2019 年全国大学生信息安全竞赛

作品报告

作品名称：基于机器学习的在线招聘欺诈检测平台

电子邮箱：thefreer@outlook.com

提交日期：2019 年 6 月 5 日

填写说明

1. 所有参赛项目必须为一个基本完整的设计。作品报告书旨在能够清晰准确地阐述（或图示）该参赛队的参赛项目（或方案）。

2. 作品报告采用A4纸撰写。除标题外，所有内容必需为宋体、小四号字、1.5倍行距。

3. 作品报告中各项目说明文字部分仅供参考，作品报告书撰写完毕后，请删除所有说明文字。(本页不删除)

4. 作品报告模板里已经列的内容仅供参考，作者可以在此基础上增加内容或对文档结构进行微调。

5. 为保证网评的公平、公正，作品报告中应避免出现作者所在学校、院系和指导教师等泄露身份的信息。

目　　录

摘要

现如今网络招聘的兴起，可以看出，企业愈来愈主动拓展眼界，积极向外寻找适合的人才。与此相协调的是，而今求职者在寻找合适职位时，不仅仅局限于所在地区的企业，对于网络招聘的依赖程度日益增高。网络有效地拉近了全国范围内企业与求职者的距离。招聘的网络化已成为一种普遍的招聘模式。

然而，在线招聘并非一片净土，与之相反，由于网络平台的开放性、不安全性以及相关制度和法律的不健全等等原因，在线招聘平台逐渐变成了欺骗者的“无法无天之地”，特别是随着互联网的普及，受害者人群规模和地域范围呈扩大趋势。到目前为止，最常见的在线诈骗案例是就业诈骗。与电信诈骗问题不同的是，这一问题尚未得到应有的重视，到目前为止仍未得到充分的探索。

针对以上需求，本项目定义并描述了这一严峻的新型网络安全研究课题的特点。与此同时，我们提供了一个公开可用的中文招聘数据集，包括8640条带标签的招聘信息，并在此基础上提出一种在线检测与分析招聘信息真假的方案，发明了特征与文本相结合的评估方法以及一系列针对于招聘领域信息的分析方法，提供了虚假招聘信息检测与分析的接口，以此为核心建立了一个在线招聘欺诈检测平台。

关键词：欺诈检测;在线招聘;就业骗局;工作骗局;数据挖掘;机器学习;自然语言处理;数据集

第一章 作品概述

1.1.背景分析

随着个人电脑使用者的增多和互联网技术的普及与发展，企业进行招聘的方式也发生了很大的改变。从早期主要是内部推荐、张贴海报，至较早期发展为在报纸、杂志、电视或广播电台上发布招聘广告，再至到外地举办大型招聘会等方式，直至现如今网络招聘的兴起，可以看出，企业愈来愈主动拓展眼界，积极向外寻找适合的人才。与此相协调的是，而今求职者在寻找合适职位时，不仅仅局限于所在地区的企业，对于网络招聘的依赖程度日益增高。网络有效地拉近了全国范围内企业与求职者的距离。招聘的网络化已成为一种普遍的招聘模式。

然而，在线招聘并非一片净土，与之相反，由于网络平台的开放性、不安全性以及相关制度和法律的不健全等等原因，在线招聘平台逐渐变成了欺骗者的“无法无天之地”。

针对在线虚假招聘及虚假招聘信息，我们给出的定义为：

在线招聘欺诈是一种恶意行为，是指通过操纵并利用在线招聘平台系统的功能并发布虚假招聘信息的行为。

虚假招聘信息是指任何不以招聘人才为目的或招聘内容具有煽动性且缺乏一定真实性的招聘信息。

对于求职者来说，一些虚假的招聘信息可能会泄露求职者的简历上的一切隐私信息；也可能使求职者陷入金融诈骗等等，造成经济损失；异或使得求职者可能进入一些“表里不一”的企业，造成经济、精神损失；更有甚者，求职者被骗入培训机构甚至传销组织，危害人身安全。

而对于公司或者组织，被欺骗者冒用的公司会受到信誉及经济的损失，对公司造成极大的伤害。

更加令人沮丧的是，如今的虚假招聘变得越来越难以与真实招聘区分开来，越来越多的求职者陷入发布虚假招聘者的圈套之中，但是又没有一个很好的方法来区分这些招聘信息的真假。因此亟待寻求一种可以有效检测虚假招聘信息的方法来解决以上问题。

1.2.相关工作

到目前为止，针对在线招聘欺诈的研究少之又少，国内更是根本没有相关的公共研究，而各大招聘平台针对虚假招聘的处理方式也只处于最基础的关键词过滤检测方式。

放眼国外，2017 年 9 月，爱琴海大学信息与通信系统工程系——信息和通信系统安全实验室的研究者：H. John Heinz III College，Carnegie Mellon University，Pittsburgh 等人，在他们的论文《Automatic Detection of Online Recruitment Frauds: Characteristics, Methods, and a Public Dataset》中对在线招聘欺诈的研究进展进行了报告，论文中提供了虚假招聘爱琴海数据集，并根据此数据集分析了虚假招聘信息可能存在的特征，并基于机器学习构建了针对与此数据集的分类模型。

值得注意的是：论文中提及的研究方法仅仅针对于英文招聘信息，这与中文的招聘信息相差甚远，但是部分针对虚假招聘的分析方式仍然值得借鉴。

1.3.特色描述

众所周知,网络的发达带来便利的同时也会带来很多问题，招聘信息真假的检测便是其中一个关乎所有求职者切身利益的问题。我们带着希望求职,当然不希望被不法分子欺骗从而泄漏个人信息甚至危及个人安全。而此项目很好的解决了这个问题，项目特色如下：

此项目以一个一次性第三方平台的身份介入，不收集用户信息，不强制使用。

用户只需将自己持怀疑态度的招聘详情页链接输入查询，即可以得到一个客观可信的招聘评价结果，包括：职位描述得分、招聘数据得分、总分、各种文本分析及招聘信息真或假。

操作简单，界面清晰友好，分析一条招聘信息只需复制粘贴一个链接，无需输入任何文字

扩展性好，核心模块以接口的形式提供给网站前端，并且可根据增加的数据更新模型，我们还提供了检测结果纠错与招聘信息举报功能，而且每次检测的结果我们都会回收保存在日志中进行进一步分析，以上扩展功能保证了分类系统的可靠性。

1.4.应用前景分析

知名招聘网站 www.hays.com 曾经发布过这样一项调查报告，其中指出中国求职者最容易通过在线求职网站找到一份新工作。根据对 308 人进行的调查结果表明 43%的中国人认为通过在线招聘平台是找到新工作的最成功方法。排在招聘平台之后的是以推荐或口碑在主的个人网络，其得票率为 36%。而社交媒体网站的得票率为 21%。整个亚洲地区的调查结果与之类似。由此可见,不仅仅是在中国,乃至全世界,由于互联网的普及, 线上求职已成为必然的趋势。

但是,谁又能保护这些求职者的个人信息呢?

