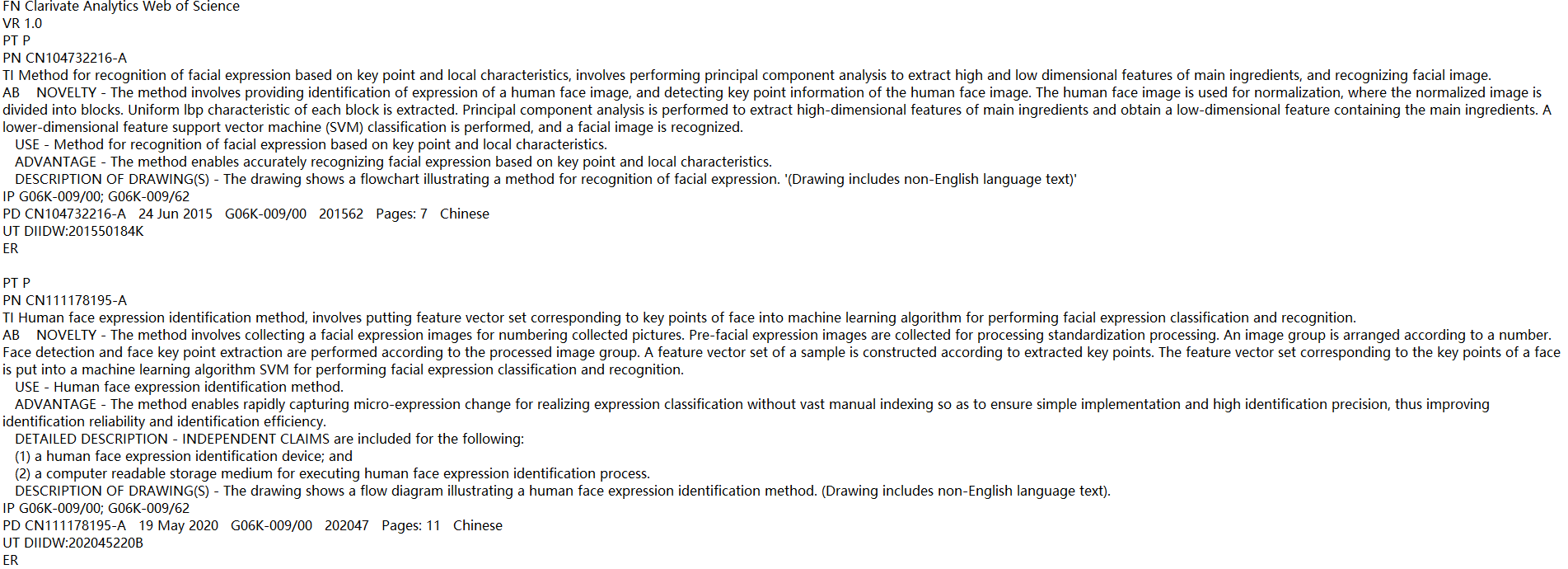
**第一步：数据获取与处理**



分为两个处理程序，第一个科研训练数据获取.py把初始的从德文特数据库中导出的专利信息作处理并存成一个标准格式的数据格式。

输入的数据格式：



它是个txt文件

输出的数据格式：



获取数据处理.py我是用来把每个专利里面的摘要没用的文字给去掉了，然后只保留了专利号、文本、ipc分类号、时间这四个信息。其实你完全可以第一步做完直接把这个表格数据存到数据库中去用，我是当时写好了就懒得改了。

