



笃行·致远

**2019第三届顺丰信息安全峰会**

2019 THE 3<sup>rd</sup> SF INFORMATION SECURITY SUMMIT



2019第三届顺丰信息安全峰会



# 基于自然语言处理的 非结构化敏感信息识别

王南飞

大数据挖掘与分析工程师

# 目录



2019第三届顺丰信息安全峰会



1  
敏感信息检测背景

2  
自然语言技术的应用

3  
总结与展望





# 敏感信息检测背景

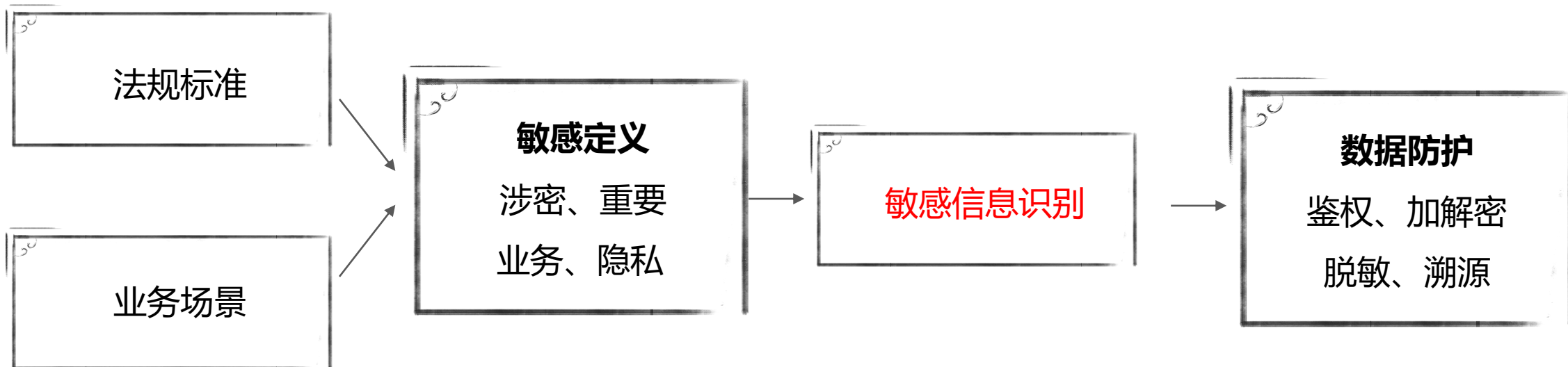




# 数据安全治理



2019第三届顺丰信息安全峰会






# 敏感信息检测的挑战

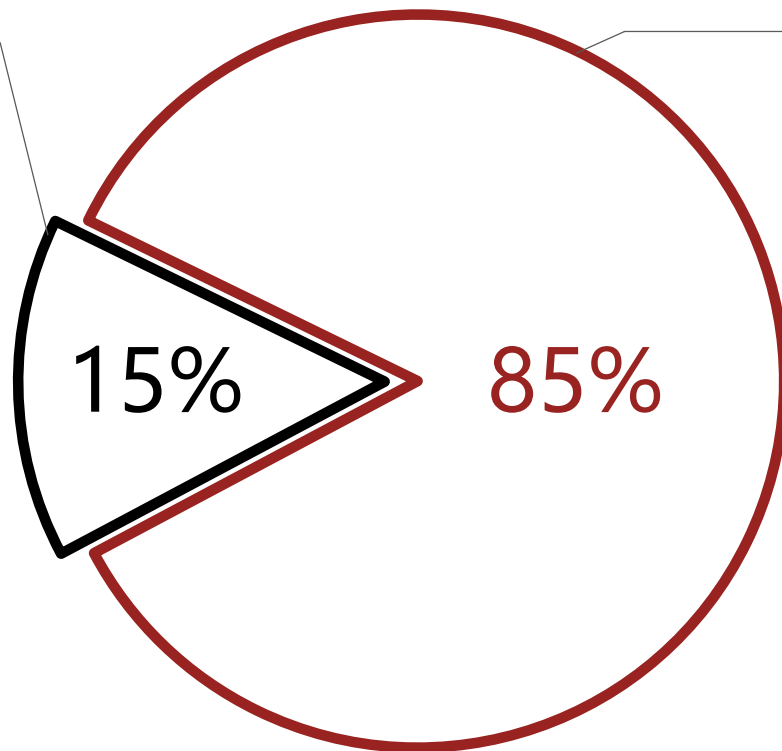


2019第三届顺丰信息安全峰会



## 结构化数据

 数据库



