课程大作业

大作业概况

《大数据分析（B)》课程大作业采用线下实践和实验课辅导结合的形式进行，重要时间节点如下：

（1）第 06 周实验课：实验环境配置辅导；

（2）第 11 周实验课：大作业中期答疑；

（3）第 15 周汇报课：大作业亮点 PPT（spotlight PPT）汇报；

（4）第 17 周考试周：大作业实验报告及相关实验代码等最终版本提交。

任务介绍

课程大作业旨在结合课程所学相关内容（如假设检验、回归分析、方差分析、贝叶斯学习、矩阵分解、数据挖掘、深度学习等），将所学内容付诸实践，同时锻炼大家阅读论文、对论文进行文献综述和算法总结、复现相关代码的能力，另外也希望通过本次大作业让大家熟悉神经网络机器翻译、推荐系统等基于深度学习方法的实现，了解深度学习环境的搭建、模型的构建、训练、测试等过程，从而为今后的学习科研等奠定一定的基础。

大作业内容包括两个选题：神经网络机器翻译（Neural Machine Translation，NMT)、推荐系统（Recommendation System)，大家可以结合自己的背景和对相关内容的了解，从两个选题中选择一个完成（**二选一**)。

选题一：神经网络机器翻译（NMT）

（1）阅读所附的 3 篇 NMT 相关的论文，进行文献综述和算法总结；

（2）根据“Effective Approaches to Attention-based Neural Machine Translation”论文（下文称之为：要复现的论文)，进行实验复现或部分复现；

（3）结合 TensorFlow 中有关 NMT 的教程，复现相关工作及结果。参考链接：
https://github.com/tensorflow/nmt/tree/tf-1.4

（4）如果自己比较熟悉 Pytorch，也可以参考以下链接进行实验复现。参考链接：
https://github.com/AotY/Pytorch-NMT

选题二：推荐系统（Recommendation）

（1）阅读所附的 3 篇推荐系统相关的论文，进行文献综述和算法总结；

（2）根据“xDeepFM: Combining Explicit and Implicit Feature Interactions for Recommender Systems”论文（下文称之为：要复现的论文)，进行实验复现或部分复现；

（3）结合 Tensorflow 中有关 xDeepFM 的教程，复现相关工作及结果。参考链接：
https://github.com/microsoft/recommenders/blob/master/examples/00_quick_start/xdeepfm_criteo.ipynb

具体要求

请仔细阅读下述大作业的具体要求，并遵照要求完成大作业。

（1）阅读所提供相关论文，进行文献综述。

【基本要求】15 分

重点阅读“要复现的论文”，关注其中引言和相关背景的介绍部分，并结合大作业所提供的其他论文对相关领域（NMT 或者 Recommendation）的主流方法进行文献综述。

在进行文献综述时，**可以以时间为序**，根据相关领域的发展脉络，介绍各个时期的典型文献以及他们所提的主要方法的概述（基本原理、优缺点等）；**也可以以类别为纲**，结合相关领域的不同典型方法，分别介绍相关方法的基本原理、代表性文献、方法的优缺点等。

【加分项】≤10 分

结合相关领域的发展趋势，对“要复现的论文”之后发表的论文（比如 2017 年以后）的典型方法进行文献综述。

（2）根据所阅读的相关论文进行算法总结。

【基本要求】15 分

（1）对相关任务（NMT 或者 Recommendation）所要解决的问题给出形式化的定义。比如通过形式化的数学方法给出相关任务的输入、输出的定义，相关任务的概率函数表达；假设使用神经网络解决该问题，神经网络建模函数的意义、神经网络的训练目标（损失函数）的定义等。

（2）对“要复现的论文”中所给出的算法或模型进行归纳总结，分别介绍相应的算法或模型的主要模块，特别是要给出相应模块的功能、原理、方法等。

【加分项】≤10 分（加分项可累加：只完成一项、完成全部两项、部分完成均有加分）

（1）结合相关领域的发展趋势，对“要复现的论文”之后发表的论文的典型算法或模型进行总结，并介绍相应的算法或模型的主要模块。

（2）算法总结能体现相关领域的发展脉络和趋势：能指出相关算法提出时拟解决的问题（也即提出的动机 Motivation），能指出相关算法的局限或者不足（也即能指出问题 Problem），能指出后续算法对前续算法的改进等。

（3）熟悉有关编程平台与深度学习框架，根据指定论文复现对应方法。

【基本要求】30 分

（1）根据所选择的深度学习框架和平台，对“要复现的论文”中的方法进行复现。

（2）可以使用网上已有的开源代码，也可以自己对相关代码进行整合改进。

（3）无论是开源代码、还是整合改进，均必须对所写（所使用）的代码的关键功能模

块使用中文进行“详细”的注释，解释该模块的功能、基本方法等；

（4）在实验报告的“复现流程”一节，详细介绍实验过程，包括所用的数据、数据的预处理、模型的实现、模型的参数说明、模型的训练过程（包括训练所用的参数）、模型的测试过程（包括测试所用的参数）等。同时要给出各流程对应的代码文件名及函数名。

【加分项】≤20 分（加分项可累加：只完成一项、完成全部两项、部分完成均有加分）

（1）给出模型训练的可视化结果并对其进行分析。

（2）结合“要复现的论文”之后提出的新算法或模型，对其加以复现，并在实验报告“复现流程”一节中整合本部分复现过程。

（4）给出自己复现方法所得的实验结果及实验分析。

【基本要求】15 分

（1）明确相关任务（NMT 或者 Recommendation）的评价指标（Evaluation Metrics），在实验报告中给出该评价指标，并对其进行解释说明。

