深度学习实验五---新闻推荐系统

SC20023067 陈勇虎
SC20023128 李国威

任务背景

实验的主要内容是新闻推荐任务。

实验的任务描述如下。根据新闻浏览历史，用户**u**和一组候选新闻，目标是根据该用户的个人兴趣对这些候选新闻进行排序。在这个过程中，新闻内容可以通过内容来建模，用户可以通过用户的新闻浏览历史来建模。然后，该模型根据候选新闻与用户兴趣的相关性来预测候选新闻的点击得分。排名结果将与真实的用户点击标签进行比较，通过AUC来衡量排名质量。

在实验提供的大的验证集数据上，本文的实验结果达到了如下所示的结果。

- AUC:0.6782
- MRR:0.3278
- nDCG@5:0.3616
- nDCG@10:0.4254

数据集说明

小数据集和大数据集的差别仅仅是**news.tsv**中**impression_id**属性的区别，故不做区分。下对大数据集的处理方法和结果进行说明，本实验中只使用了**behaviors,tsv**和**news.tsv**文件，此外，使用了glove预训练词向量，该文件数据将在后续的文件链接中提供。

数据集分析

使用的数据集包括新闻数据和用户行为数据。

新闻数据主要包含以下几个内容：

属性列名	news_id	category	subcategory	title	abstract	url	title entities	abstract entities
简要描述	新闻标识	新闻主类别	新闻副类别	新闻标题	新闻概要	新闻链接	新闻标题实体	新闻概要实体

在本次实验中，主要使用了新闻的**title**属性，因为标题的内容，在一定程度上，就可以反应出新闻的主要内容。

用户数据主要包含以下几个内容：

属性列名	impression_id	user_id	time	history	impressions
简要描述	标识	用户标识	时间	用户点击历史	用户点击标签

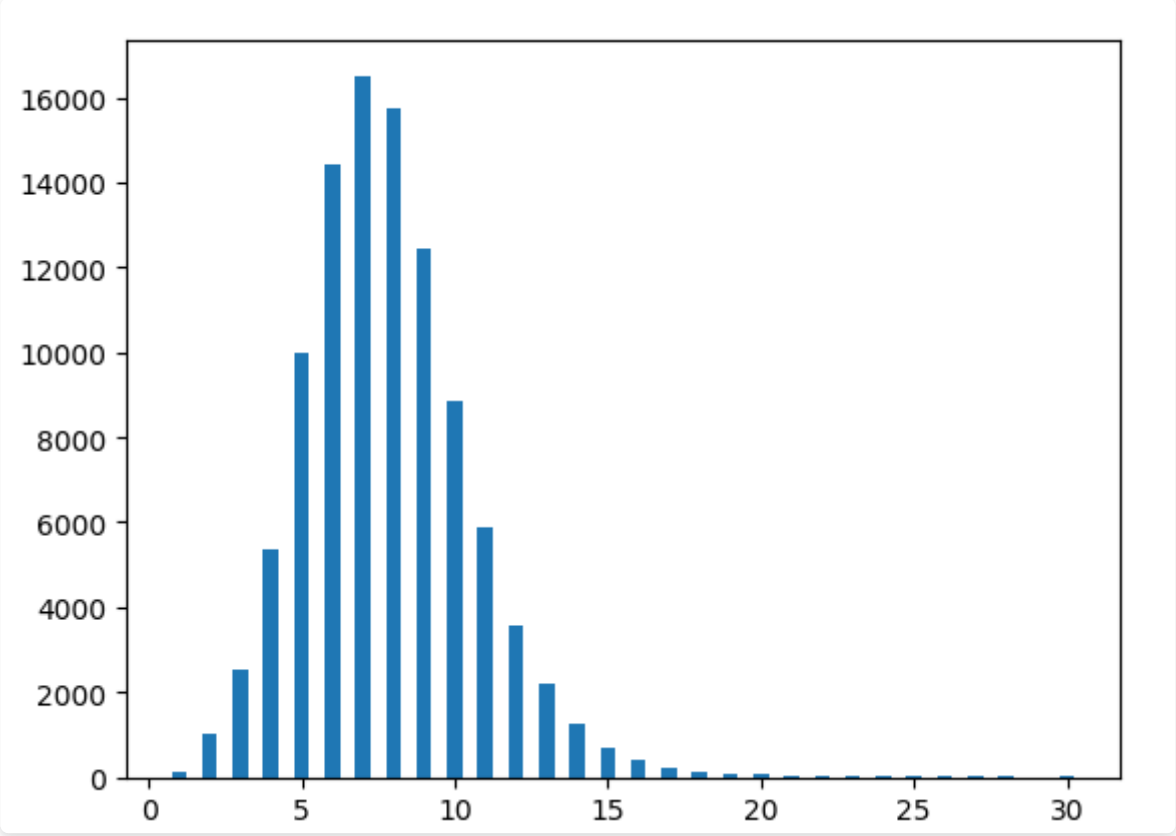
在用户行为数据中，主要使用了用户的点击历史和点击标签，通过点击的历史新闻对用户兴趣进行建模。

