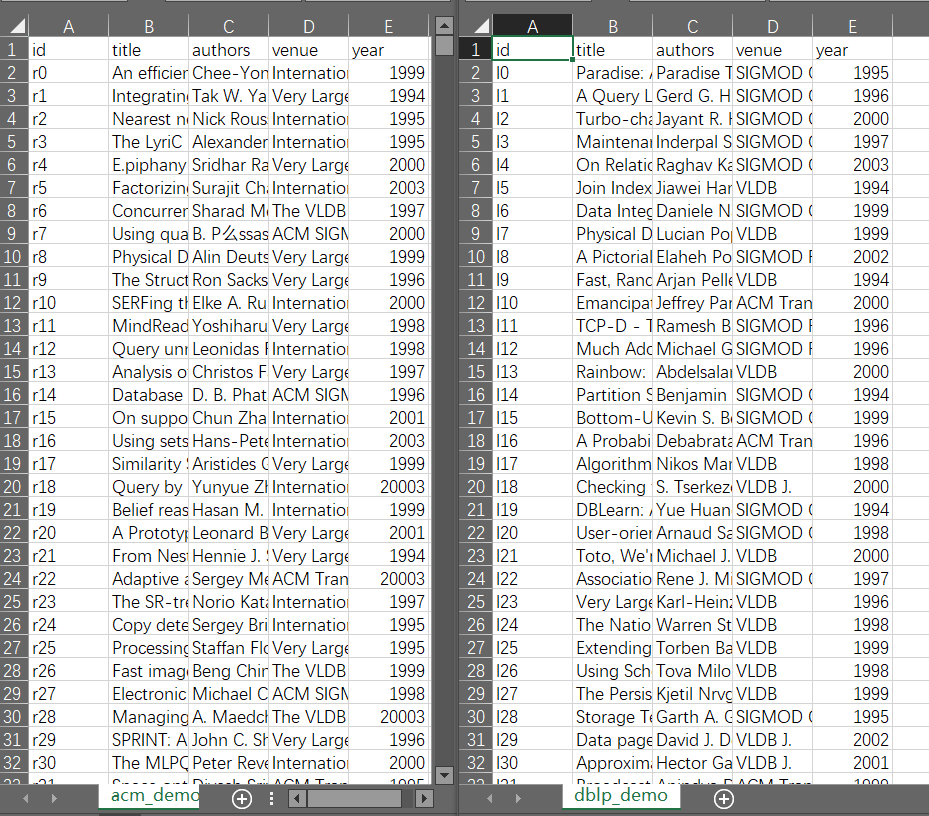
实体识别部分

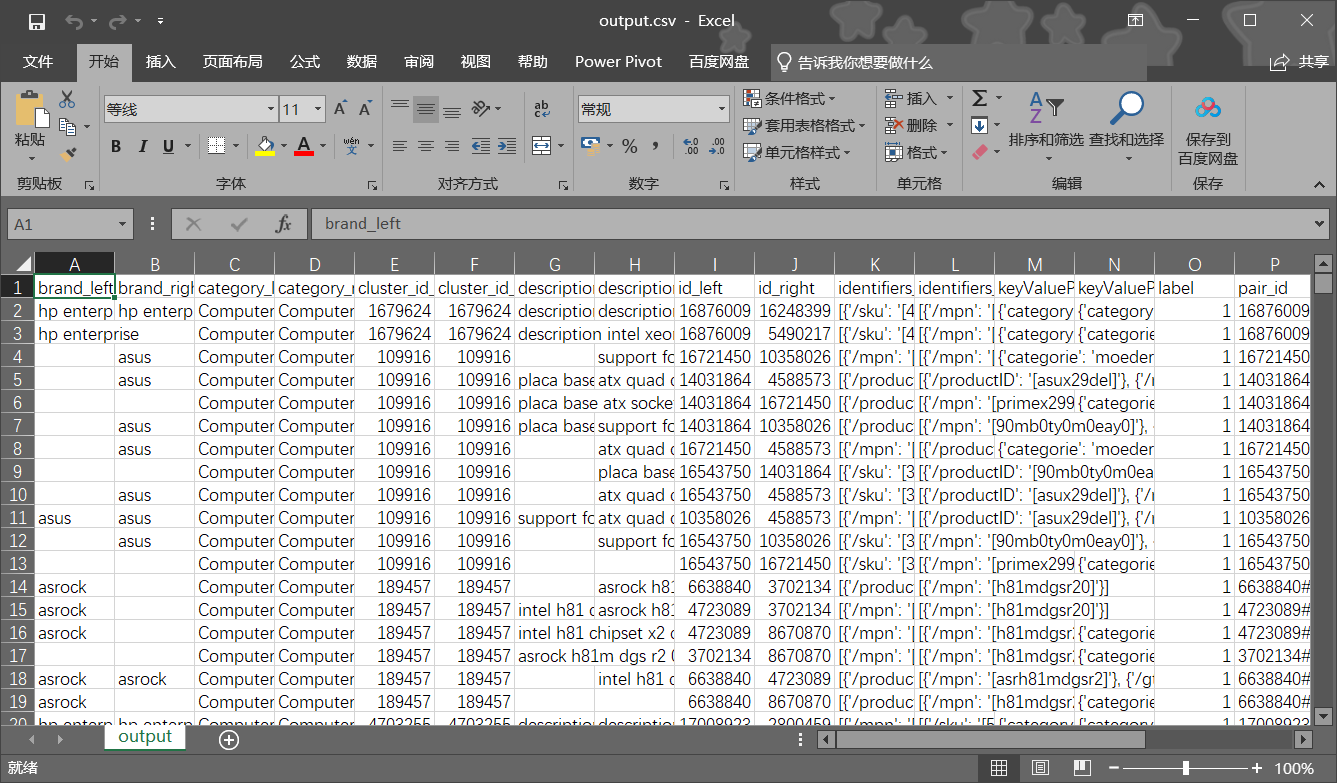
数据输入格式：

本模块针对关系型表格数据完成实体识别任务。输入数据要求包含两个待实体识别的表格信息，对于输入数据的模式不做具体要求。可以通过给出两个待匹配的表格，其中第一行为属性名。参考dblp\_demo和acm\_demo数据集，如图一所示：



图一：待匹配的两个表格数据

也可以通过在属性名前加ltable和rtable前缀的方式，分别表示左右表格的属性信息。可以参考computer数据集，如图二所示。



数据输出格式：

数据输出文件包括了两个比较的实体id以及是否为同一实体的标识label，表示为

初步设计方案：

本部分主要分为以下四个部分进行介绍：（1）模式匹配模块（2）数据分块模块（3）实体对的特征工程（4）实体匹配分类器设计。

1. 实体匹配模块

