".

Conclusions

- The optimized model showed better BLEU and ROUGE scores than the reference model.
- The two most significant changes that improved model performance were introducing a bidirectional LSTM layer and decreasing the batch size from 128 to 64.
- Increasing the sequence length of the input text resulted in lower BLEU and ROUGE scores, hence a shorter sequence length would have better model performance.
- When considering different types of input text, unseen text from the same author achieved higher BLEU and ROUGE scores than an unseen text from a different author. However, the output text still referenced the main character and settings from the training text.

Recommendations

To improve the model, some of the following methods can be explored.

1. Use multiple books from the same author to produce a larger collection of reference text.
2. Use multiple books from several authors that have similar writing styles produced during a similar period.
3. Use GloVe word embedding vectors to increase representation of words.
4. Split the raw data based on sentences and pad each sentence to a fixed length.

“By believing passionately in something that still does not exist, we create it. The nonexistent is whatever we have not sufficiently desired.”
– Frank Kafka