海量的招聘信息,隔着互联网,我们无法判别是真是假,不法分子很可能散步虚假招聘信息从而得到个人用户信息从而进行不法活动，求职者已经变得越来越不相信招聘平台的招聘信息审核机制，他们缺少一个帮助他们评判招聘信息的平台。所以,我们做的这个工作很有必要。它的身影也会随着在线招聘平台的进一步普及和发展而走进每个求职者的眼前。

第二章 作品设计与实现

2.1.系统方案

2.1.1. 数据获取

项目的第一步工作是进行数据的采集。最初的数据来源是智联网站上的招聘信息，但是由于智联的爬虫限制较为严重，故而我们最终将数据采集的对象设为了 58 同城。最终的数据总量为 2w 余条招聘信息。招聘职位涉及平时常见各种职位，涵盖了北京、上海、深圳、广州四个地区。

我们将所有信息分为三个数据表：company、recruitment、issue 进行存储。招聘信息数据库详细介绍详见项目附件一。

2.1.2. 数据初步分析与数据标记

招聘信息数据库构建完成之后，接下来的工作为对信息的真假进行人工判断。

首先我们从已有的 2w 余条招聘信息中，以城市为单位选取了 8640 条信息，其中北上广深四所城市分别占有 2160 条

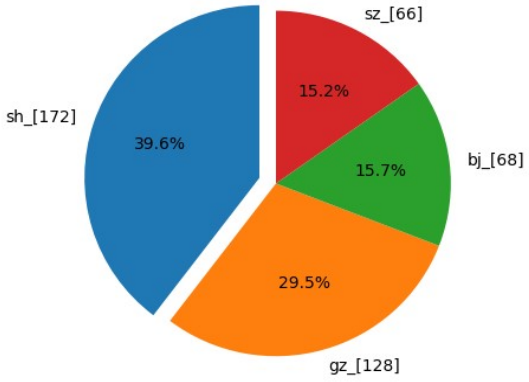
然后我们对选取的信息进行了初步分析，过滤了对人工判别的属性，详见下图，未打对勾的属性被过滤掉：

紧接着，为了保证人工判别的客观性，我们制定了人工判别数据的规则集，规则集详细介绍请参考项目附件二。

为了保证判别结果的真实性与可信性，对于每一条招聘信息，都至少有两个人进行判断，且主要依据判断规则进行。如果两人判断结果不一，将会有第三个人参与判断，或原来两个人再做商议，确定最终判断结果。

2.1.3. 数据再分析

首先我们对我们人工标记获得到的 434 条虚假信息进行了统计：



（上图为对数据集中的虚假信息分析的结果）

可以看到上海地区的在线招聘欺诈情况比较严重。其次是广州。

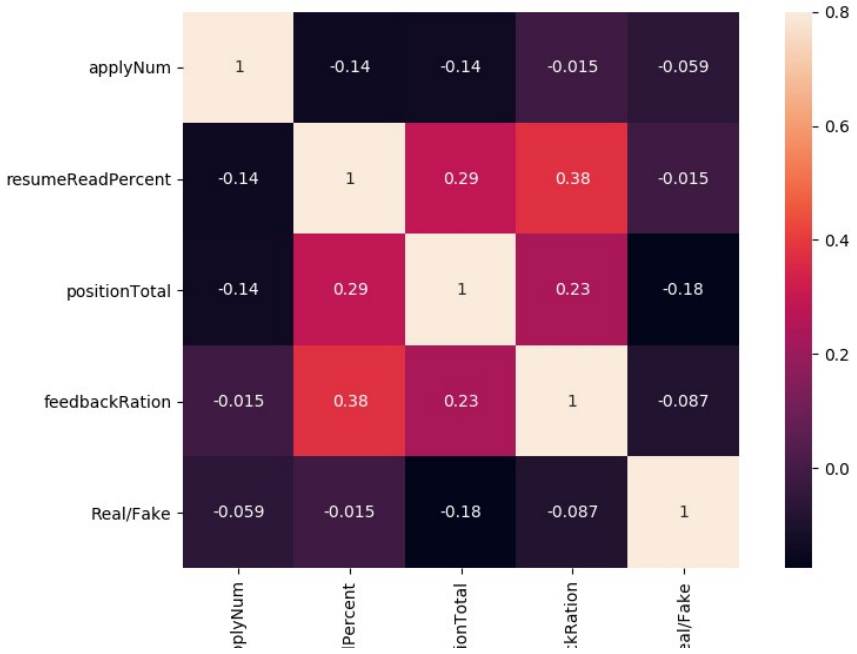
1) 对异常数据剔除

因为有些信息存在部分字段的缺失等，而最终进行测试或训练的数据必须是完整的，因此对于这些异常数据，即变量缺失度很高且变量对输出结果影响很小，则可以直接剔除（即删除列），由于我们在数据采集的过程中对缺失数据进行了严格的控制，所以这一步骤我们只删除了一些彻底缺失的特征列，并没有对数据行进行剔除。

2) 分析样本属性的相关性

对于数据集的各个字段，首先进行可视化数据分析，。最终根据可视化的结果得出，数据的属性之间并没有相关性。

我们对可能存在相关性的属性进行了热力图分析，如图：



（上图为对可能存在相关性的属性热力图分析）

可以看到没有相关性很强的属性

3) 属性的剔除、结合与分割

属性剔除：对输出结果无影响或者影响过小的剔除，此过程结合异常数据过滤阶段同步进行。

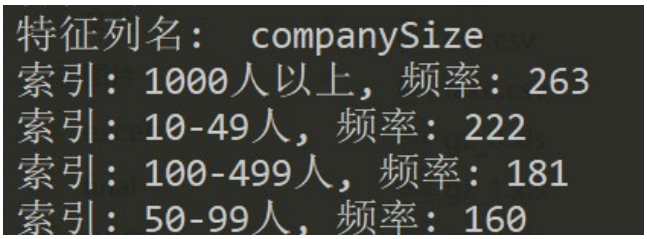
结合过程 1 异常数据剔除，最终的剔除结果为：

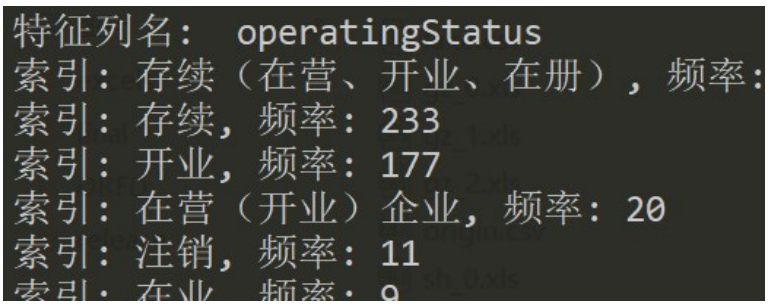
```
## 剔除的属性
FILTER = [
    "Company_ID", "companyBaseUrl", "companyTrade", "companyType",
    "teamTime", "entUrl", "companyDetailUrl", "Job_ID", "browserNum",
    "jobUpdateTime", "jobDetailUrl", "companyName", "estiblishDate",
    "companyAddr", "jobAddr", "regAuthority", "businessScope",
]
```