由于输入数据可能存在模式不同问题，本模块旨在统一输入数据的模式，便于后续数据处理。首先由用户给出全局数据模式，然后将输入的表格数据映射到该全局模式上。对于输入数据中的任意一个属性和一个待映射的属性,设计属性间的相似度度量，然后在全局数据模式中选择合适的属性进行映射，最后获得符合用户要求模式的关系型表格数据。

1. 数据分块模块

本模块旨在减少待匹配的实体对数量，减少匹配时的搜索空间，加快实体匹配效率。目前考虑采用lsh技术，获得相似度较高的元组对，并将获得的相似元组置于同一blocking中。在后续的匹配过程中，只需匹配blocking内部的实体对。

1. 实体对的特征工程

由于提供给下层实体匹配分类器的输入数据是数字类型向量，而非表格数据中的字符串序嗯Word文档列。因此本模块需要将字符串类型的实体对数据转换为数字向量，同时保持实体对相似度信息。本模块拟定构建和属性类型相关的相似度函数库，通过计算实体对属性间的相似度，构建特征向量等，进行实体对的特征工程构建。

1. 实体匹配分类器

实体匹配本质为二分类问题，对于给定的实体对判断匹配/不匹配，因此底层需要设计该任务的分类器。为了保证整体系统的可解释性，该模块拟定采用具有可解释性的随机森林分类器，并采用模型剪枝，优化搜索空间的方式提升时空可用性