## 非结构化数据

-  视频培训、社交视频
-  扫描票据、身份证扫描件
-  图片资料、电子地图
-  产品资料、说明书PDF
-  录音文件、音频邮件
-  书稿、网站内容







# 敏感信息检测的常规方法



2019第三届顺丰信息安全峰会



常规方法：

- 人工审核
- 标记水印
- 设置规则
- 关键字



不足：

- 主观性强
- 覆盖面小
- 实时性差





# 敏感信息检测的新思路



2019第三届顺丰信息安全峰会



如何从形式多样的动态文件中准确识别含有敏感信息的文件？

# NLP



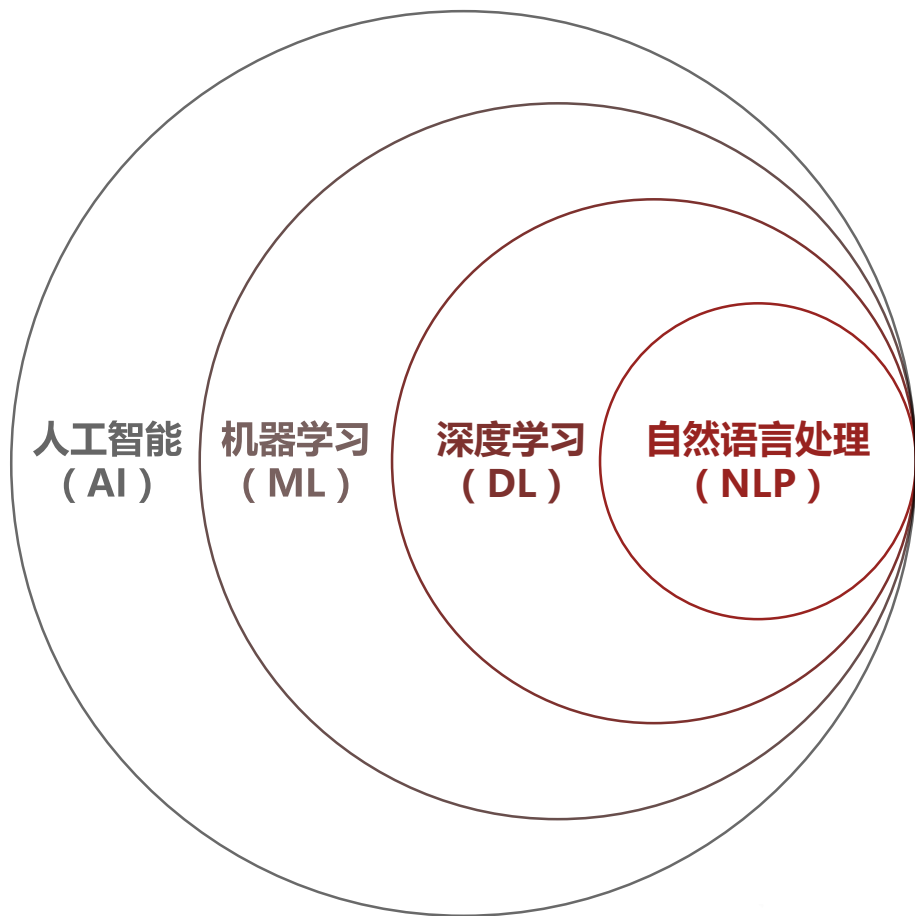




# 自然语言处理 ( NLP )



2019第三届顺丰信息安全峰会



网络舆情监控



代码漏洞检测



DGA检测



反欺诈风控



垃圾邮件过滤



...





