RNN 大作业

姓名： 赵敬业

学号： 2193712622

班级： 大数据 91

2022 年 5 月 22 日

目录

1 数据集说明

数据集是人民日报 2018、2019、2020 年三年的所有文字内容，来源于网络 [1]

2 代码说明

主要代码借鉴王迪老师的课件 LSTM 模型部分，只使用一层进行训练。
对代码的修改主要有两个地方

a) 训练过程增加了数据回收机制，降低显存的占用, 如图1。

b) 数据预处理使用 jieba 进行分词，并引入了 jieba_rate 的概念。笔者首先尝试了 jieba 分词，但是导致了显存占用过多，最终引发溢出，之后笔者引入了 jieba_rate 的概念，即在分词过后，只保留一定比率 (jieba_rate) 的最高频率的词，剩下的词使用单个汉字进行切割的方法，最终形成数据集，但是由于这是第一次尝试，jieba_rate 依然偏高，导致训练效果不理想，可以尝试考虑进一步降低 jieba_rate。代码如图2

```
58              X, Y = data_[:, :-1], data_[:, 1:]  # X是前n-1列，Y是后n-1列，X和Y的形状是(batch_size, num_steps)
59              (output, state) = model(X, state)  # output: 形状为(num_steps * batch_size, vocab_size)
60
61              del data_,X
62              gc.collect()
63              torch.cuda.empty_cache()
```

图 1: cuda 删除冗余变量

```
35  print("使用jieba分词...")
36  corpus_list=jieba.lcut(corpus_chars)#jieba分词，结果是list
37  corpus_list_no1=[i for i in corpus_list if len(i)>1 and re.match(r"[\u4e00-\u9fa5]",i)]#删掉不是中文以及只有一个的词
38  #统计最高的词频，只保留opt.jieba_rate比率的词
39  temp=Counter(corpus_list_no1).most_common(round(opt.jieba_rate*len(corpus_list_no1)))
40  corpus_list=[temp[i][0] for i in range(len(temp))]
41  chars = set(corpus_list)
42  print("jieba总共分出来了",len(chars),"个词")
43  # print("字一共有:",len(set(corpus_chars)),"个")
44  chars = set(corpus_chars).union(chars)
45  # 整个文本所有的入选词汇，包括jieba分出来的以及文中的所有字，用jieba.add_word功能权重加到非常大
46  # 然后再用jieba分词分成想要的效果，比如"中华人民共和国成立70周年"
47  # 原来分成"中华人民共和国/成立/70/周年"
48  # 现在分成"中华人民共和国/成/立/70/周/年"(只有中华人民共和国是高频词)
49  for char in chars:
50      jieba.add_word(char,10000000000000)
51      if len(char)>1:
52          jieba.suggest_freq(char, tune=True)
53  corpus_list=jieba.lcut(corpus_chars,HMM=False)
54  chars = set(corpus_list)
55  print("分词完成,最终入选的字和词有",len(chars),"个...")
56  char_to_idx = {char: i for i, char in enumerate(chars)}  # 生成char与index的对应关系
57  idx_to_char = {i: char for char, i in list(char_to_idx.items())}
58  corpus_indices = [char_to_idx[char] for char in corpus_list]  # 将文章中的字符转化为数字
```

图 2: jieba 分词流程展示

3 训练过程和结果说明

训练过程在 colab 上完成,故主要输出都在笔记本中,另外使用 Shenmuxing_lstm_Marx.txt 手动保存所有输出。

最终结果粘贴于此

epoch0 epoch 0, perplexity 275.177697, time 4670.07 sec 面对复杂局面应如何运用好辩证思维?，是中国特色社会主义制度的重要组成部分，是我国社会主义制度的根本制度，是我国社会主义制度的必然要求。　　“我们要把党的政治建设摆在首位，坚持以人民为中心，坚持以人民为中心，坚持以人民为中心，坚持以人民为中心，坚持以人民

为中心，坚持以人民为中心，把人民群众生命安全和身体健康放在第一位，我们党的政治建设和人民群众的政治生活、人民生活，我们要把人民群众生命安全和身体健康放在第一位，我们党的十八大以来，我们党的十八大以来，我们党的十八大以来，我们党的十八大以来，我们党的十八大以来，我们党的十八大以来，我们党的十八大以来，我们党的十八大以来，我们党的十八大以来，我们党的十八大以来，我们党的十八大以来，我们党

epoch1 epoch 1, perplexity 141.428752, time 4669.75 sec 面对复杂局面应如何运用好辩证思维？的原则，是我们党的政治建设的重要组成部分，是党的十八大以来，我们党的十八大以来，我国经济社会发展取得历史性成就，是我国经济发展的重要引擎。“我们要把中国特色社会主义的发展理念和发展中，我们将继续推动高质量发展。” “中国特色社会主义进入新时代，我们将继续扩大开放，推动经济全球化朝着更加开放、包容、普惠、平衡、共赢的方向发展。” “中国是中国的发展道路，也是中国的发展机遇。” “中国是中国的发展道路，也是中国的发展机遇。” “中国是中国的发展道路，也是中国的发展机遇。” “中国是中国的发展道路，也是中国的发展机遇。” “中国

epoch2 epoch 2, perplexity 118.062671, time 4670.47 sec 面对复杂局面应如何运用好辩证思维？的问题，是我们党的政治建设的重要组成部分，是党的十八大以来，我们党的十八大以来，习近平总书记在党的十九大报告中，习近平总书记在全国两会上

epoch3 epoch 3, perplexity 107.200123, time 4668.60 sec 面对复杂局面应如何运用好辩证思维？的原则，是中国特色社会主义制度的重要组成部分，是中国特色社会主义制度的重要组成部分。 “我们要坚持以人民为中心的发展思想，坚持以习近平新时代中国特色社会主义思想为指导，坚持以习近平新时代中国特色社会主义思想为指导，坚持以习近平新时代中国特色社会主义思想为指导，坚持以习近平新时代中国特色社会主义思想为指导，坚持以习近平新时代中国特色社会主义思想为指导，坚持以习近平新时代中国特色社会主义思想为指导，坚持以习近平新时代中国特色社会主义思想为指导，坚持以习近平新时代中国特色社会主义思想为指导，坚持以习近平新时代中国特色社会主义思想为指导，坚持以习近平新时代中国特色社会主义思想为指导，坚持以习近平新时代中国特色社会主义思想为指导，坚持以习近平新时代中国特色社会主义思想为指导，坚持以习近平新时代中国特色社会主义思想

epoch4 epoch 4, perplexity 100.769481, time 4669.23 sec 面对复杂局面应如何运用好辩证思维？的原则，是中国特色社会主义制度的重要组成部分，是中国特色社会主义制度的重要组成部分，是中国特色社会主义制度的重要组成部分，是中国特色社会主义制度的重要组成部分。 （作者为中国特色社会主义理论体系研究中心研究员）图片

报道　　“我们的是，我的日子，我就能看到了我的朋友。”　　“我的日子，我就能看到我的中国朋友。”　　“我的日子，我就能看到我的中国朋友。”　　“我的日子，我就能看到我的中国朋友。”　　“我的日子，我就能看到我的中国朋友。”　　“我的日子，我就能看到我的中国朋友。”　　“我的日子，我

参考文献

[1] 人民日报数据集 (2022). URL http://spiderhub.cn/data/rmrb-data. [Online; accessed 22. May 2022].