用户的兴趣建立在其历史点击的新闻中，因此对于该用户的候选新闻的点击率，历史点击信息反映出的用户兴趣，将会是一个很好的衡量指标。

对**训练集数据集**的统计结果如下：

news	users	impressions	avg_title_words	max_title_words	positive_samples	negative samples
101527	711222	2232748	7.79	30	3383656	80123718

其中，训练集出现的新闻标题长度统计结果如下图所示：



新闻标题长度统计

可见，大部分的新闻标题长度都在15个单词以下。因此在后续的数据处理中，固定标题长度为15。

数据处理细节

数据处理的实现见**final_data_preprocess_drop.py**文件。

stopwords停止词

停止词是自然语言中出现频率较高，但是对文章和页面的意义没有实质影响的一类词。在实验中提供的数据集中，新闻标题中也会出现诸如"the and of"等英文停止词。在实验中处理标题时，对于遇到的停止词，将会直接剔除。停止词词库方面，以stopwords.txt的方式提供。

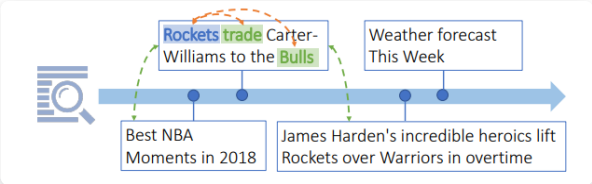
glove预训练词向量

对于新闻标题的词语，如果glove中已经含有预训练的结果，则使用glove中训练好的结果，否则，将会进行随机的初始化，用于后续的训练。

模型搭建

NRMS模型搭建

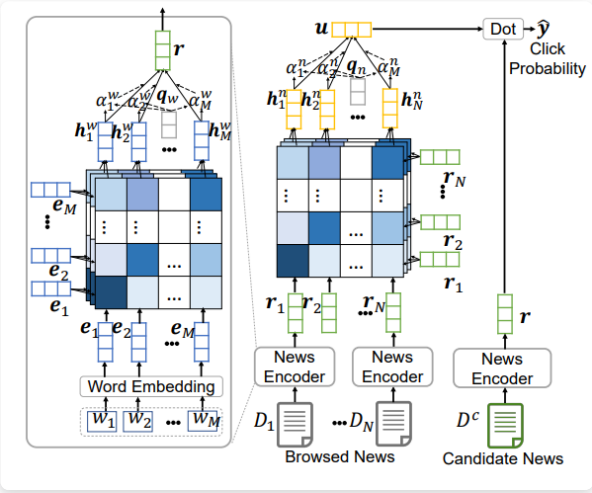
新闻的标题是对新闻内容的提炼，因此对理解新闻内容也非常重要，词语之间的联系也有利于新闻的理解。并且，不同的新闻标题之间的联系，也可以在一定程度上反映出用户的兴趣。



