# 自然语言处理实验

## 项目目标

实现一个通过少量输入来预测后续输入的词预测系统，加深自己对于自然语言处理的理解。

## 项目过程

- 数据搜集
- 数据处理
- 模型选取
- 结果展示

### 数据搜集

以卡夫卡中篇小说《变形计》作为数据集，该数据集共包含约22000个单词，可提取2600余个不重复词汇。

### 数据处理

- 从全文随机选取80%作为训练集，20%作为测试集。
- 将所有词汇转为小写字母。
- 以句子为单位分割文本，同时去除句号之外的其它标点符号。
- 依次提取所有不重复的词汇，生成词汇表。
- 使用滑动窗口法，对每个句子依次生成若干词向量。

### 模型选取

使用了 Keras 的 Sequential 模型：

```
model = Sequential()
model.add(Embedding(VocSize, 10, input_length=1))
model.add(LSTM(1000, return_sequences=True))
model.add(LSTM(1000))
model.add(Dense(1000, activation="relu"))
model.add(Dense(VocSize, activation="softmax"))
```

### 结果展示

- 进行150轮迭代训练，约40轮后loss曲线趋于平滑，150轮后达到最低点。




- 对测试集进行验证，显示误差为0.09。

```
Model loaded successfully.
Test dataset loaded successfully.
57/57 ━━━━━━━━━━━━━━━━━━━━ 1s 17ms/step - loss: 0.1044
Test Loss: 0.09336572885513306
请按任意键继续. . .
```

- 切换为预测模式后，即可基于输入的词汇给出实时预测：

```
Enter a line of text: it was not possible
to for as
Enter a line of text: it was not possible to
perform move quietly
Enter a line of text: it was not possible to perform
anything that a
Enter a line of text: it was not possible to perform anything
with while and
Enter a line of text: it was not possible to perform anything with
jaws little some
Enter a line of text: it was not possible to perform anything with jaws
that to the
Enter a line of text: it was not possible for
him all how
Enter a line of text: it was not possible for him
to he and
Enter a line of text: it was not possible for him to
stay keep think
Enter a line of text: it was not possible for him to stay
in for and
Enter a line of text: it was not possible for him to stay in
bed it the
Enter a line of text: it was not possible for him to stay in bed
and than he
```