通联数据
DATAYES

FDDC2018

“A股上市公司公告信息抽取”赛题辅导

2018.05.30

通联·智能投资服务体系

通联数据 DATAYES

业务应用层

- 智能基本面投研
- 大数据量化投研
- 基金评价与资产配置管理
- 智能投顾服务

模型算法层

智能搜索引擎 | 盈利预测 | 投研框架模型 | 量化因子 | 风险模型优化器 | AI宏观预测 | 资产配置模型 | ...

技术处理层

自然语言处理	去重/聚类	信息抽取	实体识别	事件识别	观点分析
关系图谱	标签化	情感分析	衍生指标	在线学习	机器学习

数据层

采集 / 加工 / 整合

基础数据	股票	债券	沪深高频	期权高频	公司概况	证券概况	基金	宏观行业
行业特色数据	汽车产销	房地产	医药生物	农林牧渔	能源化工	有色金属	钢铁	交通运输
大数据	热点主题	知识图谱	微信微博	新闻公告	券商研报	电商	公告	

目录

通联数据 DATAYES

赛题背景和意义

上市公司公告-概念和价值

通联数据 DATAYES

- 上市公司公告，是指上市公司按照证监会要求，通过指定平台向社会公众公布公司相关信息。
- 在股市的投资研究过程中，上市公司的公告披露是投资者的重要参考依据，尤其对于专业的机构研究员，挖掘公告重要信息是每日投研的必要过程。