新闻以及词语间联系

在新闻编码器端，NRMS通过multi-head self-attention从新闻的标题中学习新闻的表示，在用户编码器端，通过用户的历史点击新闻学习用户的表示，最后通过multi-head self-attention找到其中的联系，此外，使用了一个附加的attention层选择出新闻和用户表示中的重要联系。

模型框架图如下所示:

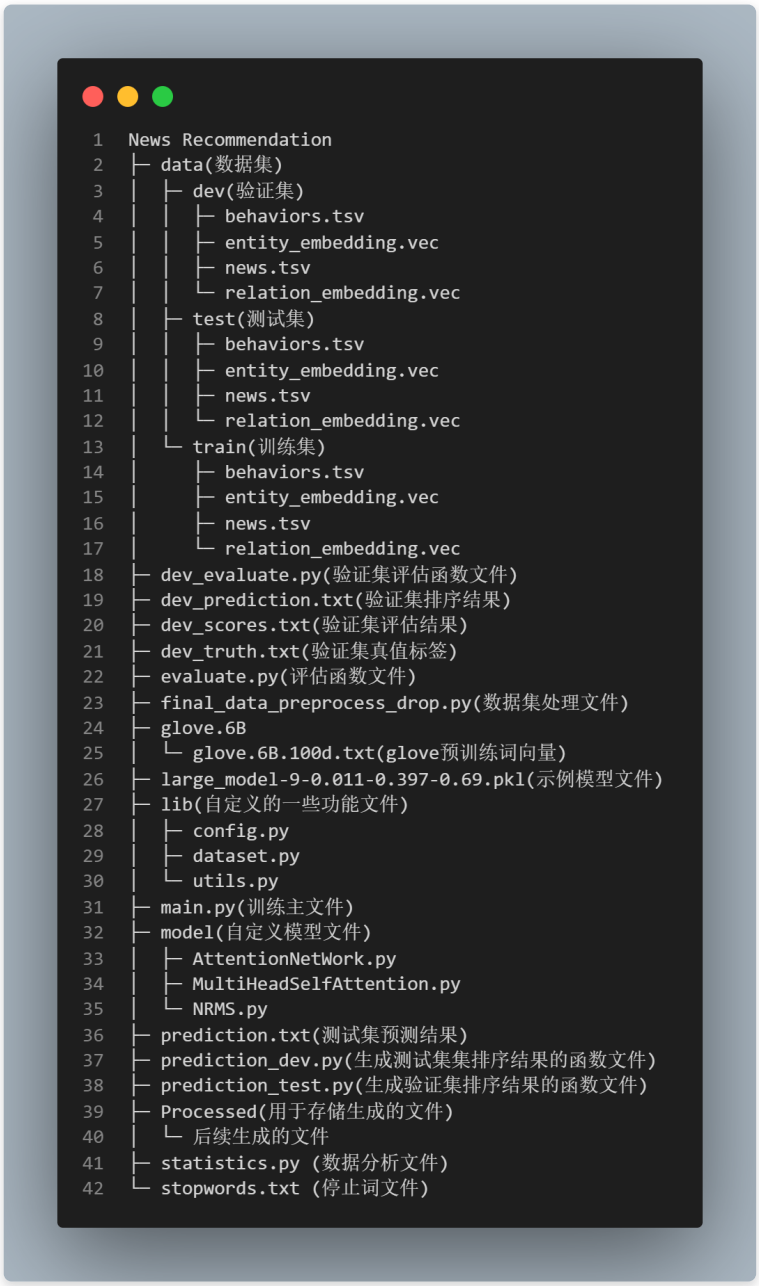


NRMS模型框架

在模型训练中采用负采样的方式，因此问题转化为一个多分类的问题，误差函数使用交叉熵函数。由于问题转化成了一个伪1+K的分类问题，因此K值的设定也是一个tradeoff的问题，实验中暂时取K = 4。