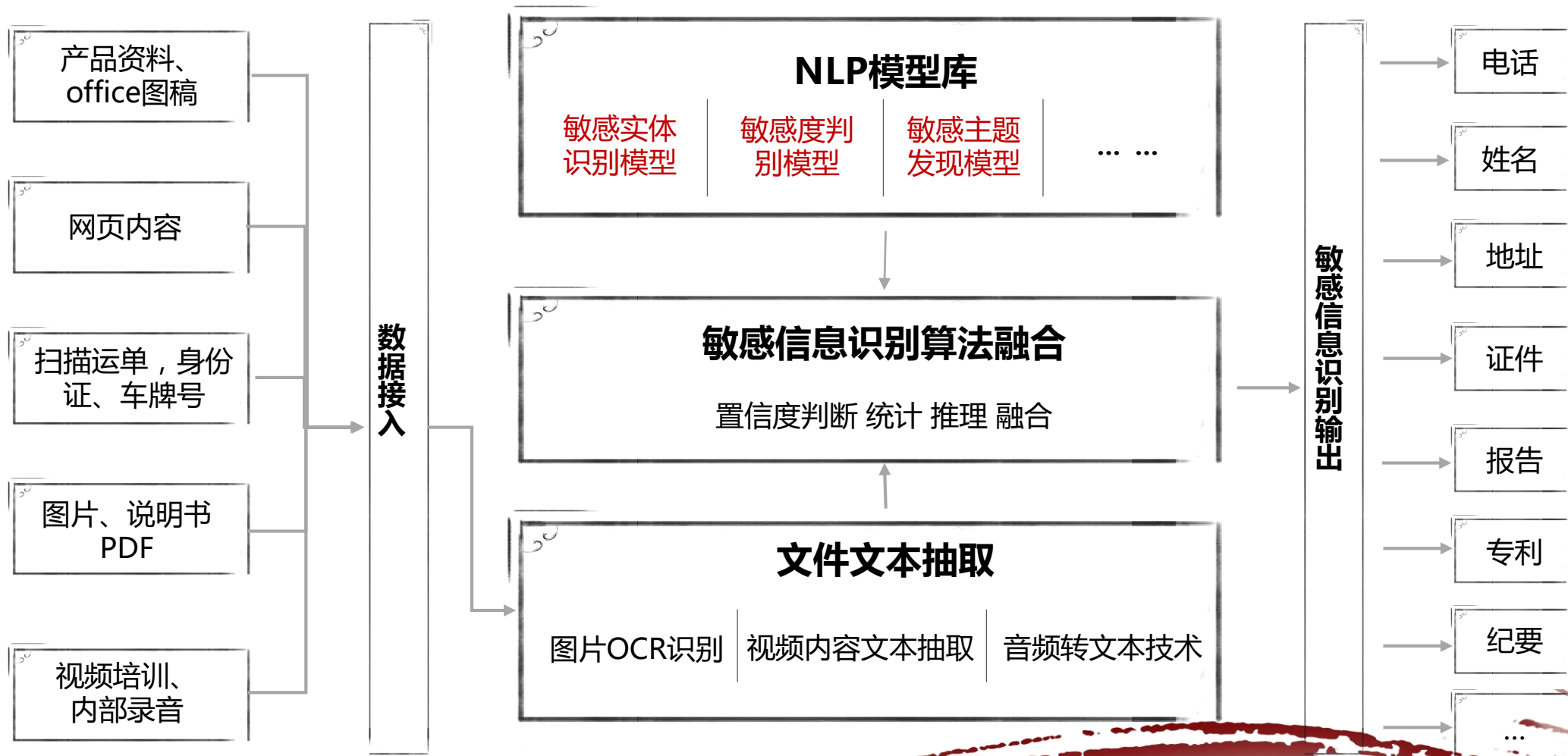
## NLP敏感信息检测



# 敏感信息检测整体流程



2019第三届顺丰信息安全峰会

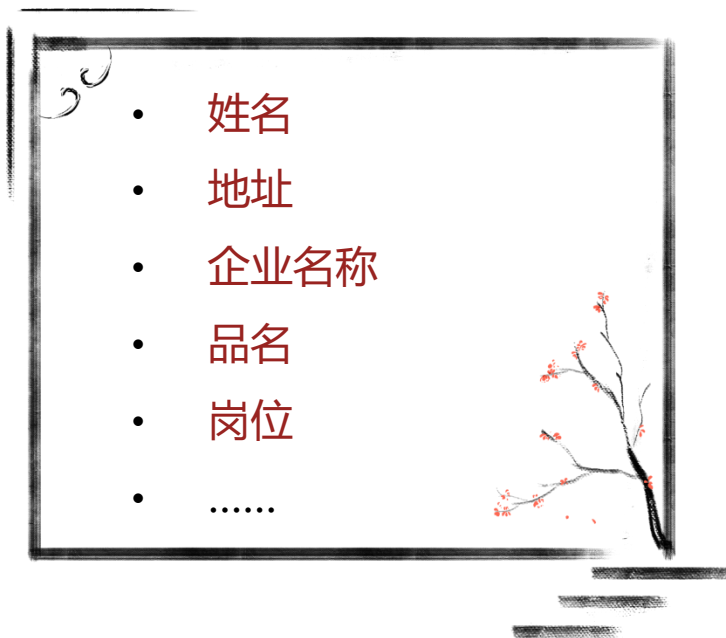
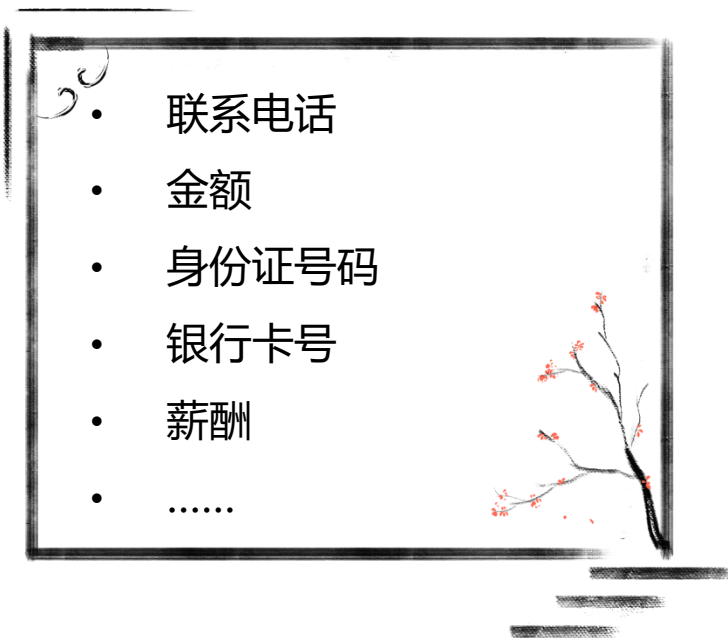