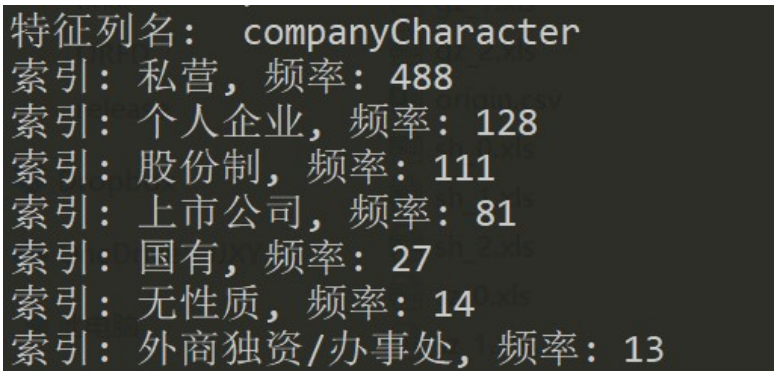
（上图为最终被剔除的属性名称，共 17 个）

属性结合：两属性之间相关性过大，可以将两属性转换为一个属性，且保留解释性较强的那一个，根据过程 2 的分析结果可以看出没有需要结合的属性。

属性分割：对于两个或多个属性间相关性较强且无法进行属性结合的属性可以对其进行分割，分割为多个解释性强的属性，在这一步骤我们主要通过 One-Hot 编码对单类别与多类别文本进行编码以及通过关键词匹配从文本描述中提取二进制向量，如图：







（上面三张图展示的是部分类别属性及其在原始平衡数据集之中出现的频率）

1. companyIntro
   - 有无
   - 是否过短（描述长度统计值）
   - 公司描述是否里提及招聘 **
2. jobDescription
   - 是否过短（描述长度统计值）
   - 无用信息是否过多（各种处理之前长度 - 各种处理之后长度）
   - 描述是否提到联系方式 **
   - 感叹号及其他特殊元素数目
   - 描述是否出现职业名称和标准描述五大要素
3. jobWelfare
   - 描述是否出现联系方式关键词

（上图展示的为通过关键词提取二进制向量的规则）

4) 文本属性分析

我们对进行过简单处理的文本属性进行了分析，确定了根据文本属性的内容需要对其进行编码的方法，首先公司描述对结果影响不大，故对文本的处理忽略它，而职位描述、职位标题、职位子标题很重要所以我们决定使用 Bert 预训练模型对其进行编码。

2.1.4. 特征工程

1) 数据预处理

数据再分析过程中所分析的过程，在此步骤都进行了代码实现，将特征处理为可以输入特征构建模块的格式。

除了对特征的处理，我们还对文本数据进行了单独处理：由于通用的停用词并不适用于招聘领域的文本，而且经过实验我们发现停用词匹配的概率非常的小，所以我们略去了停用词过滤这一个步骤，取而代之的是对职位描述段落进行分句，具体流程如下：

1. 我们以正则表达式为基础制定了一个分句的规则
2. 我们对所有训练数据中的职位描述进行分句

3. 根据分句结果统计了，句段平均长度（7.8 取为 8）

```
原始句段最大长度 89.0
原始句段平均长度 7.817747299920222
原始句段最小长度 3.0
```

（初步分析原始职位描述文本）

4. 再根据句端平均长度重新进行分句，分句结果去掉长度小于平均长度的句段

5. 重新统计了过滤之后的句段平均长度（14.4 取为 15），标准句段长度设为 8/2 + 15/2 = 11.5 取为 12，句段总长度出现的频率，频率相同的有三个，我们选取了最短的那一个，并且将其改为 168，因为 168 可以被 12 整除

```
原始句段最大长度 89.0
原始句段平均长度 14.411832014949342
原始句段最小长度 0.0
```

（根据句段平均长度过滤之后分析原始职位描述文本）

```
句段总长度出现频率:
0       19
384     12
155     12
286     12
10      11
197     11
863     10
78      10
578      9
22       8
34       7
```

（统计处理之后句段总长度频率）

6. 最后我们将标准句段长度设为 168，标准句段数目设为 168/12

2) **特征构建**

a) 纯数字及文本数字：处理之后直接使用

b) 类别类文本：包括单文本类别及多文本类别，进行 One-Hot 编码

c) 关键词匹配 bool 向量：使用我们自己构建的针对于招聘领域的关键词词库以及正则表达式匹配，匹配结果向量作为特征，图为我们的关键词词库及正则表达式：

名称	修改日期	类型	大小
comp_intro.txt	2019/5/19 星期日 2...	文本文档	1 KB
contact.txt	2019/5/25 星期六 1...	文本文档	1 KB
job_intro.txt	2019/5/26 星期日 1...	文本文档	1 KB
require.txt	2019/5/26 星期日 1...	文本文档	1 KB
salary.txt	2019/5/19 星期日 1...	文本文档	1 KB
welfare.txt	2019/5/25 星期六 2...	文本文档	5 KB
work_time.txt	2019/5/19 星期日 1...	文本文档	1 KB

（我们自己构建的针对于招聘领域的关键词词库）

```
## 正则工具
PATTERNS = {
    # "segment": r'\W|\d|\s',
    "segment": r'[\u3002\uff1b\uff0c\uff1a\u3001\uff1f\uff01\u201c\u201d\uff08\uff09\u300a\u300b【】\{\}\(\)?,]*',
    "digital": r'([0-9]*\.?[0-9]+)',
    "contact": r'(\d{1,11})+.?(手|call|机|号|联|联系|拨打|拨|讯|微.{1}信|薇.{1}信|威.{1}信|q|Q|QQ|Qq|qq|qQ|话|电)+.*
    "email": r'([0-9a-zA-Z]+\.*)([0-9a-zA-Z]+)\@+[0-9a-zA-Z]+\.[0-9a-zA-Z]{3}(\.[0-9a-zA-Z]{2}|)',
    "url": r'([http|ftp|https])?(:\/\/)?([0-9a-zA-Z]*)?\.?([0-9a-zA-Z]+\.[0-9a-zA-Z]+)',
    "work_time": r'(2[0-4]|[0-1][0-9]|[0-9])[:时]{1}([0-5][0-9]|[0-9])',
}
```

（正则工具）

d) 文本向量：我们将处理之后的职位描述与职位标题、子标题连接在一起，输入 Bert 模型进行句子编码，输出向量作为此招聘信息的文本向量。

3) 特征提取与特征选择

由于我们的数据集几乎没有冗余信息，我们并没有对特征进行提取与选择

2.1.5. 训练模型

我们以数据集中虚假信息数目（434）为基准，从数据集中选取了 868 条数据作为原始平衡数据集，并将从中选取 800 条作为原始训练集，68 条作为原始测试集；为了得到最优模型，首先使用标准参数对不同的模型（六种）进行训练，得到训练结果之后对训练结果进行评估对比选出最优模型。