（2）基于“要复现的论文”中的数据集，对上述复现的算法或模型进行实验，给出复现实验结果。

（3）将复现实验结果与“要复现的论文”中给出的结果进行比较，说明是否存在差异，并给出实验结果的分析；假设复现结果与论文中的结果存在差异，对可能的原因进行分析。

【加分项】≤10 分

对“要复现的论文”之后提出的新算法或模型的复现，并与上述基本要求中的结果一起，在实验报告中给出本加分项对应的实验结果和相应的分析。

（5）撰写实验报告以及汇报用亮点 PPT（spotlight PPT）。

【实验报告】10 分（评价实验报告撰写是否规范、内容是否全面丰富、逻辑是否清晰、重点是否突出）

（1）实验报告可以以中文撰写、也可以以英文撰写。要求重点突出、逻辑清晰。

（2）实验报告的格式参考正式的 paper，建议包括：

(a) 报告题目：神经网络机器翻译、推荐系统（二选一）
(b) 个人信息：包括姓名及学号；具体专业方向（不能只是电子信息）；电子邮箱
(c) 中文摘要及关键词
(d) 英文摘要及关键词
(e) 引言
(f) 1. 文献综述（其中 1 为建议编号，下同）（可细分为子章节，下同）
(g) 2. 算法总结
(h) 3. 评价指标
(i) 4. 数据集
(j) 5. 复现流程
(k) 6. 实验结果及分析

(l) 7. 结论
(m) 8. 所完成的加分项
（以表格方式给出所完成的加分项，并给出实验报告中的对应子章节索引）
(n) 9. 心得体会
(o) 参考文献
(p) 附录：给出包括实验报告在内的大作业相关文件清单及相应说明
（即上传到网络学堂的文件内容；代码可放于一个目录，并对该目录作说明）

（3）实验报告的表格、图片等要给出相应的表题、图题，并顺序编号，并在正文中相应地方给出引用；参考文献应在正文中给出相应的引用。

【亮点 PPT（spotlight PPT）】15 分（评价 PPT 的美观度、工作亮点总结、口语表达与汇报展示的效果、时间掌控的情况等）

（1）请大家准备 2 页的 PPT，第一页为报告题目、姓名、学号、专业方向、电子邮箱等；第二页介绍大作业的工作亮点。

（2）工作亮点（PPT 第二页）：除了基本算法/模型的介绍之外，应突出自己工作的亮点部分：可以是模型的亮点、实验结果的亮点、除了基本要求之外完成的加分项的亮点、甚至是实验报告撰写的亮点、实验结果呈现形式的亮点、心得体会的亮点等等，总之能够凸显自己工作特色的所有东西都可以作为亮点给出来。

（3）完成大作业后，会花一次课程（拟安排在第 15 周）的时间，让大家在课堂上展示和介绍自己的亮点 PPT，**每个人严格控制为 1 分钟**（时间控制的情况也会体现在分数中）。

（4）亮点 PPT 需使用 pptx 准备，16:9 格式（不能用 pdf）。

（5）亮点 PPT 需在规定时间（具体时间请等待通知）之前上传到网络学堂，以便汇总后在课堂上依次展示，以节约时间。

（6）在截止日期前上传实验结果。

【每晚交 1 天（0-23 小时 59 分 59 秒算 1 天）扣 5 分】

将实验报告、汇报用亮点 PPT、所复现代码以及相关说明文件（如代码运行环境需求说明、代码运行方法说明等），打包成一个 zip 文件上传到网络学堂。

有关资源

相关数据集提供清华云盘下载地址：（访问密码：bigdata2021）

https://cloud.tsinghua.edu.cn/d/41b2f7a3fed34a3fadbf/

（1）NMT 子目录相关数据说明

nmt-tf-1.4.zip： 相关参考资料

iwslt15_en_vt： English-Vietnamese 小数据集

wmt16_de_en： English-German 大数据集

（2）Recommendation 子目录相关数据说明

xdeepfmresources.zip：　　推荐系统数据集

其他大家关心的问题

Q1：可以更换数据集么？

A1：可以使用其他数据集，但不建议这么做，因为会导致最后不好比较结果。

Q2：需要复现到和原论文一样的程度么？

A2：不需要。如果没有达到原论文的结果性能，也没有问题；但是最终成绩可能会参考这部分的情况，特别是大家一定要给出可能的原因分析。本大作业重点关注大家通过本次大作业能学到什么，而不是简单的一个实验结果。

Q3：代码会查重么？实验报告会查重么？

A3：代码不会查重，但是实验报告会查重。大家可以参考和下载已有的开源代码来完成大作业，但是一定要在实验报告中、说明文档中注明代码的来源。另外，必须要好好学习所下载的代码，要按照上述要求把整个实验过程和原理在实验报告中写清楚，不能只是跑了一个代码得到一个结果而已。

Q4：听说文献综述会影响实验报告的查重率？

A4：文献综述不是把别的论文中的内容简单拷贝粘贴过来，而是需要用自己的语言来对所阅读的论文进行总结，从论文所解决的问题、所提的方法、方法的优缺点等角度进行总结，具体要求见前述说明。通过自己的语言来进行文献综述，对实验报告查重率的影响应该可以控制得很小。

Q5：自己的电脑跑不起来论文中给定的数据集，课程会提供 GPU 资源么？或者换成小的数据集？

A5：由于学校没有为本课程配备服务器和 GPU 等计算资源，所以得同学们自己考虑去找计算资源。如果自己实验室没有 GPU 资源，大家可以考虑使用免费的 GPU 环境：google colaboratory（不过需要同学们自己去探索）；也有一些其他平台提供免费 GPU，大家可以去搜一下；另外，也可以问问身边的同学看能否帮忙。可以换小数据集（具体见 Q1）。