## 短文本敏感信息识别



2019第三届顺丰信息安全峰会



如何从短文本中，检测出人名，地址，企业名称等重点信息？





# 敏感实体识别模型



2019第三届顺丰信息安全峰会



张亮说，顺丰的总部在深圳市南山区，快递服务很好。



实体识别

张/**N**亮/**N**说/**O**，/**O**顺/**Org**丰/**Org**的/**O**总/**O**部/**O**在/**O**深/**A**圳/**A**市/**A**  
南/**A**山/**A**区/**A**，/**O**快/**O**递/**O**服/**O**务/**O**很/**O**好/**O**./**O**

- 姓名 /N
- 企业 /Org
- 地址 /A
- 其余序列 /O



后处理

{张亮/Name}说，{顺丰/Org}的总部在{深圳市南山区/Address}，快递服务很好。

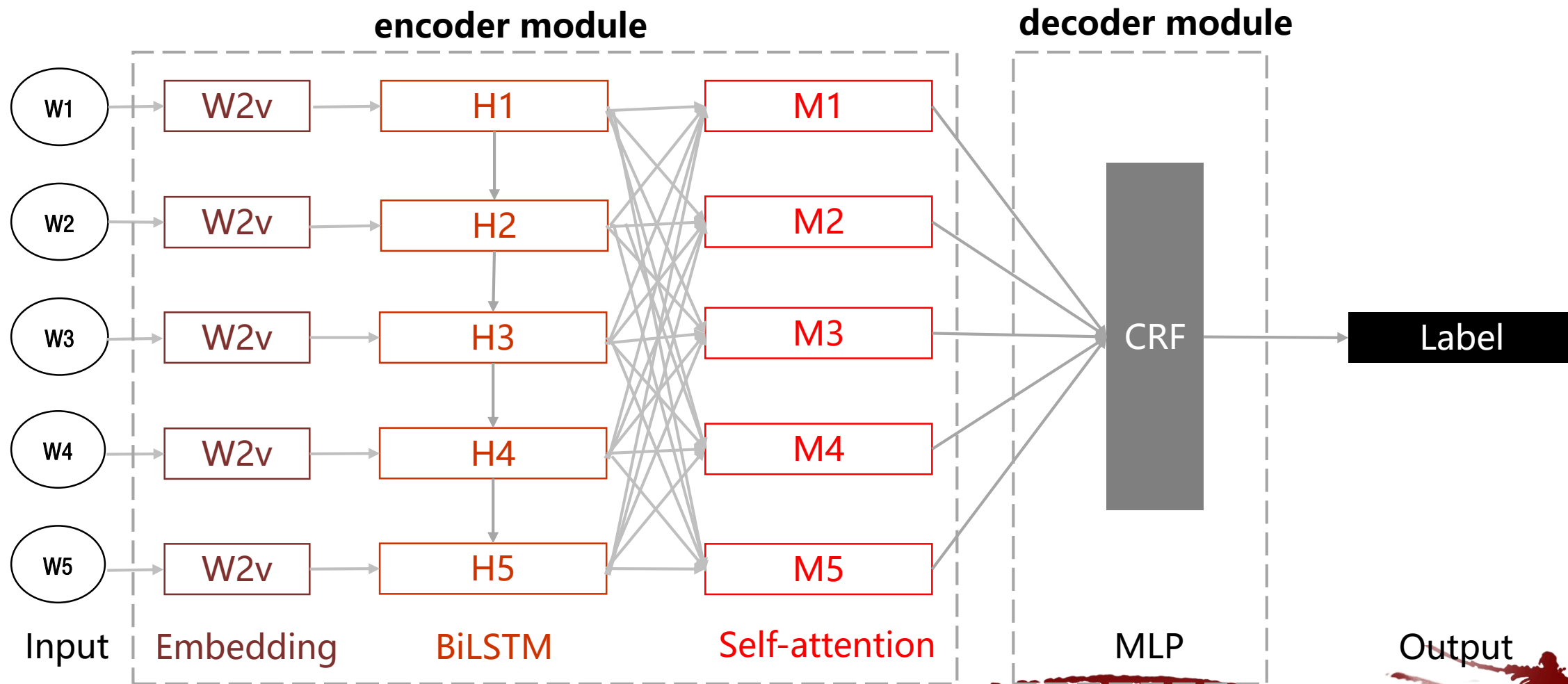




# 敏感实体识别模型



2019第三届顺丰信息安全峰会







# 应用效果



2019第三届顺丰信息安全峰会



样本校验

表名称: dbo.number

字段名: name

字段说明: 姓名

数据类型: varchar(255)

