

# RNN 实现情感识别任务

## Deadline:

5 月 22 日前将报告和源码打包成压缩文件按照姓名+学号+homework3 的格式发送到助教邮箱 [xinfeiliu@mail.ustc.edu.cn](mailto:xinfeiliu@mail.ustc.edu.cn)

IMDB 影评数据集，带有 positive 和 negative 两类标签

下载地址：<https://ai.stanford.edu/~amaas/data/sentiment/>

也可以使用 keras 库中预处理好的 imdb dataset

```
from keras.datasets import imdb
```

## 对文本分析实现情感识别

### 1.1 任务目标

训练 RNN、transformer 或 bert 等序列神经网络，来对 imdb 数据集的文本进行情感识别。

### 1.2 开发环境

安装 Pytorch, keras 等

### 2.3 实现流程

推荐使用 `huggingface` 中 `transformers` 库中包含的分词器或预训练模型:

```
from transformers import BertModel
```

```
model = BertModel.from_pretrained("bert-base-uncased")
```

```
outputs = model(**inputs)
```

- 数据预处理，设置最大文本序列长度等。
- 分词 Tokenization: 将文本划分成 token，包含 word\_based、空格分词、subword\_based 以及 byte pair encoding(BPE)等方法。
- Word Embedding: one hot、word2vec。
- 构建网络模型，训练。可以灵活选 RNN 模型的结构，单向、双向 RNN，LSTM、Tansformer 等。

### 3 评价指标

- 测试集上情感识别的准确率。