公司业绩

财报、经营数据、重大合同

- 分析公司基本面
- 构建公司财务模型

市场预警

增减持、限售股解禁

- 表现对股价的信心
- 限售股解禁对股价造成压力

融资情况

定向增发、IPO

- 股本变动
- 股价溢价/倒挂分析

事件驱动

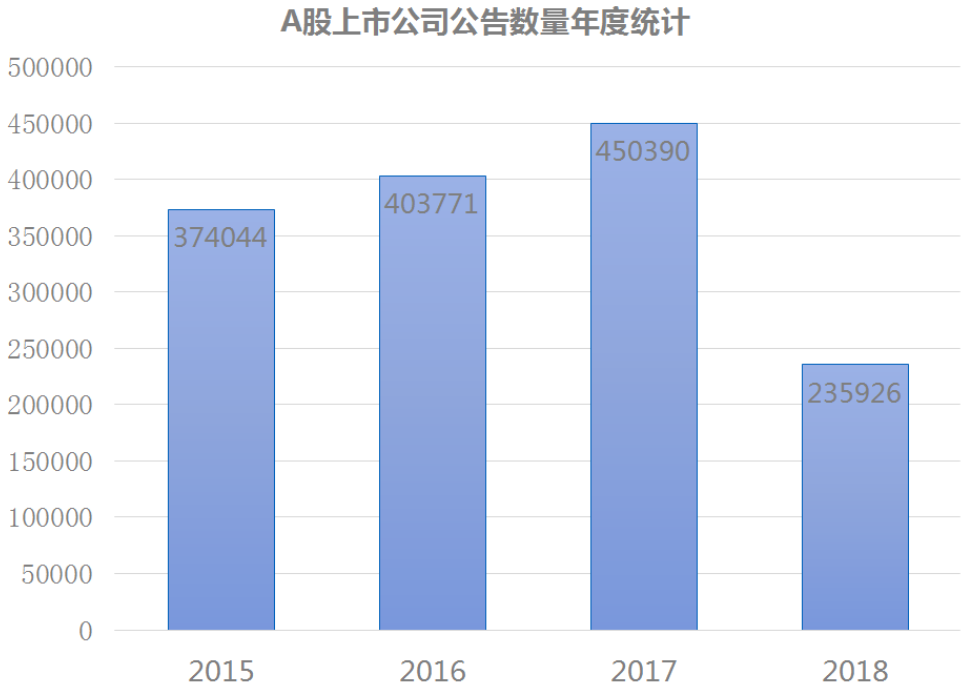
分红送转、资产重组

- 构建事件驱动策略
- 高送转

赛题背景和意义

通联数据 DATAYES

上市公司公告-难点



海量数据

- 仅A股公告，一年的数据量超过为40万篇
- A股、港股、新三板、债券
- 其他各种类型的金融文档

数据复杂性

- 部分篇幅很长(年报、定向增发、资产重组)，包含大量信息，耗时费力

人工处理难度很大

赛题背景和意义

通联数据 DATAYES

上市公司公告-结构化数据抽取

- 使用自然语言处理、机器学习算法对公告信息进行自动抽取
- 大幅提高投研效率，为投资者创造巨大的价值
  - 高性能
  - 准确性
- 自然语言处理-信息抽取问题的挑战
  - 与实际投资需求紧密结合
  - 国际级导师亲临指导
  - 丰厚的优胜奖金和顶尖金融机构的青睐

目录

通联数据 DATAYES

信息抽取介绍

通联数据 DATAYES

信息抽取介绍

通联数据 DATAYES

以分类模型为例

目标：抽取重大合同-合同金额

输入：签署了《邹县发电厂煤场扬尘治理改造项目(EPC)》，合同金额为27,590万元人民币

..... /合同 /金额 /为 /27,590 /万元 /人民币 .....

..... *X*(合同) *X*(金额) *X*(为) *X*(27,590) *X*(万元) *X*(人民币) .....

F(*X*(27,590)) = *{label_1, lable_2 ...}?*

目录

通联数据 DATAYES

Part 1

通联数据 DATAYES

各公告类型业务与字段说明

股东增减持

通联数据 DATAYES

- 由上市公司披露的，上市公司股东在交易所系统进行增持股票或减持股票的信息
- 公告id+股东名称+变动日期唯一确定一条增减持记录
- 抽取位置：文本 & 表格

字段	是否可能为空	约定单位	释义	说明
股东全称	否		股东全称	个人或公司
股东简称	是		股东简称	特指公司简称
变动截止日期	否	XXXX-XX-XX	股东增减持变动的日期	
变动价格	是	元	增减持的交易价格	如无直接公布无需计算
变动数量	否	股	增持或减持的数量	
变动后持股数	是	股	增持或减持的股数	如无直接公布无需计算
变动后持股比例	是	1	增持或减持股数占总股本的比例	如无直接公布无需计算

重大合同

- 由上市公司披露的，上市公司或其子公司的项目中标、合同签署信息
- 在项目和合同中存在甲方、乙方的概念，一般来说，上市公司及其子公司作为乙方出现
- 公告id+甲方+乙方唯一确定一条重大合同记录
- 抽取位置：文本

字段	是否可能为空	约定单位	释义	说明
甲方	是		指提供合同方，或招标方	公司或机构
乙方	否		竞标或投标方	公司或机构
项目名称	是		项目名称，通常表述为“xxx项目”	
合同名称	是		合同名称，通常表述为“xxx合同”	
合同金额上限	是	元	项目或者合同的金额 项目金额通常指中标价 合同金额为合同中约定的金额	合同总金额涉及外币的：如果原文中将外币折合成人民币，优先抽取人民币金额，否则抽取外币金额。样例数据未对外币金额进行汇率换算。
合同金额下限	是	元		
联合体成员	是		联合竞标、施工的合作方	公司或机构

重大合同

1. “合同金额上限”&&“合同金额下限”：根据披露方式不同，进行不同的赋值

情况	只披露上限	只披露上下限	同时披露上下限	披露精确值
说明	赋值给“金额上限”	赋值给“金额下限”	分别赋值	上下限都赋精确值

2. “联合体成员”：可能需要抽取多个成员进行拼接
    1. 公告中描述为“联合体为A公司，B公司，C公司”时，将A公司，B公司，C公司分别标记为联合体成员。
    2. 公告中描述为“A公司，B公司，C公司共同中标”时，将A公司标为乙方，B、C公司标为联合体。
    3. 签署合同，多个乙方时，选一个作为乙方，其余放至联合体字段。
    4. 拼接时使用“、”分割