姓名(99%)

非敏感数据(1%)

样本数据

阳倩

毋辉

任丽娟

竺萍

东丹丹

公超

晋莹

时梅

巢波

姜强

确认

样本校验

表名称: dbo.number

字段名: adress

字段说明:

数据类型: varchar(255)

地址(97%)

非敏感数据(3%)

样本数据

甘肃省昆明县得阳计街k座 292264

西藏自治区通辽县清河兴安盟路l座 713620

贵州省兴安盟市长寿郑州路T座 612919

宁夏回族自治区辛集县东丽福州路c座 726786

辽宁省南宁市沈河割路L座 510399

新疆维吾尔自治区郑州区高港嘉禾街N座 228497

河北省惠州市大兴张家港街r座 106435

西藏自治区鹏市蓟州兴安盟路M座 740609

山西省宇县大兴晋路O座 166806

福建省明县新城镇街t座 667829

确认



# 长文本敏感信息识别



2019第三届顺丰信息安全峰会



- 合同
- 专利
- 企业营业报告
- 工作汇报
- 财务报告文件
- 市场调研材料
- ...

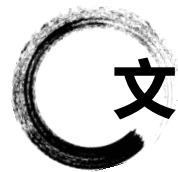


- 文本量巨大
- 敏感信息稀疏
- 主题丰富
- 应用场景多
- ...