使用的六种训练模型分别是：逻辑回归、随机森林、决策树、Bagging、GBDT 及 Ada

而模型评估的标准主要有 Accuracy，Precision，Recall，ROC_AOC，F1 以及 Fit time

2.1.6. 模型评估与优化

为了提高模型的泛化能力，在模型训练及最优模型选取阶段我们采用了 K-Fold 交叉验证原则；

在对最优模型参数优化阶段我们采用网格搜索优化的方式；

在模型测试阶段，我们选择在原始测试集上对模型优化前与优化后分别进行测试，而且也测试了优化后模型在交叉验证评估的表现；

最后我们在完全不平衡数据集上对模型的泛化能力进行了测试，选取数据虚假与真实比例为：1/6

2.2.实现原理

2.2.1. 分类模型介绍

（1） 逻辑回归

逻辑回归是一种有监督的统计学习方法，主要用于对样本进行分类。

在线性回归模型中，输出一般是连续的，例如

$$y=f(x)=ax+b \quad y=f(x)=ax+b$$

对于每一个输入的 x，都有一个对应的 y 输出。模型的定义域和值域都可以是[-∞, +∞]。但是对于逻辑回归，输入可以是连续的[-∞, +∞]，但输出一般是离散的，即只有有限多个输出值。例如，其值域可以只有两个值{0, 1}，这两个值可以表示对样本的某种分类，高/低、真/假、阴性/阳性等，这就是最常见的二分类逻辑回归。因此，从整体上来说，通过逻辑回归模型，我们将在整个实数范围上的 x 映射到了有限个点上，这样就实现了对 x 的分类。因为每次拿过来一个 x，经过逻辑回归分析，就可以将它归入某一类 y 中。

（2） 随机森林

随机森林就是通过集成学习的思想将多棵树集成的一种算法，它的基本单元是决策树，而它的本质属于机器学习的一大分支——集成学习（Ensemble Learning）方法。随机森林的名称中有两个关键词，一个是“随机”，一个就是“森林”。“森林”容易理解，一棵叫做树，那么成百上千棵就可以叫做森林了，这样的比喻还是很贴切的，其实这也是随机森林的主要思想--集成思想的体现。“随机”的含义我们会在下边部分讲到。

其实从直观角度来解释，每棵决策树都是一个分类器（假设现在针对的是分类问题），那么对于一个输入样本，N 棵树会有 N 个分类结果。而随机森林集成了所有的分类投票结果，将投票次数最多的类别指定为最终的输出。

2.2.2. 文本编码模型-Bert

Bert 由多个 Transformer 模型的 Encoder 连接而成，它结合了 ELMo 语言模型与 Transformer 编码器模型的优点。

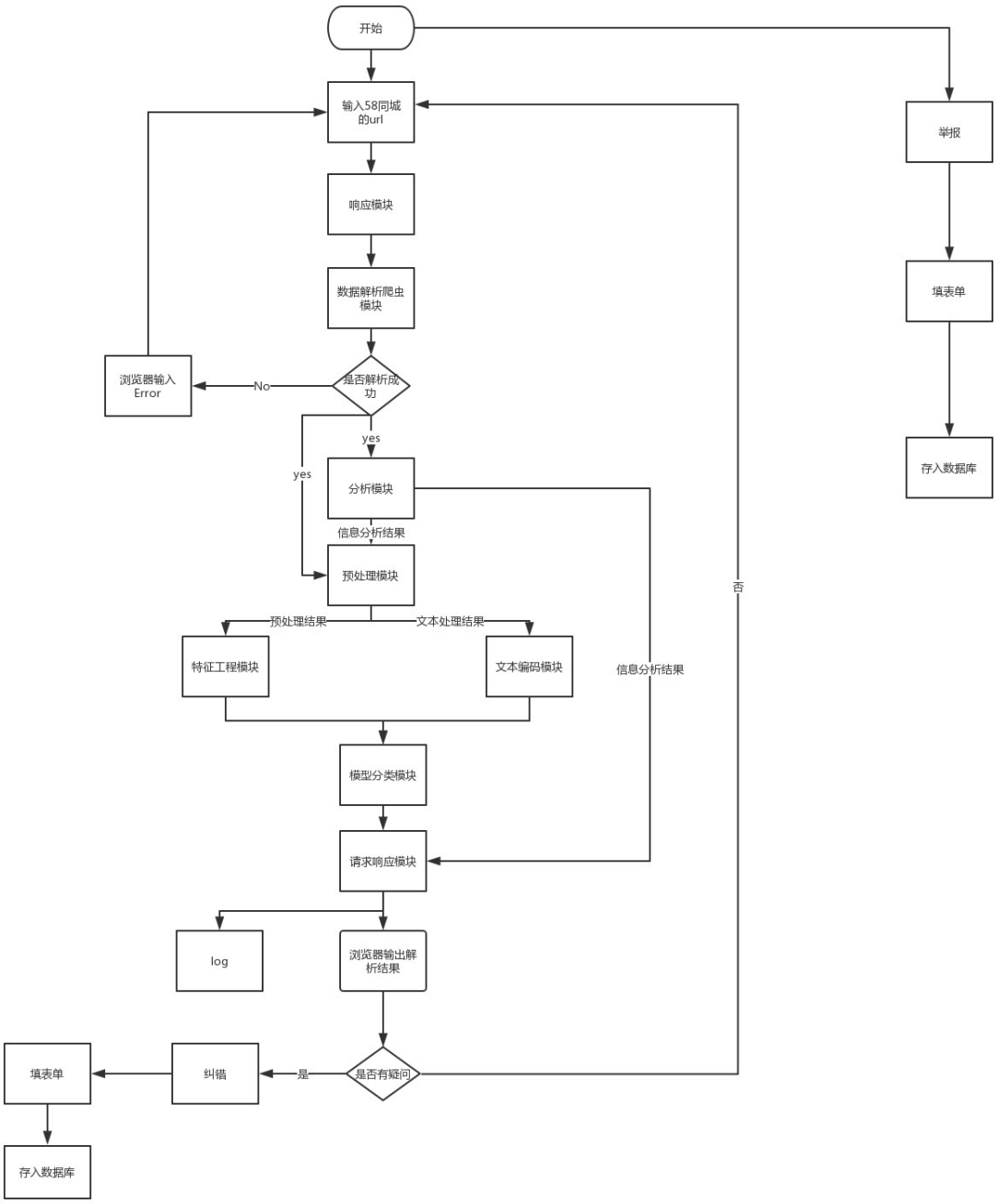
在此基础上，对 Bert 模型进行两过程预训练：第一个任务是随机地扣掉 15%的单词，用一个掩码 MASK 代替，让模型去猜测这个单词；第二个任务是，每个训练样本是一个上下句，有 50%的样本，下句和上句是真实的，另外 50%的样本，下句和上句是无关的，模型需要判断两句的关系。