运行方式说明

- 创建**glove.6B, data, Processed**等文件夹，以及准备原始的数据集，使目录结构如下所示：



```
News Recommendation
├─ data(数据集)
│  ├─ dev(验证集)
│  │  ├─ behaviors.tsv
│  │  ├─ entity_embedding.vec
│  │  ├─ news.tsv
│  │  └─ relation_embedding.vec
│  ├─ test(测试集)
│  │  ├─ behaviors.tsv
│  │  ├─ entity_embedding.vec
│  │  ├─ news.tsv
│  │  └─ relation_embedding.vec
│  └─ train(训练集)
│     ├─ behaviors.tsv
│     ├─ entity_embedding.vec
│     ├─ news.tsv
│     └─ relation_embedding.vec
├─ dev_evaluate.py(验证集评估函数文件)
├─ dev_prediction.txt(验证集排序结果)
├─ dev_scores.txt(验证集评估结果)
├─ dev_truth.txt(验证集真值标签)
├─ evaluate.py(评估函数文件)
├─ final_data_preprocess_drop.py(数据集处理文件)
├─ glove.6B
│  └─ glove.6B.100d.txt(glove预训练词向量)
├─ large_model-9-0.011-0.397-0.69.pkl(示例模型文件)
├─ lib(自定义的一些功能文件)
│  ├─ config.py
│  ├─ dataset.py
│  └─ utils.py
├─ main.py(训练主文件)
├─ model(自定义模型文件)
│  ├─ AttentionNetWork.py
│  ├─ MultiHeadSelfAttention.py
│  └─ NRMS.py
├─ prediction.txt(测试集预测结果)
├─ prediction_dev.py(生成测试集集排序结果的函数文件)
├─ prediction_test.py(生成验证集排序结果的函数文件)
├─ Processed(用于存储生成的文件)
│  └─ 后续生成的文件
├─ statistics.py (数据分析文件)
└─ stopwords.txt (停止词文件)
```

文件结构

- 运行**final_data_preprocess_drop.py**，处理数据集，结果保存在**Processed**文件夹中
- 运行**main.py**加载处理好的数据集并训练
- 运行**prediction_dev.py**生成对验证集的排序结果
- 运行**prediction_test.py**生成对测试集的排序结果
- 运行**dev_evaluate.py**或者**evaluate**可对排序结果进行auc等指标的评估
- 运行**statistics.py**用于显示训练集的一部分统计结果

实验设置

```python
class BaseConfig():
    device = 0
    num_epochs = 20
    batch_size = 128
    num_workers = 4
    learning_rate = 0.001 # 0.0002
    dropout = 0.2

    pad = '<pad>'
    unk = '<unk>'
    pad_idx = 0
    unk_idx = 1
    negative_sampling_ratio = 4 # K = 4
    num_words = 57478 # 41342 small 57478 large
    num_words_title = 15 # 48
    num_words_history = 20 # 50

    entity_embedding_dim = 100 # 200
    word_embedding_dim =  100
# 300 Glove embedding
    category_vec_dim = 100
    query_vector_dim = 200

class NRMSConfig(BaseConfig):
    dataset_attributes = {
        "news": ['category', 'subcategory', 'title', 'abstract'],
        "record": []
    }
    num_attention_heads = 20

nrms = NRMSConfig()
```

实验设置情况

主要参数简单说明：

- negative_sampling_ratio: 实验中的K值
- num_words_title: 标题最大的词汇量
- num_words_history: 历史点击新闻的最大个数
- 其余参数不难理解，不做赘述。

验证集排序结果

验证集排序结果，见**dev_score.txt**

```
AUC:0.6782
MRR:0.3278
nDCG@5:0.3616
nDCG@10:0.4254
```

验证集评估结果

提交内容

提交文件目录

- 项目代码(原始数据集，以及处理好的数据集，模型等见后续链接)
- 验证集排序结果及评估文件**(prediction.txt，dev_score.txt)**
- 测试集排序结果**(prediction.txt)**

项目其他资源链接

为方便后续的审核，全部的代码，模型，处理好的数据集等可在下面的链接中获取。

百度网盘链接：https://pan.baidu.com/s/1ZQMgOXzGPXd3Ih7lJquGRw

提取码：sdxx

Reference : Neural News Recommendation with Multi-Head Self-Attention