如何从一篇文章的文本中，判断出文章是否是一篇较为敏感的文件？

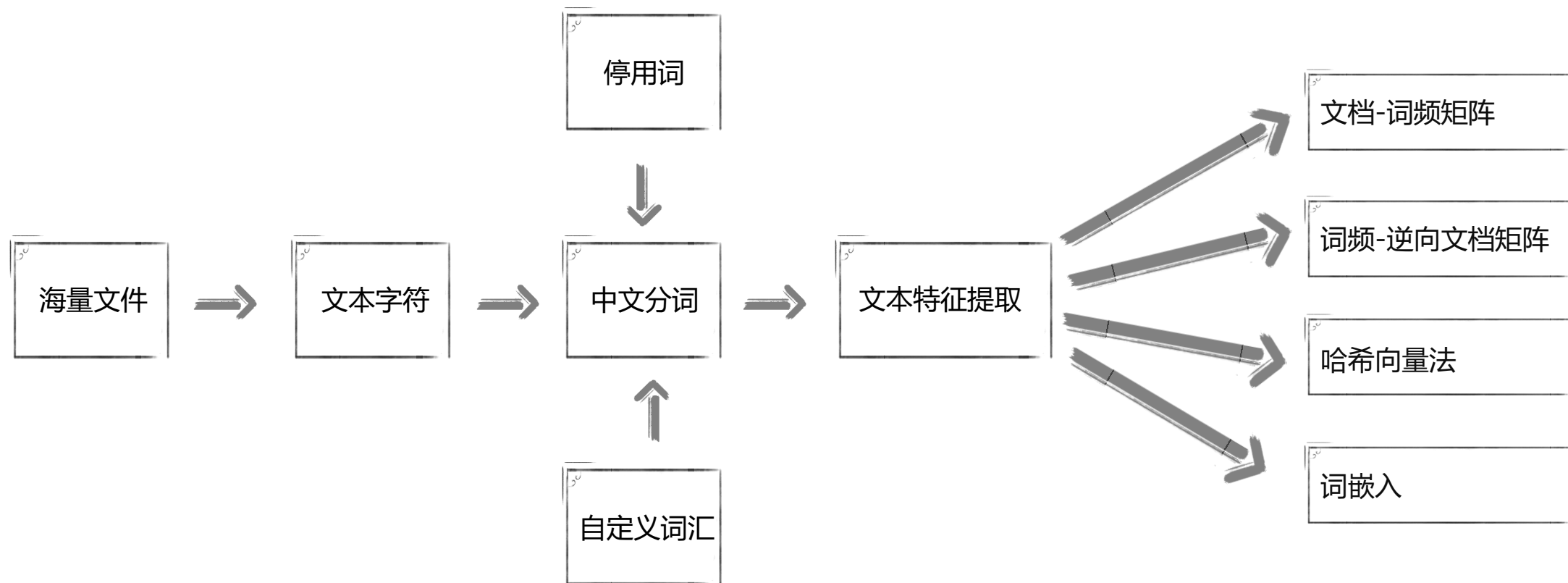




# 文本数值化处理



2019第三届顺丰信息安全峰会

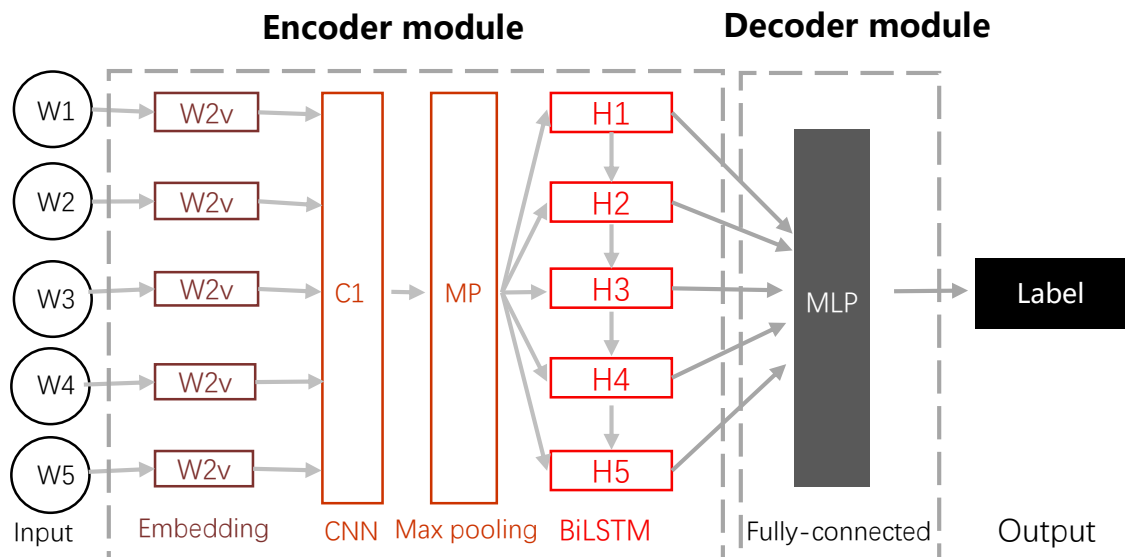




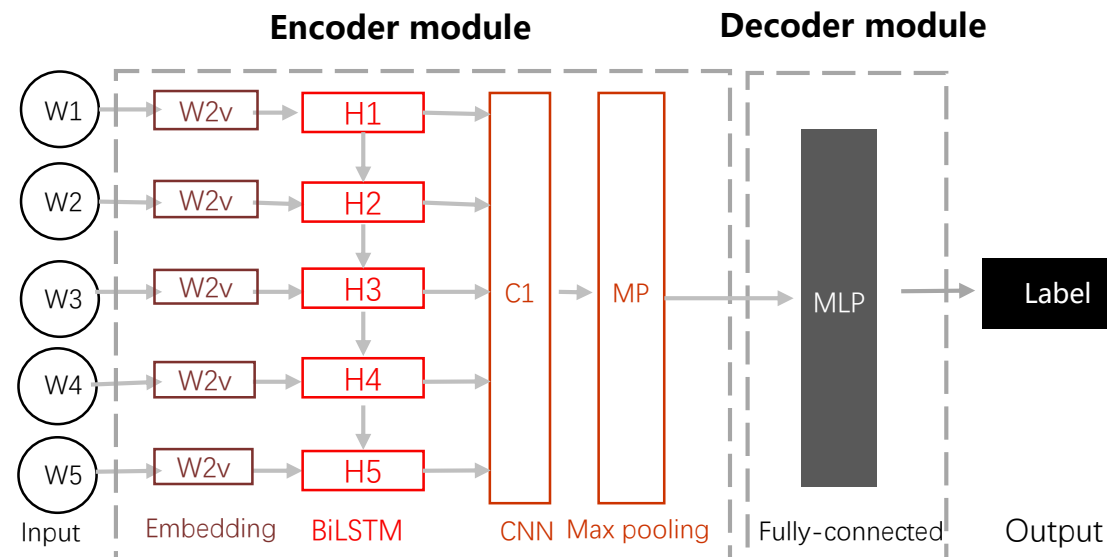
# 敏感度判别模型



2019第三届顺丰信息安全峰会



CNN-LSTM Model



LSTM-CNN Model





## 模型结果验证



2019第三届顺丰信息安全峰会



### CNN-LSTM顺丰整体敏感信息识别模型性能

敏感	81.6% 精准率	74.1% 召回率	67.7% F1值
----	--------------	--------------	--------------

非敏感	75.5% 精准率	81.3% 召回率	78.3% F1值
-----	--------------	--------------	--------------