定向增发

通联数据 DATAYES

- 定向增发：上市公司向符合条件的少数特定投资者非公开发行股份的行为
- 公告id+定增对象唯一确定一条定增记录
- 抽取位置：文本 & 表格

字段	是否可能为空	约定单位	释义	说明
增发对象	否		是指认购本次非公开发行股票的投资者	企业、机构或者自然人
发行方式	是		通过发行价格描述进行判断， 描述为"发行价格为xx元/股",判断为"定价"; 描述为"发行价格不低于XXX"，判断为"竞价"	
增发数量	是	股	指增发对象认购本次非公开发行的股票数量	可能是一个确定的数，也可能是一个范围。当是一个范围时，优先抽取上限
增发金额	是	元	指增发对象认购本次非公开发行股票的出资金额	可能是一个确定的数，也可能是一个范围，当是一个范围时，优先抽取上限； 增发金额并不单纯指增发对象以现金认购的金额， 如果增发对象的认购方式既包含现金， 又包含其他认购方式，则增发金额指的是两种认购方式的价值总计
锁定期	是	月	指增发对象认购的股份自发行结束之日起到可以上市交易或转让之间的期限	锁定期可能为多个增发对象共用
认购方式	是		各个增发对象认购本次非公开发行股票的认购方式	

定向增发

通联数据 DATAYES

1 、本次非公开发行股票相关事项已经获得公司第五届董事会第三十五次（临 时 ）会议、第六届董事会第六次（临时 ）会议、第六届董事会第八次（临时 ）会 议、第六届董事会第九次（临时 ）会议、 2015 年第六次临时股东大会和 2016 年 第四次临时股东大会审议通过。根据有关法律法规的规定，本次非公开发行股票 方案尚需上市公司股东大会审议通过、中国证券监督管理委员会核准后方可实施。

2 、本次非公开发行的发行对象为公司实际控制人饶陆华、桂国才、孙俊、深圳市国银资本投资管理有限公司 - 国银资本稳健 1 号证券投资基金、郭伟、祝文闻、陈长宝，本次公司与饶陆华、桂国才的交易构成关联交易。

认购方式

3 、本次向特定对象非公开发行的股票合计 8,908 万股，全部以现金认购。依据公司与各发行对象签署的附条件生效的股份认购协议及其补充协议，各发行对象认购情况如下：

发行对象	认购金额（万元）	认购股份（万股）	认购比例
饶陆华	101,578.40	4,760	53.44%
陈长宝	24,754.40	1,160	13.02%
桂国才	21,297.32	998	11.20%
孙俊	10,029.80	470	5.28%
祝文闻	16,858.60	790	8.87%
深圳市国银资本投资管理有限公司 - 国银 资本稳健 1 号证券投资基金	8,536.00	400	4.49%
郭伟	7,042.20	330	3.70%
合计	190,096.72	8,908	100.00%

增发对象

增发数量：单位万股需要转换为 股。

增发金额：单位万元需要转换为 元。

若公司股票在定价基准日至发行日期间发生派息、送红股、资本公积金转增股本等除权除息事项，本次发行的数量将作相应调整。本次非公开发行股票数量以中国证监会最终核准发行的股票数量为准。

4 、公司本次非公开发行的定价基准日为公司第五届董事会第三十五次（临时 ）会议决议公告日（ 2015 年 10 月 29 日 ）。本次非公开发行价格为 21.34 元 / 股，不低于定价基准日前 20 个交易日公司股票交易均价的 90% 。公司股票在定 价基准日至发行日期间如有派息、送股、资本公积金转增股本等除权除息事项，将对发行价格进行相应调整。

