**使用LDA主题模型对问询函文本内容进行主题分类**

一．LDA主题模型介绍

LDA（Latent Dirichlet Allocation），是Blei等人提出的基于概率模型的主题模型算法，LDA是一种非监督机器学习技术，可以用来识别大规模文档集或语料哭肿的潜在隐藏的主题信息。它采用了词袋（bag of words）的方法，这种方法将每一篇文档视为一个词频向量，从而将文本信息转化为了易于建模的数字信息。该方法假设每个词是由背后的一个潜在隐藏的主题中抽取出来。对于语料库中的每篇文档，LDA定义了如下的生成过程：

（1）对每一篇文档，从主题分布中抽取一个主题；

（2）从上述被抽到的主题所对应的单词分布中抽取一个单词；

（3）重复上述过程直至遍历文档中的每一个单词。

图示

描述已自动生成语料库中的每一篇文档与（通过反复试验等方法事先给定）个主题的一个多项分布相对应，将该多项分布记为。每个主题又与词汇表（vocabulary）中的个单词的一个多项分布相对应，将这个多项分布记为。上述词汇表是由语料库中所有文档中的所有互异单词组成，但实际建模的时候要剔除一些停用词（stopword），还要进行一些词干化处理等。和分别有一个带有超参数（hyperparameter）和的 Dirichlet先验分布。对于一篇文档中的每一个单词，我们从该文档所对应的多项分布中抽取一个主题，然后我们再从主题所对应的多项分布中抽取一个单词。将这个过程重复次，就产生了文档，这里的是文档的单词总数。这个生成过程可以用如下的图模型表示：

二．模型分析过程

1.获取原始文本

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成首先，通过Python爬虫手段从交易所网站爬取所有的监管问询函发出清单。一共获取到2326条问询信息，其中每一条问询信息包含：时间、公司代码、函件类别、公司简称、函件类型、网址和函件编码7栏信息。

此外，函件类别中共有7类：问询函（830条）、定期报告事后审核意见函（518条）、重大资产重组预案审核意见函（435条）、并购重组审核意见函（297条）、定期报告信息披露监管问询函（165条）、年报审核问询函（80条）和监管工作函（1条）。

文本

描述已自动生成最后，使用requests模块根据得到的网址信息，以函件编码为唯一识别符，从网上上下载所有的函件文档并存储到指定文件夹：

2.文本预处理

首先图片包含 文本

描述已自动生成，由于下载的函件文档均为pdf格式，因此需要对其进行解析以获取文本内容方便后续的处理。这里使用pdfplumber模块对每一个函件文档进行解析，最后保存为txt文档：

接下来，从已经解析好的txt文档中读取文本内容，使用jieba分词模块进行文本清洗、分词、去除停用词的操作。为了优化专用领域的分词效果，需要补充特定的专用词典到jieba的分词词典当中。这里使用到的补充词典有6个，其中“财经金融词汇”和“经济金融类词汇”来自搜狗输入法官网所提供的专有词典；另外4个词典来自GitHub相关项目的开源词典：

文本

描述已自动生成另外，为了提高LDA模型进行主题分类的效率，同时考虑本文所关注的主题分类文本对象的特征，在文本预处理中的结果中进保留名词（n）、机构团体（nt）、其他专名（nz）和状态词（z）。在这部分的最后，使用jieba提供的jieba.analyse.extract\_tags（基于 TF-IDF 算法）对文本内容进行了一次主题关键词的提取。

最后得到的列表texts存储了所有问询函文本经过预处理之后的文本数据。

3. LDA主题分析

目前，多种机器学习包均提供了LDA主题模型的模块，这里使用Sklearn中的LatentDirichletAllocation模块。

文本

描述已自动生成

主题模型中最主要的一个参数就是主题数量的确定。由于是无监督的学习，无法像监督学习那样通过验证集和测试集来对模型训练结果进行评估和进行调参。目前LDA分类效果的评定主要有如下几种方式：

1. 主观经验判断：不断调试、操作性强、最为常用；
2. 基于困惑度Perplexity：但需要测试集；
3. 贝叶斯统计标准方法；
4. 非参数方法：Teh提出的基于狄利克雷过程的HDP法。

