1.机构名称相似度

2.机构关键词相似度

3.机构社交关系相似度

1. 需要机构的名称，取就可以

功能：计算相似度

1. 获取机构的关键词信息，将关键词信息表达成一个向量

功能1：获取作者的关键词信息，获取机构的关键词信息

功能2：将关键词信息提取为一个有效区分机构的向量

功能3：计算相似度

1. 获取机构的合作机构，所属作者，找相似性

功能1：从数据库中获取机构的合作机构

功能2：从数据库中获取机构的作者

功能3：计算相似度

54所 陈享远

中国电子科技集团公司第五十四研究所

庄电子部第54研究所

机电部石家庄54所

电子工业部54研究所

电子部54研究所

电子部石家庄54研究所

如何表征 科研人员/科研机构 的研究领域：

是否可以把所有的论文标题，摘要凑起来去训练主题模型，然后拿该研究员的论文之类信息去得到主题分布，判断相似度 可以试试，应该比较好实现，就是数据量的问题

数据爬取下来，按每个机构各种名称list 进行爬取，爬下来的数据有一些问题：1.爬下来的并不是像namelist列的那样规整 2.一些单位地址解析下来为空 3.一些记录中可能根本没有包含目标机构。

工程上：现在可以现将数据解析为每个论文中第一作者的单位，对应的论文信息，其他机构为该机构的合作单位，论文作者为该单位的员工，论文摘要和标题为该机构的文本内容