Science Fiction Text Generation

The main objective of this project is generating science fiction through NLP text generation. The dataset given was raw text including multiple science fictions. Mainly, Word2Vec and LSTM are used in this task.

Preprocessing

For preprocessing, the raw text is split into list of sentences using regular expressions by avoiding abbreviations, acronyms, ellipses and decimal numbers. Even though there exist sentence tokenizer from NLTK and SpaCy, regular expression turns out to be fastest with large dataset. After getting list of sentences, each sentence is tokenised into words again. The step of splitting into sentences could be skipped, however, since Word2Vec is performed for word embedding, it was necessary to meet the input requirement in this project. Furthermore, stop words removal and lemmatisation is skipped on purpose since it is expected to see those two in the science fiction's writing style.

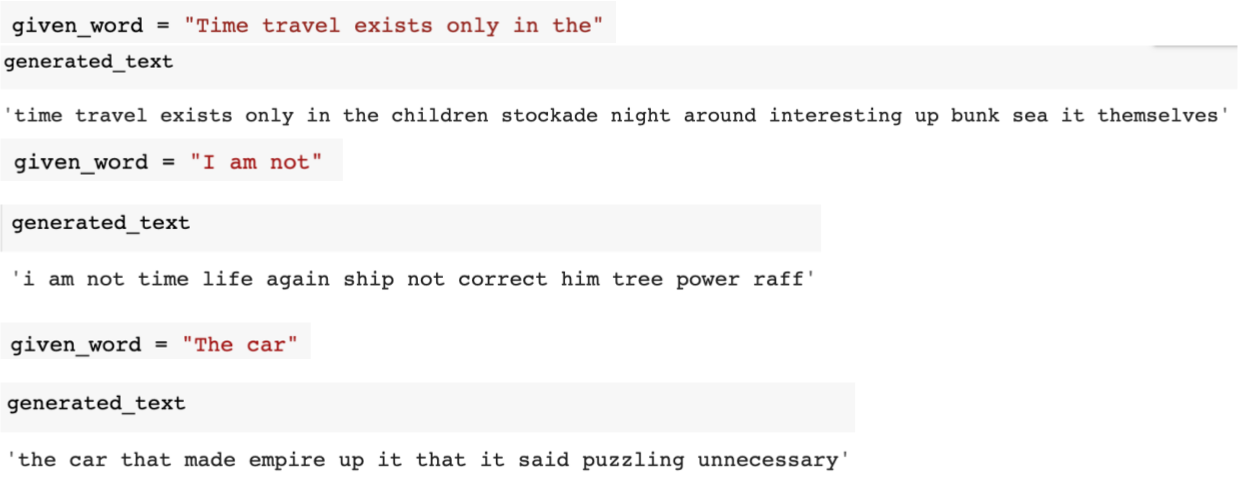
Problems & Solutions

The challenge of this task was the model was generating random characters instead of words. It was suspected that the model was not learning enough to generate words. Thus, different parameters are tried. Specifically, increasing the learning rate of LSTM – 0.01, 0.001, 0.0001 – and increasing the size of Word2Vec – 50, 100 – and increasing the batch size from 128 to 2048 help the model to learn more. Moreover, to make the model deeper and more complex, two Dropout and one more LSTM layers are added. After adjusting all these approaches, the final hyperparameter is set as follows,

- Word2Vec: size=50, min_count=1, window=5 and iter=3
- LSTM: 2 LSTM layers with num_units=2, 2 Dropout layers with dropout_rate=0.3 and learning_rate=0.01
- Batch_size=2048 and epochs=50

Results

With the final architecture, it succeeded to generate words as shown in Figure 3. It generated the words very well, however, sometimes the sentence is grammatically incorrect. This is because of the lack of data; thus, it is expected to see more accurate text generation with more books of science fiction in the future.



```
given_word = "Time travel exists only in the"
```

```
generated_text
```

```
'time travel exists only in the children stockade night around interesting up bunk sea it themselves'
```

```
given_word = "I am not"
```

```
generated_text
```

```
'i am not time life again ship not correct him tree power raff'
```

```
given_word = "The car"
```

```
generated_text
```

```
'the car that made empire up it that it said puzzling unnecessary'
```

Figure 3: Generated Texts from Given Words