**主题模型初步分析结果**

**语言：**python3.7

**调用包：**Gensim

**数据：**ABCDI爬虫数据

大数据：5000条文本

区块链：490条文本

人工智能：70条文本

物联网：3480条文本

云计算：3480条文本

**处理：**

1、读入一个方面的全部文本，对每个文本内容进行分词，去掉停用词（停用词表是哈工大学、百度、四川大学的停用词表以及一些我们后来根据输出结果自己制定的符号的集合）

2、文本内容处理后构建字典，再构建词袋，调用gensim.models.LdaModel，设置训练出10个的主题，学习率为50。

**结果：（以大数据结果为例）**

大数据相关

0.031\*"数据" + 0.024\*"日志" + 0.018\*"中" + 0.016\*"查询" + 0.010\*"表" + 0.010\*"配置" + 0.010\*"SQL" + 0.010\*"执行" + 0.009\*"点击" + 0.009\*"操作"

0.014\*"业务" + 0.013\*"蚂蚁" + 0.013\*"数据" + 0.012\*"服务" + 0.010\*"金融" + 0.010\*"技术" + 0.008\*"金服" + 0.008\*"能力" + 0.007\*"提供" + 0.007\*"中"

0.011\*"数据" + 0.010\*"OSS" + 0.009\*"类型" + 0.008\*"oss" + 0.008\*"中" + 0.007\*"文件" + 0.007\*"命令" + 0.006\*"返回" + 0.006\*"支持" + 0.006\*"gt"

0.064\*"实例" + 0.027\*"数据库" + 0.017\*"数据" + 0.014\*"同步" + 0.014\*"创建" + 0.014\*"云" + 0.013\*"选择" + 0.013\*"支持" + 0.012\*"RDS" + 0.012\*"中"

0.087\*"nbsp" + 0.007\*"mysql" + 0.007\*"中" + 0.005\*"log" + 0.005\*"Kubernetes" + 0.005\*"aliyun" + 0.005\*"执行" + 0.005\*"─" + 0.005\*"SOFATracer" + 0.004\*"size"

0.069\*"gt" + 0.062\*"lt" + 0.012\*"参数" + 0.007\*"类型" + 0.007\*"返回" + 0.007\*"key" + 0.006\*"amp" + 0.006\*"示例" + 0.005\*"id" + 0.005\*"请求"

0.019\*"服务器" + 0.017\*"配置" + 0.014\*"文件" + 0.013\*"服务" + 0.012\*"云" + 0.010\*"安装" + 0.010\*"系统" + 0.010\*"请" + 0.009\*"阿里" + 0.007\*"中"

0.058\*"迁移" + 0.025\*"单击" + 0.019\*"颜色" + 0.019\*"数据" + 0.017\*"组件" + 0.015\*"样式" + 0.013\*"输入" + 0.013\*"手动" + 0.013\*"数值" + 0.013\*"节点"

0.019\*"中" + 0.015\*"调用" + 0.009\*"数据" + 0.009\*"请求" + 0.009\*"服务" + 0.009\*"消息" + 0.008\*"接口" + 0.008\*"日志" + 0.007\*"方法" + 0.007\*"客户端"

0.030\*"监控" + 0.016\*"报警" + 0.015\*"请求" + 0.014\*"规则" + 0.012\*"数据" + 0.010\*"服务" + 0.010\*"时间" + 0.010\*"流量" + 0.009\*"访问" + 0.008\*"云"

每一条输出表示一个主题，每个主题由若干个重要词语组成，词语前的数字表示该词的概率。如第一个主题由数据、日志、中、查询、表、配置、SQL、执行、点击、操作等关键词组成，我们大致可以将其主题自定义为大数据查询；第二个主题由业务、蚂蚁、数据、服务、金融、技术、金服、能力、提供、中等关键词组成，我们可以将其主题自定义为蚂蚁金服的数据服务。其他主题同理自定义，以确定其主题内容。

**存在问题：**

1. 训练数据少，所以训练出来的结果差强人意，需要用更多数据进行建模。
2. 停用词表需要进一步完善，且可以加入一些除词的规则；另外分词部分可以自定义字典，加入新的分词。
3. 运行速度比较慢，可以通过改变主题数或者学习率来改善。

**应用：**

得到的LDA模型可用于主题预测，给定新的文档预测文档主题分布，可用于分类。

**疑惑：**

是应该将ABCDI所有的文本一起进行主题分析，还是ABCDI每一部分分别进行主题分析，得到一个大主题下的小主题？ （这里两种情况都写了相应的代码）

**附件：**

停用词表

爬虫数据

代码

主题模型结果