EE226 期末大作业选题1：学术网络的节点分类与链路预测

本选题在Kaggle平台上开展，需要在Kaggle上提交结果参与性能评测与排名，同时需要在Canvas上提交其他材料。项目详情、数据格式、评测方法等请在Kaggle竞赛页查看。

1. 项目描述

在此项目中，我们将会给出来自人工智能和数据挖掘领域的10个顶级会议（2016年-2019年）的42,614位作者和相应的24,251篇论文及其出版物的引用信息（无关引用已被剔除）。你需要对提供的数据进行处理，并建立学术网络，以下是两种建议的方法：

1）建立一个**同构**网络，其中每个节点代表一个作者，每个有向边代表两个相连作者之间的引用关系或共同作 者。 这将需要一些额外的数据处理，以将论文之间的关系转换为作者之间的关系。 2）构建一个**异构**网络，其中包含两种类型的节点，一种类型的节点代表作者，另一种类型的节点代表论文。 在该网络中，作者节点和论文节点之间的边表示该论文为对应的作者所著，而两个论文节点之间的每个有向边代表引用关系。

你可以选择任何一种方法来建立学术网络，甚至可以自主探索其他的方法。

此项目含有两个子问题：

问题一：作者节点的多标签分类。在网络中，每个作者节点都有单个或多个标签，而每个论文节点（如果出现在网络中的话）都有单个标签：如果他/她在会议A中发表了论文，则他/她将被标记为A。 如果他/她已在多个不同的会议上发表论文，则他/她将具有多个标签。 在这个问题中，我们提供20%的节点（包括作者节点和论文节点）的标签信息（论文对应的节点属于哪一会议）。 你的任务是预测其余80%**作者**节点的标签。

问题二：作者节点之间的链路预测。我们提供的信息是2016年-2019年的作者和论文信息，而一些活跃的作者在2020年也会有新的著作发表，新的合作或引用关系就会在已建立的学术网络中的作者节点之间产生新边。你的任务是根据已提供的信息，对测试集中的每一对作者节点进行预测， 如果在2020年这两个给定的节点间将连边，则将其标记为1，否则将其标记为0。

2. 性能指标

对于两个问题，我们均将测试集分为公开集和私有集两部分，数据量各占50%，在Kaggle竞赛截止前，你的所有提交结果均只显示在公开集上的性能，你可以据此选择3个结果参与私有集的测试。我们最终将根据私有集上的性能进行评判，不过性能的略微波动不会影响评判结果和最终成绩。

问题一：Mean F1-score

$$[F1 = 2\frac{p \cdot r}{p + r} \quad \text{where} \quad p = \frac{tp}{tp + fp}, \quad r = \frac{tp}{tp + fn}]$$

其中p即准确率，r即召回率。

问题二：AUC

3. 文件提交

选择本项目后，需要加入对应的Kaggle Inclass Competition（两个问题都要加入），并且除了在Kaggle上进行编程和提交结果外，还需要注意以下时间点：

1) **4.14（第八周周三）**前，确定分组和选题，如对项目有问题，可在此时间段内反馈。

2) **4.28 （第十周周三）**前，提交中期报告，包含introduction, related work, research plan, expected outcome, options.

3) **6.9（第十六周周三）**前，提交以下材料：

- [ ] 完整的程序代码和对应的readme文档，注明使用的python包和运行命令（Notebook没有运行命令则可以不写），如果是非Kaggle环境下完成，请注明python版本和各包的版本。
- [ ] 英文项目报告，以论文形式撰写，包含introduction， related work， problem definition，algorithm， results， conclusion。

4. 参考文献

本项目推荐使用图神经网络（Graph Neural Networks, GNNs）进行网络嵌入（Network Embedding），并使用节点的嵌入向量进行节点分类和链路预测。参考文献如下：

- 所有拓扑网络（通常是同构的）：DeepWalk(KDD2014), node2vec(KDD2016), SDNE(KDD2016), GAT(ICLR2018), DGI(ICLR2019).
- 异构网络: metapath2vec(KDD2017).