最后的格式：



**数据分析：**

我数据导入数据库是手动导入的，没写代码。

这里放一下数据库设计格式：



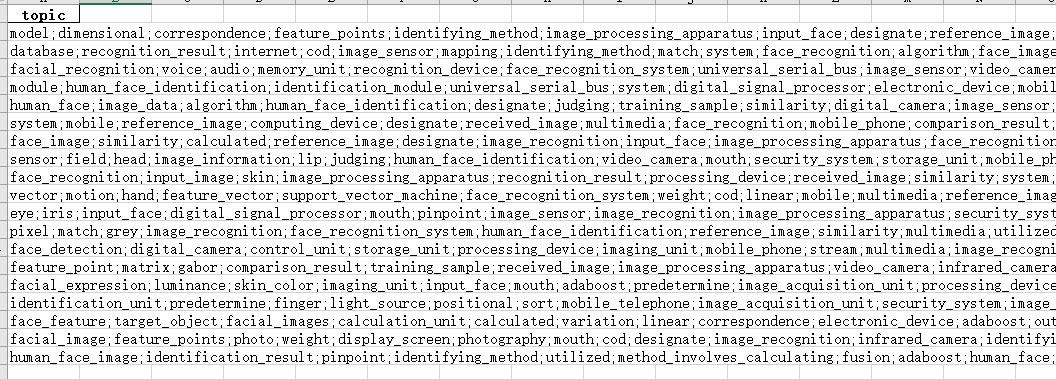
然后就是几个重要的数据分析的算法介绍：



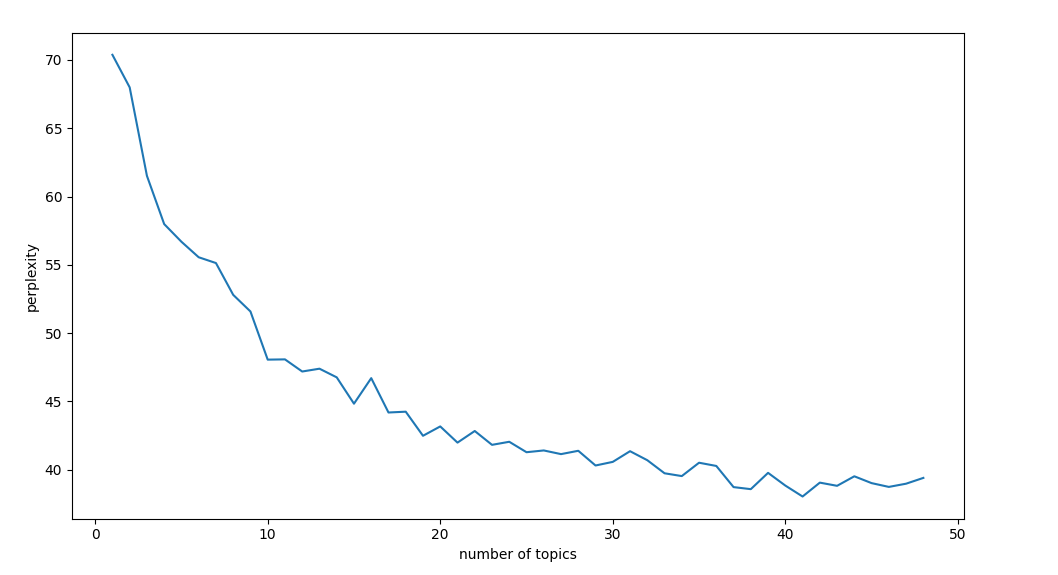
显而易见这个代码的作用就是对专利文本摘要作主题模型分析并得出每个主题对应的主题词和相应的困惑度曲线的。

主要的流程步骤：先划分时间窗，从数据库中选取自定义的时间窗口，然后利用LDA模型对这个时间窗里面的所有专利摘要文本进行主题模型分析，注意：这里的时间窗选定后专利文本数量就一定了，需要自定义只有主题数。最后输出对应主题词和困惑度曲线。