发行方式：定价

如本次非公开发行的价格低于公司本次非公开发行之发行期首日前 20 个交 易 日公司股票交易均价的 70% 的，则本次非公开发行的价格以发行期首日前 20 个交易日公司股票交易均价的 70% 为准。

5 、公司本次非公开发行拟募集资金总额 190,096.72 万元，扣除发行费用后用于智慧能源储能、微网、主动配电网产业化项目、新能源汽车及充电网络建设与运营项目、智慧能源系统平台项目、 110MW 地面光伏发电项目。

6 、本次非公开发行的股票自发行结束之日起三十六个月内不得转让。

锁定期：36

7 、本次非公开发行股票不会导致公司控股股东及实际控制人发生变化。

8 、本次发行完毕后，不会导致公司股权分布不符合上市条件之情形。

9 、为兼顾新老股东的利益，本次发行前滚存的未分配利润将由本次发行完成后的新老股东按照持股比例共享。公司利润分配政策、最近三年利润分配、未来三年

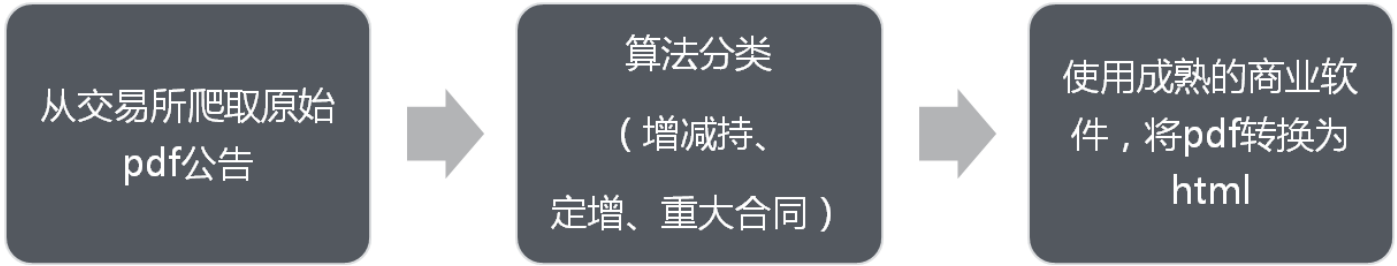
Part 2

通联数据 DATAYES

公告PDF与HTML格式

公告PDF与HTML格式

通联数据 DATAYES



- PDF版本：交易所公开数据，命名：公告id.pdf
- 便于人眼阅读，了解公告内容
- 程序读取需要一定工程化的工作量，但是可以保留所有字符的全部信息，包括位置坐标

- HTML版本：使用商业软件从pdf转换而来，命名：公告id.html
- 人眼可读性较pdf略差，但是便于程序读取解析
- 第三方软件转换时，难免会存在格式上的错误（例如将文本识别为标题），需要选手想办法在模型层面解决

公告PDF与HTML格式

通联数据 DATAYES

```
<div id="SectionCode_2" title="一、股东减持情况" type="paragraph">LF
 <div id="SectionCode_2-1" title="1、股东减持股份情况" type="paragraph">LF
 <div id="SectionCode_2-2" title="2、股东本次减持前后持股情况" type="paragraph">LF
  <div type="content">LF
   <table cellspacing="0">LF
    <tbody>LF
    <tr>LF
    <tr>LF
   </tbody></table>LF
  </div>LF
  <div type="content">LF
   <hidden name="a4">LF
   </hidden>LF
  </div>LF
 </div>LF
</div>LF
<div id="SectionCode_3" title="二、其他相关说明" type="paragraph">LF
 <div type="content">LF

  1、上述股东本次实施减持情况与披露的减持计划一致，符合《证券法》、《上市公司收购管理办法》、《深圳证券交易所创业板股票上市规则》、
  《深圳证券交易所创业板上市公司规范运作指引》和证监会《上市公司股东、董监高减持股份的若干规定》（证监会公告[2017]9
  号）、深圳证券交易所《深圳证券交易所上市公司股东及董事、监事、高级管理人员减持股份实施细则》等相关法律、法规以及相应承诺的要求。LF
 </div>LF
 <div type="content">LF
  2、公司将继续关注上述股东股份减持计划实施的进展情况，并按照相关法律法规的规定及时履行信息披露义务。LF
 </div>LF
</div>LF
```

