数据集

本次作业提取了 169,343 篇 arXiv 上计算机科学相关的论文，这些论文涉及到 40 个子领域。每篇论文提供了标题与摘要，以及一个 128 维的特征向量表示，该向量通过对标题和摘要中所有单词的 skip-gram embedding 进行平均而获得。我们也提供了论文间的引用关系，论文发表的年份。数据集中的论文按年份分为训练集、验证集和测试集：2017年及其之前的论文作为**训练集**，2018年的论文作为**验证集**，2019年及其之后的论文作为**测试集**。我们提供了训练集和验证集论文的标签，但未提供**测试集**的标签（需要自己实现类似于作业二中的分类算法来预测）。

如何读取

```
# Read abstract, category, year of each paper
papers = pd.read_csv(f'./dataset/papers.csv.gz', compression='gzip')

# Read the embedding vector of each paper
feats = pd.read_csv(f'./dataset/feats.csv.gz', compression='gzip',
header=None).values.astype(np.float32)

# Read the citation relations between papers
edges = pd.read_csv(f'./dataset/edges.csv.gz', compression='gzip',
header=None).values.T.astype(np.int32)
citer, citee = edges
```