AVG	78.5% 精准率	77.7% 召回率	73.0% F1值
-----	--------------	--------------	--------------

### LSTM-CNN顺丰整体敏感信息识别模型性能

敏感	91.5% 精准率	80.6% 召回率	85.7% F1值
----	--------------	--------------	--------------

非敏感	74.3% 精准率	88.2% 召回率	80.6% F1值
-----	--------------	--------------	--------------

AVG	82.8% 精准率	84.4% 召回率	83.2% F1值
-----	--------------	--------------	--------------





## 应用效果



2019第三届顺丰信息安全峰会



### KMS邮件通知

亲爱的 周奇：

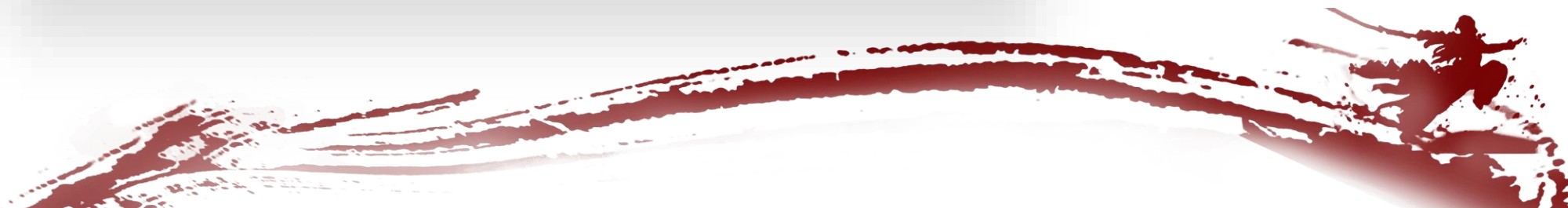
您好！经系统安全检测，你上传的知识附件“153439-0.xlsx.txt”识别为敏感，预测置信度为69.0%，请您关注并设定的正确的知识密级。谢谢！

正常：未包含敏感信息，密级可设定为公开；

敏感：包含敏感信息，密级建议设定为内部或机密。

【知识安全，从我做起】[点击查看详情](#)