这两个任务各有一个 loss，将这两个 loss 加起来作为总的 loss 进行优化。训练结束后的 Transformer 模型，包括它的参数，称为通用语言表征模型。

Berth 优势

- 通用性好: 由于 BERT 预训练阶段采取的两阶段过程, 并且 BERT 支持迁移学习, 使得绝大部分 NLP 任务都可以使用 BERT 来提升效果
- 减少开销: BERT 预训练通过超大数据、巨大模型、和极大的计算开销训练而成. 训练这样一个通用的模型需要非常大的开销, 然而谷歌官方已经提供了多个适用于不同场景及不同类型的 BERT 预训练模型, 其中包括中文
- 模型优越: ELMo 的语言模型是双向的, 但是模型无法预训练,也就是无法进行微调迁移, 而 Transformer 虽是可微调的预训练模型,但它是单向的. BERT 结合了 ELMo 模型和 Transformer 模型的优越之处.
- 相比以往的词嵌入技术, BERT 考虑了上下文, 包括词与词, 句与句之间的关系, 使得 BERT 可以在以往的词嵌入基础之上来创建语境化的词嵌入.
- 效率高: 不需要分词，不需要自己训练或者下载词向量，不需要 Bi-LSTM，就那么粗鲁的把 BERT 的微调任务数据改成我们的数据，就可以得到比条件随机场好得多的结果。

2.3.软件流程



（网站平台流程图）

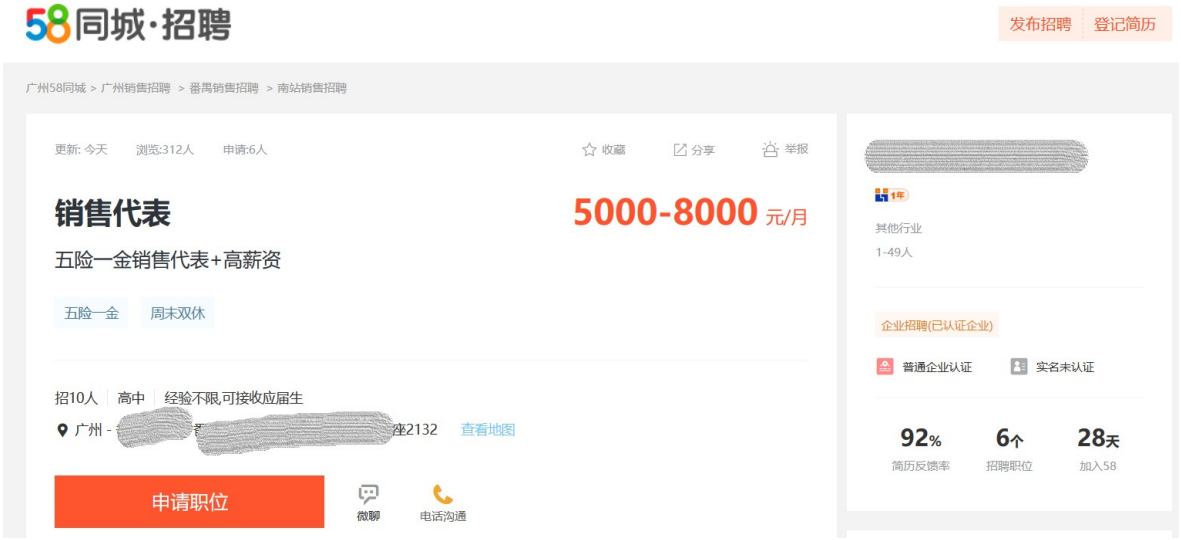
2.4.功能

本作品的主要功能是对招聘信息的真假进行判断，用户需提供所需鉴别的求职信息所在的网址，系统可自动获取求职信息，并对求职信息进行判断。判断结果系统会反馈给用户，并且为了进一步优化模型，我们还提供了检测结果纠错、招聘信息举报以及检测结果自动记录功能。

具体演示过程如下所示：

2.4.1. 选择链接

打开你需要检测的 58 同城招聘详情页链接



（任意 58 同城招聘详情页）

2.4.2. 检测

复制链接到检测页面，在输入框输入链接，点击按钮

（输入招聘链接检测）

稍等几秒，就会收到检测结果

（检测与分析结果）

2.4.3. 纠错

假如你对检测结果不满意可以选择结果纠错，提交表单

（纠错）

（纠错表单）

2.4.4. 举报

假如你想举报一个招聘信息，也可以选择举报，提交表单

（举报）

（举报表单）

2.4.5. 说明

其中，当最终得分大于或等于 50 分时，检测结果为真，反之为假；左边的表格为对职位描述文本的分析；右边的表格为对公司介绍文本的分析。

值得说明的是,对于在原本的详情页可以看得到的信息且无需进一步分析的信息，我们并没有搬运过来（以防侵权），我们尽力让网页显示的友好、简洁，让操作一键化、人性化。

2.5.指标

在模型训练阶段，评价模型优劣的指标有：而模型评估的指标主要有Accuracy，Precision，Recall，ROC_AOC，F1以及Fit time

首先我们在平衡数据集上面对选取的五种分类器进行了性能评估，以下为评估结果：

2.5.1. 特征向量分类器

（1）在标准参数下，采用 k-fold 交叉验证策略（k = 10）将数据集分为训练子集和交叉验证子集，结果如图 1 所示：

分类器	Accuracy	Precision	Recall	ROC_AUC	F1	Fit Time
Random Forest	0.835	0.843	0.835	0.835	0.833	0.016
Bagging	0.822	0.826	0.822	0.822	0.821	0.049
Gradient Boosting	0.821	0.823	0.821	0.821	0.821	0.181
Ada Boost	0.775	0.776	0.775	0.775	0.775	0.094
Decision Tree	0.770	0.774	0.770	0.770	0.869	0.007
Logistic Regression	0.711	0.714	0.711	0.711	0.710	0.041

（图 1：标准参数-平衡数据集-k_folder-特征向量分类模型评估）

可以看到 Random Forest 的表现最好。

（2）针对上面的评估结果，我们单独对 Random Forest 进行了参数网格优化，对优化后的模型进行了交叉验证，结果如图 2 所示：

分类器	Accuracy	Precision	Recall	ROC_AOC	F1
Random Forest	0.874	0.876	0.874	0.874	0.874

（图 2：优化后参数-平衡数据集-k_folder-特征向量随机森林模型评估）

可以看到模型在交叉验证评估取得的成绩得到了比较好的提升

2.5.2. 文本向量分类器

与特征向量类似，对于文本向量分类模型的评估，我们依旧在标准参数下，采用 k-fold 交叉验证策略（k = 10）将数据集分为训练子集和交叉验证子集，结果如图 3 所示：