主题及其主题词：一行是一个主题，



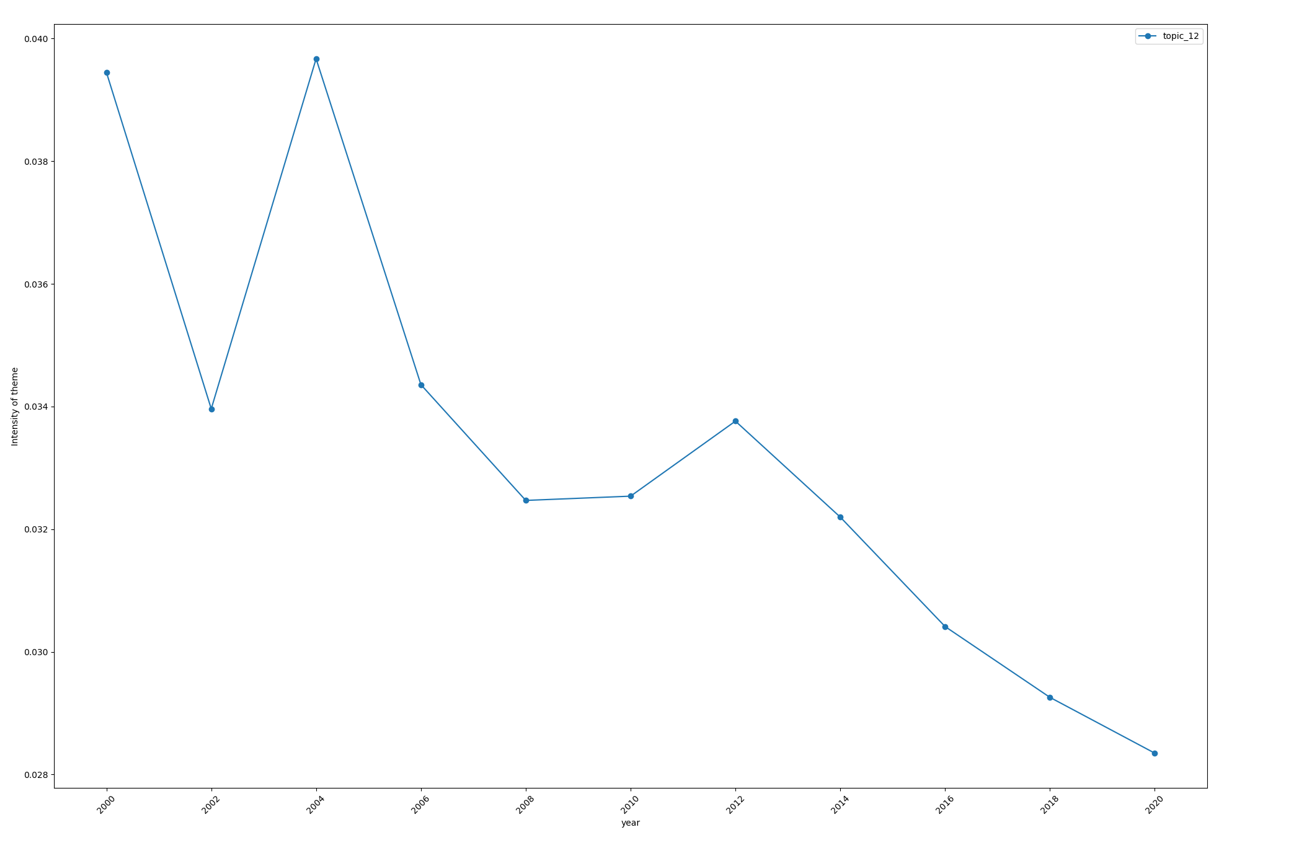
困惑度曲线：





计算各个主题强度随着时间变化而变化的值并画出折线图。

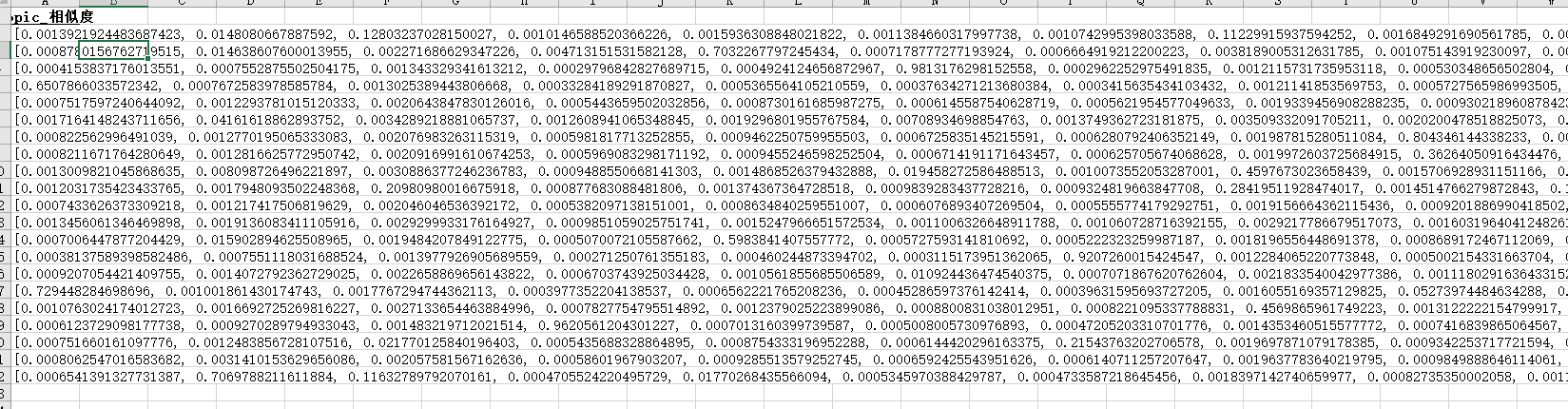
折线图：





计算两个时间窗下的主题余弦相似度并输出

输出格式：

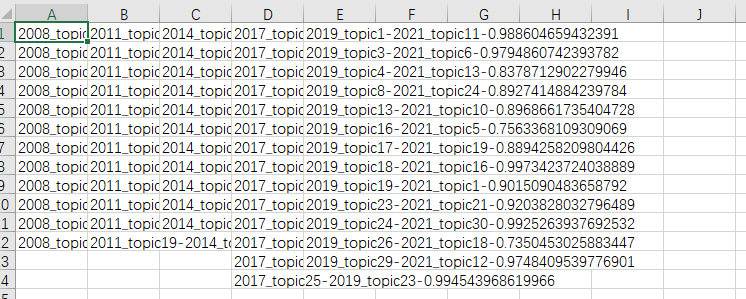


N\*M的矩阵，我这里用excel存的，N表示第一个时间窗主题数，M表示下一个时间窗主题数



通过关联过滤规则找出有效关联主体对，具体流程可见论文。

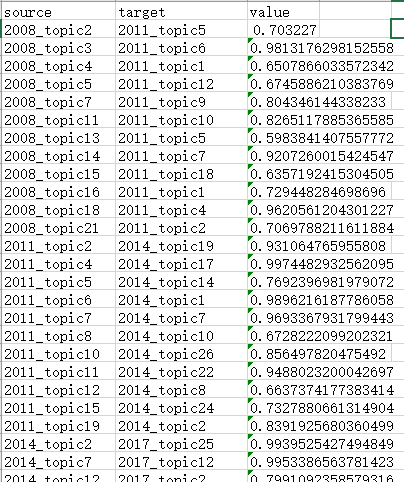
最后输出格式：





根据这些有效节点画图

BDP数据格式：



对流程或者方法搞不明白可以看一下我的论文，里面有详细说明过程。