公告PDF与HTML格式

通联数据 DATAYES

HTML标签说明

标签层级	标签名称	释义	子属性	子属性释义
0		公告标签	title	公告标题
1		公告段落	title	段落标题
2		公告正文		
3	<hidden>	页码	name	页码计数，从0开始
3	<table>	表格		
4	<tbody>	表格		
5	<tr>	表格行		
6	<td>	表格列(单元格)	rowspan	单元格跨度
7	<image>	图片		

- 标签层级序号大的标签中不可以包含层级低的标签
- 段落(paragraph)中可以包含子段落
- 正文(content)中**不会**再包含content
- 表格(table)，页码(hidden) 必须在content标签中
- 表格(table)中只有一个tbody标签，并且不会包含子表格
- 一个表格(tbody)会有多行(tr)，一行(tr)会有多列(td)
- 一个表格行(tr)或表格列(td)中不会再有其他的tr或td
- 图片(image)必须在content标签或td标签中，image在本次比赛中无需使用，可以直接过滤

Part 3

通联数据 DATAYES

赛题数据-技术难点

技术难点

通联数据 DATAYES

1. 本次比赛中训练数据以原始文本-最终结果的端到端形式给出，没有在原文中标注位置
2. 本次比赛中提供的所有训练、测试数据，全部由人工生产，并经过二次校验；但即使在这种情况下，仍然有部分数据存在错误。
3. 本次比赛中提供了html格式的语料，转换中造成部分格式的错误需要选手在建模时予以考虑
4. 一篇公告中存在多条记录
5. 本次比赛中的部分字段需要经过转换或计算才能得到最终结果

目录

通联数据 DATAYES

规则解答

通联数据 DATAYES

赛制说明：

赛事阶段		赛事安排	注意点	时间点
初赛	一阶段	发布训练数据集、测试数据集A	开发算法	5月18日12：00 -7月8日12：00
	二阶段	发布测试数据集B	1天时间，提交结果	7月8日13：00 -7月10日12：00
	代码审核	审核代码， Top100进复赛	实名认证 Top120提交代码审核	7月10日12：00 -7月12日12：00
复赛	一阶段	发布复赛数据集， 更换若干公告类型	天池平台调试程序，保证可以成功运行	7月13日12:00 -8月8日22:00
	二阶段	平台统一运行程序进行评测， Top 5 进决赛	程序提交后不可修改， Top 10团队提交代码审核	8月9日-8月22日
决赛	决赛	现场答辩	提交答辩PPT、参赛总结、算法核心代码	8月29日-8月30日

规则解答

通联数据 DATAYES

评分规则：

对参赛选手提交的结果，对每个字段、按如下方法进行判别和统计

类别	判断标准	标记
Possible	标准数据集中该字段不为空的记录数	POS
Actual	选手提交结果中该字段不为空的记录数	ACT
Correct	主键匹配 且 提交字段值=正确字段值 且 均不为空	COR

评分公式

指标	公式
字段召回率：	$\text{Recall} = \frac{COR}{POS}$
字段准确率：	$\text{Precision} = \frac{COR}{ACT}$
字段F1 ：	$\text{F1} = \frac{2 \times Recall \times Precision}{Recall + precision}$
类型得分	$\text{Score} = \frac{1}{n}\sum_{i=1}^{n} F1_i$ ，其中n为该类型字段总数、F1_i为第i个字段的F1得分
最终得分	$\text{Score} = \frac{1}{C}\sum_{i=1}^{C} Score_i$，其中C为数据集中的公告类型总数，$\text{Score}_i$为第i个类型的得分

关注FDDC2018

通联数据
DATAYES