分类器	Accuracy	Precision	Recall	ROC_AUC	F1	Fit Time
Logistic Regression	0.760	0.762	0.760	0.760	0.760	0.513
Gradient Boosting	0.745	0.748	0.745	0.745	0.745	5.184
Ada Boost	0.696	0.699	0.696	0.696	0.696	3.111
Bagging	0.678	0.682	0.678	0.678	0.675	2.663
Random Forest	0.691	0.698	0.691	0.691	0.691	0.131
Decision Tree	0.642	0.643	0.642	0.642	0.642	0.457

（图 3：标准参数-平衡数据集-k_folder-文本向量分类模型评估）

我们对逻辑回归模型进行了参数网格优化，但是发现参数优化的结果和标准参数的结果几乎一致，具体原因可能是由于训练集的规模太小而数据本身的维度太高导致的结果。

第三章 作品测试与分析

3.1.测试方案

3.1.1. 测试原始测试集

这里需要强调一点，由于我们的测试集规模较小，只有 68 条，所以我们也对比了优化之前以及优化之后的模型对测试集的性能，

（1） 使用原始测试集测试 Random Forest 模型，结果如图 4 所示

分类器	Accuracy	Precision	Recall	ROC_AOC	F1
Random Forest	0.853	0.875	0.824	0.853	0.848

分类器	Accuracy	Precision	Recall	ROC_AOC	F1
Random Forest	0.912	0.967	0.853	0.912	0.906

（图 4：优化后参数-平衡数据集-测试集-特征向量随机森林模型评估）

其中图中表一为未优化模型在测试集上的表现；表二为优化之后模型在测试集上的表现。

（2）Random Forest 模型在测试集上的混淆矩阵如图 5 所示：

混淆矩阵

```
1 tn, fp, fn, tp = (33, 1, 5, 29)
```

Real	Fake	实际类别
29	5	Real
1	33	Fake

（图 5：随机森林测试集混淆矩阵）

综合图 4 和图 5 可以看到在测试集上，优化之后的 Random Forest 模型取得了很好的效果；

（3） 逻辑回归模型在测试集上的表现与混淆矩阵如图 6 所示：

分类器	Accuracy	Precision	Recall	ROC_AOC	F1
Logistic Regression	0.838	0.818	0.882	0.838	0.845

测试集混淆矩阵

```
1 tn, fp, fn, tp = (27, 7, 4, 30)
```

Real	Fake	实际类别
30	4	Real
7	27	Fake

（图 6：优化后参数-平衡数据集-测试集-文本向量逻辑回归模型评估）

其中图中表一为模型在测试集上的评分；表二为混淆矩阵

3.1.2. 测试完全不平衡数据集

在完全不平衡数据集（1:6）上面，我们对我们刚刚训练的随机森林模型进行了测试，结果如图 7 所示：

分类器	Accuracy	Precision	Recall	ROC_AOC	F1
Random Forest	0.904	0.990	0.897	0.922	0.941

混淆矩阵

```
1 tn, fp, fn, tp = (411, 23, 268, 2336)
```

Real	Fake	实际类别
2336	268	Real
23	411	Fake

（图 7：优化后参数-完全不平衡数据集-特征向量随机森林模型评估）

其中图中表一为模型在完全不平衡数据集上的评分；表二为混淆矩阵

3.2.结果分析

（1） 对图 2 和图 4 进行分析可以得出如下结论：

优化之后的随机森林模型在测试集上对特征向量分类的表现相对来说很好，相比未优化模型的表现提升很多。

（2） 对图 6 进行分析可以得出如下结论：

逻辑回归模型在测试集上对文本向量分类上面表现良好，原因可能是测试集规模比较小而数据维度高的原因，但是由于数据采集与人工判别需要占用过多资源，我们暂时没有能力使用更大的数据集对其进行训练，后续可能会继续进行相关工作。

（3） 对图 5 和图 6 进行分析可以得出如下结论：

随机森林的特征向量分类模型倾向于将信息分类为假，而逻辑回归的文本向量分类模型更倾向于将信息分类为真，得出的这个结论恰好与招聘信息的现实情况相吻合，很多虚假招聘信息往往在文字描述上面试图以假乱真，而真实的特征数据却是虚假招聘发布者无法左右的。

（4） 对图 7 分析可以得出如下结论：

在完全不平衡数据集上面，随机森林的表现依然良好，这说明我们训练得到的模型泛化能力很强，在面对大量未知数据依旧可以保持很好的性能，这也证明了我们的研究是有效的，并且是有应用前景的。

综合以上测试评估，我们可以得到如下结论：两个分类器的正确率都在 0.8－0.9 的范围内波动，这说明每 10 条信息就可能有 1－2 条被误判。另外值得说明的是，受限于现有的资源，我们训练所使用的数据集还是太少，假如要进一步的优化模型，就需要我们以及对这个项目感兴趣的人们继续努力了。

第四章 创新性说明

在理论方面，本作品基于机器学习方法，提出一种在线检测与分析招聘信息真假的方案，发明了特征与文本相结合的评估方法以及一系列针对于招聘领域信息的分析方法，将机器学习完美的应用在了招聘信息保护领域；

在应用方面，本作品基于机器学习训练了两个性能非常好的分类器模型，并提供了在线检测招聘信息真假的接口，以此接口为核心实现了一个基于 B/S 架构的在线招聘欺诈检测平台；

在其他方面我们还提供了可用于进一步对虚假招聘研究的数据集以及针对于招聘领域的关键词词库，并开源了相关代码。

第五章 总结

在本文中，我们从项目流程为主干详细的介绍了在线招聘欺诈检测的思路方案与项目实现过程以及相关原理，并对项目结果进行了严格的分析。本文介绍的思路方案可作为对招聘领域进一步研究的基础，我们公布了我们使用的数据集以及相关代码，非常希望我们的工作可以进一步引发和推动虚假招聘领域的相关研究。

本作品的创新在于在此之前并没有可检测网络招聘信息真假的方案，而有些招聘网站中的招聘信息存在着很严重的欺诈问题，这些虚假的招聘信息可能会对求职者造成巨大的损失或伤害，目前的招聘网站上虽然对招聘信息的发布提出了相对严格的控制，但是还是有些虚假信息躲过了招聘网站的初步筛选，而一旦发布在招聘网站上，就可能发生一些难以控制的危害。而对于普通人来讲，仅仅根据网站提供的招聘信息，有时也很难判断是真是假，因此实现一种可以检测招聘信息真假的方案来帮助求职者是一项非常有意义的工作。

关于判断招聘信息真假原理，我们采用了机器学习的方法，用平衡数据集对于不同的分类模型进行训练，最终得出最优模型。对于信息中的主要特征向量采用随机森林方法训练的模型在所有模型中分类的效果最好，而对于信息中的文本向量采用逻辑回归训练的模型在所有训练模型中表现最好。前端接受用户输入的招聘信息详情链接