知识管理系统



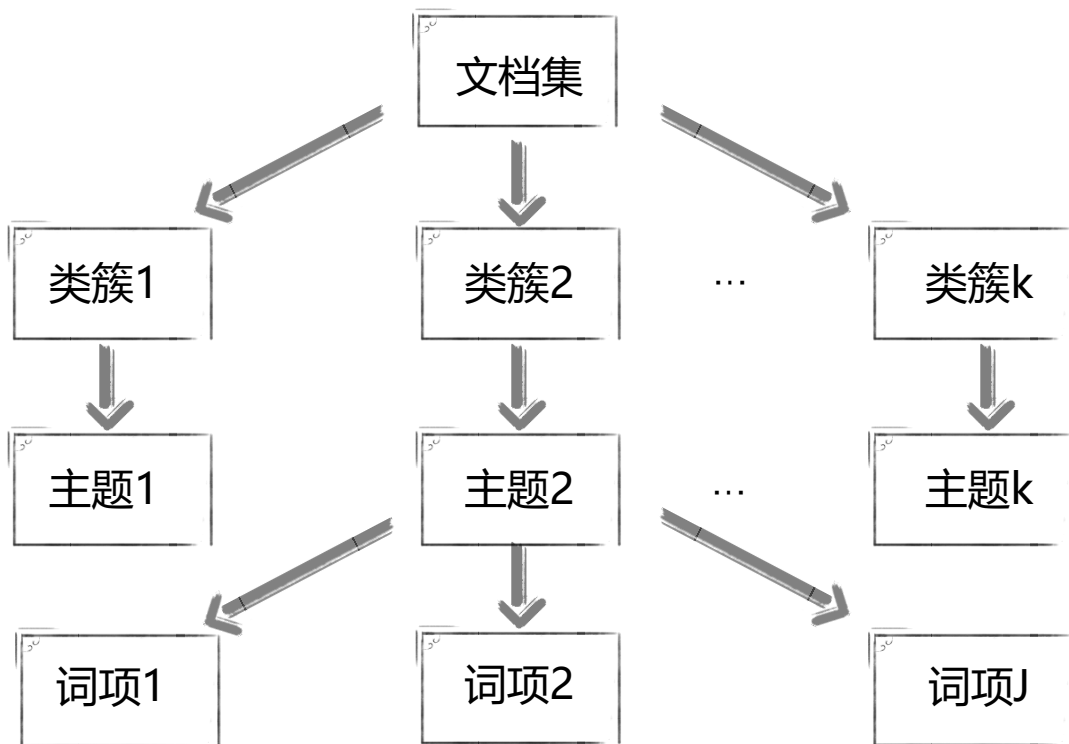




# 敏感主题发现模型



2019第三届顺丰信息安全峰会

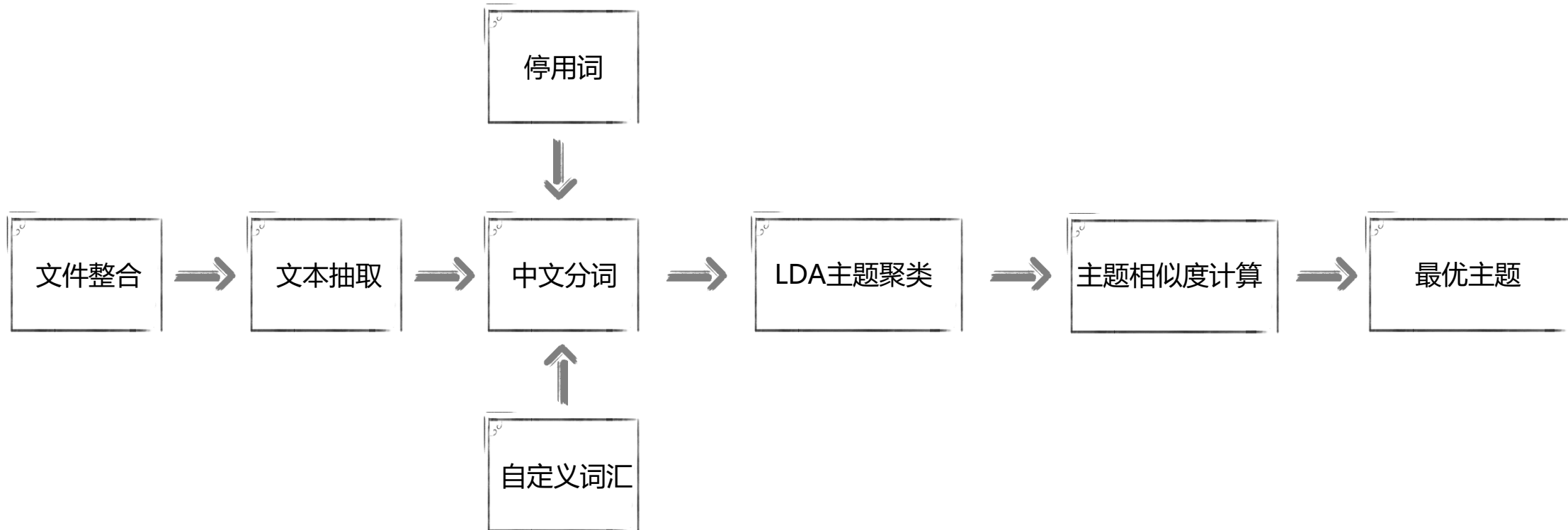




# 敏感主题发现模型



2019第三届顺丰信息安全峰会





## 敏感主题发现模型

笃行·致远

## 2019第三届顺丰信息安全峰会





## 总结与展望





## 模型优化方向



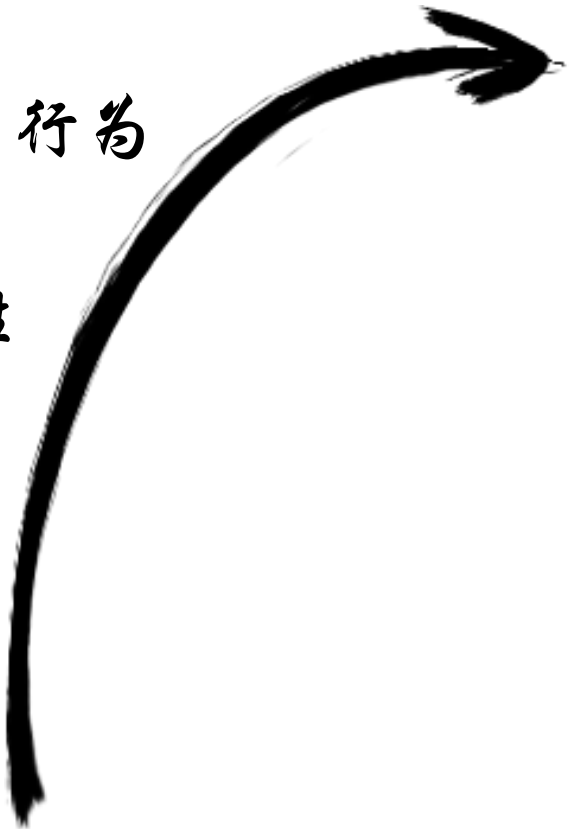
2019第三届顺丰信息安全峰会



用户行为

文件属性

文本内容



NLP是敏感信息识别的一把利剑，

敏感信息识别是数据安全的基石。