keywords: 根据生成 keywords；具体定义的规则见文件中所述，规则可作适当调整；输出 cate key subkey （它们的等级由高到低）

cleanTitle: 去除 title（title 已分词）中首尾出现的停词、多余空格并全部转为小写

combine_features: 得到每个 paper 的 words 集合，集合中元素包括：title keyword 按照上述处理过程之后划分而成的短语、单词；不同级别的词频按照级别赋予权重 weight（权重可调节，视输出效果而定）

Orgnames: 得到 Organization 的名称关键词，对于 journal 和 conference 两个部分包括 fullname shortname（对 fullname 进行词干化处理，设置 stem=T, 而 shortname 不做修改设置 stem=F）；对于 affiliation 的名称处理同时去除 university school lab department institution 这些机构名称指示的词汇（实际采用的是正则表达式的模糊匹配，具体请看代码）

comOrg: 这里的 Organization 指的是 journal 和 conference，这一部分实现的功能是对于每个 organization，得到它包括的所有 paper 的 words 组合；同时加上它们名字的关键词；即组合 combine_features 和 Orgnames 的输出结果；

get.sparsemat: 生成对应于 doc-term 的稀疏矩阵（slam 包）【其中 doc 的定义可以是 paper、journal 都可以】

topicmodels: 生成 topicmodels 包的 doc-term 矩阵 dtm，将出现频率过高的 term（我发现有"journal" "study" "analysis"这种无意义词）去除（这里我用了 tf 准则，原因是我觉得 tfidf 的性能在这里表现不好，可以自行实验）；

dist_dtm: 表示 doc 之间的 cosine 距离矩阵，用 dissimilarity 函数可生成

选 topic 个数：用 5 折交叉验证找 perplexity 的最小点处以及 loglikelihood 的最大点平衡来选择 topic 个数，大约在 20-40 范围内；

得到 topic 之后可以训练模型得到每个 doc 的 topic；用 posterior 函数输出每个 doc 对应的 topic 分布和每个 topic 对应的 term 分布；进一步可以算出每个 doc 之间按照 topic 定义的 distance（dist_dto）

train:

trains 数据集是按照组合 confirmed 和 deleted 两个在 sample 中同时出现的案例组成的（第一句代码）

get.author.paper：对于 train 的每个 author 找 paperauthor 数据集中的 paper 集合，找到对应 paper 的 words 组合作为对应 author 的 words 组合

modify.aut.papers: 由于 paperauthor 有噪音（只能认为它有一定的真实性），因此生成的 author keywords 组合鱼龙混杂；这一部分拟通过对于不同的 paper 赋予不同的权重来找到相对接近真实性的 author keywords 组合，我的想法是找到这些 paper 的在 paperauthor 的数据集中连接的 author 们，然后找当前 author 与这些 author 们的 coauthor 数目，在一定程度上可以反映当前 paper 的重要程度；目前由于复杂度等一些问题未完工，待续

valid.paper.com: 这一部分找 train 中对应的 paper 关键词；我按照权重也纳入了 paper 隶属的 journal 和 conference 的关键词（即 comOrg 的输出）

train1: train1 是原训练集加入按照以上定义的关键词和文档（doc 在这里就是 train 里的每一个 author 和每一个 paper）生成的距离；

测试：按照 logistic 回归，我现在用的是 auc 值；我的结果显示（仅作比较）：

（1） 纳入 journal 和 conference 关键词之后的 paper 关键词预测效果较好

（2） 用 doc-topic 生成的距离比用 doc-term 直接生成的距离表现较逊

有说明不清楚的地方再联系我~

我认为可以改进的地方：

（1） 了解数据性质

（2） 利用图的特征

（3） 探索去噪音的可行途径

（4） 关键词的划分

（5） 权重设计