传送给后端，后端则根据链接解析到详细的招聘信息，对招聘信息检测分析之后，将所有结果返回给前端，前端完成对用户的反馈，即完成了招聘信息真假检测与分析的整个过程。

从作品应用以及发展前景看，本作品有着巨大的市场需求，因为网络招聘成为一种越来越重要的招聘方式，而大多数的求职者都无法直接判断他们所看到的招聘信息的真假，而在线检测系统可给求职者一个相对可靠的判断依据。

整个作品采用 B/S 架构，在用户交互方面，本作品方便快捷操作简单，仅需用户提供招聘信息的网址，本系统就能返回判断结果：包括分析结果、检测结果及招聘信息分数，并且提供了纠错与举报的功能，根据用户的反馈与举报，增加训练样本，从而提高分类系统的可靠性。

但是本作品还有些问题是我们应当正视的，首先即是训练数据集的大小，我们最终采用的虚假招聘信息的数据集仅有400余条，这对于检测所有类型的招聘信息是远远不够的，但是由于时间精力有限，我们目前能获得的打了虚假标记的招聘信息也仅有这么多。但是本作品的系统提供了比较妥善的解决方案，即系统的鉴别能力会随着用户的增多而变得更加强大，系统会根据用户的举报与反馈增加新的训练样本来训练模型，最终使得模型更加可靠，使得判断结果更加准确。

参考文献

[1]Sokratis Vidros, Constantinos Kolias, Georgios Kambourakis. Online recruitment services: another playground for fraudsters. Computer Fraud & Security, Volume 2016, Issue 3, 2016, Pages 8-13.

[2]Vidros, Sokratis (Department of Information and Communication Systems Engineering, University of the Aegean, Karlovassi, Samos; 83200, Greece), Kolias, Constantinos, Kambourakis, Georgios. Akoglu, Leman Source: Future Internet, v 9, n 1, March 3, 2017

附件

附件一

58 同城数据库详细说明及分析

数据表格式

所有信息被分为三个表：company、recruitment、issue

- company：只与公司相关的信息字段
- recruitment：网页上招聘详情页只与招聘相关的信息字段
- issue：公司与招聘职位的联系集——发布，公司发布职位，存放两者相关联的信息字段

三个表的详细请看下面三个图：

Name: company

Columns: Add Remove Up Down

#	Name	Datatype	Length/Set	Unsign...	Allow N...	Zerofill	Default	Comment	Collation
1	Company_ID	VARCHAR	200	☐	☐	☐	No default		utf8mb4_ge
2	positionTotal	INT	10	☐	☑	☐	NULL		
3	companyBas...	VARCHAR	1000	☐	☑	☐	NULL		utf8mb4_ge
4	companyNa...	VARCHAR	200	☐	☑	☐	NULL		utf8mb4_ge
5	companyTra...	VARCHAR	200	☐	☑	☐	NULL		utf8mb4_ge
6	companyCha...	VARCHAR	100	☐	☑	☐	NULL		utf8mb4_ge
7	feedbackRat...	INT	10	☐	☑	☐	NULL		
8	companySize	VARCHAR	100	☐	☑	☐	NULL		utf8mb4_ge
9	companyAddr	VARCHAR	300	☐	☑	☐	NULL		utf8mb4_ge
10	companyIntro	VARCHAR	2000	☐	☑	☐	NULL		utf8mb4_ge
11	businessSco...	VARCHAR	2000	☐	☑	☐	NULL		utf8mb4_ge
12	companyType	VARCHAR	100	☐	☑	☐	NULL		utf8mb4_ge
13	creditCode	VARCHAR	100	☐	☑	☐	NULL		utf8mb4_ge
14	estiblishDate	DATETIME		☐	☑	☐	NULL		
15	operatingSt...	VARCHAR	100	☐	☑	☐	NULL		utf8mb4_ge
16	orgNumber	VARCHAR	100	☐	☑	☐	NULL		utf8mb4_ge
17	regAddress	VARCHAR	200	☐	☑	☐	NULL		utf8mb4_ge
18	regAuthority	VARCHAR	200	☐	☑	☐	NULL		utf8mb4_ge

Name: recruitment

Columns: Add Remove Up Down

#	Name	Datatype	Length/Set	Unsign...	Allow N...	Zerofill	Default	Comment	Collation
1	Job_ID	VARCHAR	200	☐	☐	☐	No default		utf8mb4_genera
2	jobTitle	VARCHAR	100	☐	☑	☐	NULL		utf8mb4_genera
3	jobSubTitle	VARCHAR	200	☐	☑	☐	NULL		utf8mb4_genera
4	jobSalary	VARCHAR	100	☐	☑	☐	NULL		utf8mb4_genera
5	applyNum	INT	10	☐	☑	☐	NULL		
6	resumeRead...	INT	10	☐	☑	☐	NULL		
7	browserNum	VARCHAR	100	☐	☑	☐	NULL		utf8mb4_genera
8	jobWelfare	VARCHAR	200	☐	☑	☐	NULL		utf8mb4_genera
9	jobRequire...	VARCHAR	100	☐	☑	☐	NULL		utf8mb4_genera
10	jobAddr	VARCHAR	200	☐	☑	☐	NULL		utf8mb4_genera
11	jobDescripti...	VARCHAR	2000	☐	☑	☐	NULL		utf8mb4_genera

Name: issue

Columns: Add Remove Up Down

#	Name	Datatype	Length/Set	Unsign...	Allow N...	Zerofill	Default	Comment	Collation
1	Company_ID	VARCHAR	200	☐	☐	☐	No default		utf8mb4_gener
2	Job_ID	VARCHAR	200	☐	☐	☐	No default		utf8mb4_gener
3	jobUpdateTi...	DATETIME		☐	☑	☐	NULL		
4	needNumber	VARCHAR	100	☐	☑	☐	NULL		utf8mb4_genera
5	jobDetailUrl	VARCHAR	1000	☐	☑	☐	NULL		utf8mb4_genera

数据表字段详细介绍如下：

1. company

- Company_ID：公司 ID
- positionTotal：公司发布职位数，这里有多少，招聘表和发布联系集关于这个公司的信息数目就有多少
- companyBaseUrl：公司 58 同城基本信息页
- companyName：公司名字
- companyTrade：公司行业，由于 58 同城网页改版，原本这个字段是有数据的，现在全是未知
- companyCharacter：公司性质：民营、国企等等
- feedbackRation：公司反馈率，这个对公司发布的所有职位都是相同的，代表没收到一份简历投递这个公司对这个简历进行标记操作处理的概率，标记操作就是查看简历修改简历的状态为：可面试、不合适、待定等等状态。也可以理解为，100 个人投递这家公司，其中有 100 * feedbackRation 人收到了这家公司的回应
- companySize：公司规模，也就是员工规模，如 100-499 人
- companyAddr：公司详细地址
- companyIntro：公司介绍

