深度学习实验五

任务背景：

本次实验的主要内容是新闻推荐任务：

首先， 新闻的时效性强 。新闻平台每天都会发布大量新的新闻，已经发布的新闻会被最新发的内容迅速覆盖、随后在页面上“消失”。这就导致有些新闻可能刚发布就被淹没，没有人发现，也没有人阅读。

其次，新闻文章的标题和正文往往都包含了丰富的内容。因此，新闻推荐技术的一大关键在于使用NLP技术从新闻的标题与正文中掌握并理解新闻内容在传递什么信息。

此外，新闻平台上往往不会显示用户的文章阅读排序。新闻推荐系统需要从用户的浏览与点击行为上掌握用户的阅读喜好，并对用户的阅读兴趣进行建模。然而， 对用户的阅读兴趣进行建模是很难的 ，因为用户的阅读兴趣通常比较广泛，并随着时间的推移而变化

任务内容：

本次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文章进行排序。在这个过程中，新闻文章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。

数据集：

小数据集是为了同学们更好更快的理解使用数据，载入模型，但是最终的结果是依据大数据集上的。当然也可以直接跳过小数据集。

- behaviors.tsv 用户的点击历史和impression记录（content为样例）

behaviors.tsv(共有5列)

Column	Description	Content
Impression ID	impression的ID	123
User ID	user ID	U131
Time	impression的时间 “MM/DD/YYYY HH:MM:SS AM/PM”	11/13/2019 8:36:57 AM
History	用户之前的历史记录	N11 N21 N103
Impressions	用户的点击历史的记录	N4-1 N34-1 N156-0 N207-0 N198-0

- news.tsv 新闻文件的信息（content为样例）

news.tsv

Column	Description	Content
News ID		N37378
Category		sports
SubCategory		golf
Title		PGA Tour winners
Abstract		A gallery of recent winners on the PGA Tour.
URL		https://www.msn.com/en-us/sports/golf/pga-tour-winners/ss-AAjnQjj?ocid=chopendata
Title Entities	包含在这条新闻的标题中的实体	[{"Label": "PGA Tour", "Type": "O", "WikidataId": "Q910409", "Confidence": 1.0, "OccurrenceOffsets": [0], "SurfaceForms": ["PGA Tour"]}]
Abstract Entites	这则新闻摘要中包含的主题	[{"Label": "PGA Tour", "Type": "O", "WikidataId": "Q910409", "Confidence": 1.0, "OccurrenceOffsets": [35], "SurfaceForms": ["PGA Tour"]}]

- entity_embedding.vec 从知识图中提取实体在新闻中的embedding（conten为样例）

the 100-dimensional embeddings of the entities and relations learned from the subgraph (from WikiData knowledge graph) by TransE method.

ID	Embedding Values
Q42306013	0.014516 -0.106958 0.024590 ... -0.080382

提交结果：

希望同学们仔细阅读result的生成方法避免出现差错！！！

同学们需要提交模型生成的每个impression中的新闻排名结果。结果生成一个名为prediction.txt的文件。在这个文件中，每一行包含一个impression的ID和候选新闻的排名列表。每行格式为：

ImpressionID [Rank-of-News1,Rank-of-News2,...,Rank-of-NewsN]

比如，给出一个

ImpressionID	Candidate News
24481	N125045 N87192 N73556 N20417

这种impression的预测结果可以是:

24481 [4,1,3,2]

这意味着该impression中候选新闻文章的排名顺序为N87192、N20417、N73556、N125045。

最终的提交文件将包括prediction.txt文件，代码文件，不需要提供数据，还有实验报告，实验报告中需要注明在整个大数据集上面的AUC结果，详细的测试数据和结果将在12.8日公布，该实验时长为3周，截至12.30日23.59。

测试代码同evaluate.py文件，如果不放心自己生成的prediction结果是否符合格式要求，大家可以根据evaluate文件自己测试。

给分标准

本次实验需要使用pytorch，可以自由调库。

实验的完成度以及实验报告部分【实验报告中需要包含自己的数据分析，建模过程等】占比75%，

auc的结果好坏占比25%，由于需要差异化分数，所以该部分需要进行一个rank，对所有同学的auc结果进行比较之后打分。