本文采用第（1）种方式，即通过不断调整T来进行比对。为了使经验判更加合理，采取的方式是：从每一个主题中随机挑选某个词语，然后在主题之间进行调换，再由另一人观察每一个主题，看是否能够找出那一个外来词（从其他主题调换进来的）。如果能够正确分辨出外来词，说明此时的主题分类是比较准确的。经过不同T的尝试，最后确认的主题数量T的合理值为5。

三．主题分类结果

经过上述的模型分析过程，得到了所有文档的主题分类结果。首先来看5个主题的Top20分布的词汇（按分布概率排序）：

|  |  |  |
| --- | --- | --- |
| 主题 | Top20词汇 | 数量 |
| 1 | 股票上市 董事会 总股本 高送转 核实 转增股本 知情人 方案 议案 收函 书面形式 含税 比例 内幕 资本公积金转增股本 减持 经营规模 派发现金红利 公司 函件 | 53 |
| 2 | 战略 德合 小微企业 德信 典当 金融 经营权 贷款人 产权交易 委托人 产业链 部分股权 信托计划 协议书 汽车空调 车辆 领域 受托人 后级 建信 | 8 |
| 3 | 项目 内容 投项 失败率 产量 油价 流动资金 股权融资 哈萨克斯坦证券交易所 漏损 能源行业 修改版 江苏汇金控股集团有限公司 咨询服务 布伦特原油 山西广和山水文化传播股份有限公司 媒体报道 战神 目目 集资 | 24 |
| 4 | 基准日 定价 预案 均价 交易日 股票交易 资质 债权人 股份发行 办理 书面 问题 价格 换股 股份 支付能力 符合国家 银座 发电量 数额 | 27 |
| 5 | 公司 情况 原因 标的 交易 资产 合理性 报告期 会计师 业务 格式 预案 计提 净利润 事项 上市公司 金额 控股股东 公告 股权 | 2214 |

从上表可以看出，归类在主题5的问询函文件数量最多，占比超过90%。这一主题下的关键词，主要是基本信息披露如公司基本信息、经营情况、财报数据和公告事项。这一主题所涉及的内容对几乎所有的企业都是适用的，因此可以看到证监会对这方面监管内容是经常性和全面性的。

数量第二多的是主题1。这一主题下值得关注的是和股权相关的关键词较多。因此可以认为被分类到这一主题下的问询函主要是关注与股权相关的事项。

主题3的主题关键词涉及一些特定行业的企业，可以看出是和能源相关的。包括这一行业内企业的项目投资、财务状况、融资等等事项。主题4主要涉及市场交易行情的关键词，体现这一类问询函对日常行情的特别关注。

主题2的关键词涵盖一些不常发生的事项，如产权交易、信托等。这类事项在市场中发生较少，因此数量最少。

从下表也可以看出，在证监会的函件类型分类下，除监管工作函之外，各个类型下最主要的关注主题同样均为主题5。也就说明，在各类型的问询函中，关注的重点仍然是对主题5所涵盖的公司基本情况、财务信息和经营情况等相关事项。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 问询函 | 定期报告事后审核意见函 | 重大资产重组预案审核意见函 | 并购重组审核意见函 | 定期报告信息披露监管问询函 | 年报审核问询函 | 监管工作函 |
| 主题1 | 52 |  |  |  |  |  | 1 |
| 主题2 | 6 | 1 | 1 |  |  |  |  |
| 主题3 | 22 |  | 1 | 1 |  |  |  |
| 主题4 | 10 |  | 9 | 8 |  |  |  |
| 主题5 | 740 | 517 | 424 | 288 | 165 | 80 |  |

从以上对各个主题的分析可以看出，证监会的监管特点是对基本信息披露的监管非常关注，是监管的重中之重。其次是对一些特色的交易事项予以特别关注，如与股权、并购重组相关的事项。下表是对这部分LDA文本主题分类的结果汇总，并按照重要性进行了标示：

|  |  |
| --- | --- |
| 主题 | 主要关注内容 |
| 1 | 股权相关的事项★★★ |
| 2 | 稀有事项，如产权交易、信托★ |
| 3 | 特定行业的相关事项★★ |
| 4 | 市场交易行情★★ |
| 5 | 基本信息、经营情况、财报数据★★★★ |

图表, 饼图

描述已自动生成图表, 饼图

描述已自动生成图表, 饼图

描述已自动生成图表, 饼图

描述已自动生成图表, 饼图

描述已自动生成

https://github.com/Wunala/LDA-topic-model\_jieba\_finacial-text