- companyType：同样由于 58 改版，现在全为未知
- teamTime：经营期限，我全设为了"_"，其实可以默认为成立日期-至今，但是假如这样默认很不严谨
- entUrl：公司天眼查网址，受改版影响，现在无法获取，全为未知
- companyDetailUrl：这个是公司自己提供的官方网址，一般没有
- businessScope：公司经营范围
- creditCode：公司统一社会信用代码
- estiblishDate：公司成立日期，假如公司成立日期不存在，都将其设为：1800-12-31
- operatingStatus：公司经营状态：开业、存续等等
- orgNumber：组织机构代码
- regAddress：注册地址
- regAuthority：注册机构
- regCapital：注册资本

2. recruitment

- Job_ID：职位 ID
- jobTitle：工作标题，一般较为正常，描述职位的行业或者特定名称，这个标题应该是公司按照 58 要求设定的
- jobSubTitle：工作子标题，一般不太正常，这个子标题应该为公司自定义，一般为描述工资、福利等等信息，是个很重要的字段
- jobSalary：薪资，一般为一个范围
- applyNum：申请人数，即投递简历人数
- resumeReadPercent：简历阅读百分比，这个和反馈率类似，顾名思义表示投递一个简历被阅读到的概率，在 58 同城网页上面，有的这个职位的简历阅读百分比，有的显示公司的反馈率，显示规则不清楚。但是这个字段和反馈率关系密切，上面说了反馈率是阅读之后再对简历进行标记，而阅读率仅仅是阅读，既包括被标记的简历也包括未被标记的，所以一般阅读率 >= 反馈率，这个规律从数据库中也看得出
- browserNum：招聘信息浏览量，58 同城可能对这个数字的加载进行了 AJAX

处理，当初写代码的时候没注意到，所以现在所有浏览量都为 0

- jobWelfare：工作福利，格式为：包吃_ 包住 _ 节日福利 _ 员工聚餐 _ 近地铁
- jobRequirement：工作要求，为学历和工作经验要求，格式为：学历不限_经验不限
- jobAddr：工作详细地址，格式为：广州_ 白云_ 嘉禾望岗_ 广州白云嘉禾望岗空港大道润林大厦 506
- jobDescription：工作描述
- jobCity：爬取信息所选的城市：bj、sh、gz、sz

3. issue

- jobUpdateTime：公司更新这条职位的时间，只精确到天，其实详情页可以精确到分钟，但是我没有存
- needNumber：公司发布这条职位提供的岗位数目
- jobDetailUrl：职位招聘详情页链接

附件二

判断规范

1. 每一条信息后面都会有 **Real/Fake** 这一字段，真就填 1，假就填 0
2. 判断结束或者定期，判断相同数据的两个人要对比各自的判断结果，有争议的结果要进行讨论，如果两个人讨论还是不可以得出结果，那就把 **Real/Fake** 这一字段置为空

首先剔除对判断无效无用的数据，然后制定判断规则。

1. 去除无用属性

打对号的要留下的属性，没打对号是要被筛除的属性

company-58

- ☐ Company_ID
- ☑ positionTotal
- ☐ companyBaseUrl
- ☑ companyName
- ☐ companyTrade
- ☑ companyCharacter
- ☑ feedbackRation
- ☑ companySize
- ☑ companyAddr
- ☑ companyIntro
- ☐ companyType
- ☐ teamTime
- ☐ entUrl

- ☐ companyDetailUrl

company-ent

- ☑ businessScope
- ☑ creditCode
- ☑ operatingStatus
- ☑ orgNumber
- ☑ regAddress
- ☑ regAuthority
- ☑ regCapital

recruitment

- ☐ Job_ID
- ☑ jobTitle
- ☑ jobSubTitle
- ☑ jobSalary
- ☑ applyNum
- ☑ resumeReadPercent
- ☐ browserNum
- ☑ jobWelfare
- ☑ jobRequirement
- ☑ jobAddr
- ☑ jobDescription

- ☐ jobCity

issue

- ☐ jobUpdateTime
- ☑ needNumber
- ☐ jobDetailUrl

2. 剩余属性的判定规则

issue

属性名	简要说明	异常点
needNumber	职位需求人数，或者成为职位提供岗位数目	无

recruitment

属性名	简要说明	异常点
jobTitle	无	无
jobSubTitle	这个字段最容易看出问题，正常的招聘为对职位的补充描述，或者为职位要求的简要概括	1. 子标题与标题主题不对应；  2. 在子标题里提及工资或其他福利信息； 
jobSalary	由于没有参考的职位平均工资也没办法获得，所以根据工资范围判别不太	3. 除了工资属性，在子标题、职位描述处多处强调工资，并且这几处换算出来的工资各自不相同且差异很大。 如：

	现实	
applyNum	无	无
resumeReadPercent	首先要纠正一个很容易进入的误区，查看简历的意思是下载并查看，而公司即使不下载简历也可以看到你的信息。不能仅仅通过百分比高低就冒然进行判别	4. applyNum < n，n 是一个阈值，大于 n 说明这个招聘收到了很多申请，反之；那么正常的招聘由于申请人数小那么他的阅读百分比一般都要很大，这个规律简单点说就是，随着 n 的减小，阅读百分比逐渐接近于 100； 5. applyNum > n，与 1 相对应，随着 n 的增大，阅读百分比逐渐减小，具体减小到什么程度，取决于这个招聘提供的岗位数目为多少。 综上，加入阅读百分比的值不满足上面两条规律，可以认为异常，n 值我认为可以设置为：needNumber * 5
jobWelfare	俗话说得好：便宜没好货，那么对于职位福利来说，公司可以提供的福利应该是与薪资成正比的，举个例子，一个月 5000 的工资，然后公司除了基本福利还提供很多要公司花钱的额外福利，显然是不可能的。	6. 福利太差，基本的福利都不全的 7. 福利太好，很多收集简历的人员，大多不太了解行业内的福利，他们往往一股脑的提供几乎所有的福利或者提供很多诱人的福利，例如房补、车补、保底薪资之类，还是那句话，一个月 5000 的工资，福利怎么可能会这么好
jobRequirement	学历要求及工作要求，这个要结合职位内容进行判断，你说，假如要找一个会计，然后学历不限，工作经验不限不太现实吧	8. 对于一些显然需要具备一定专业知识的职位，却声称学历不限工作经验不限等等 9. 虽然职位本身确实不需专业知识，可以要求学历不限工作经验不限，但是提供的工资又比较高，很可疑。
jobAddr	一般都为正确地址	无
jobDescription	最关键的属性，一般的职位描述应包括以下几个方面（都类似）： 1. 职责描述：描述职位工作、责任 2. 任职资格/要求：职位要求 3. 工作	10. 提示加其他联系方式直接联系（这种情况很少见，只要见到就是假的） 11. 除了前面说的 5 个方面，其余出现的都是没有用的信息，这类信息出现的越多越可疑 12. 职位描述有明显的煽动性，也就是表达出力求你的加入这种感情

	时间 4. 详细介绍福利 5. 薪资介绍	13. 职责描述里面出现关键词，类似：在家就可以做，做自己的老板，为自己的未来做主，收入高，容易赚钱，无需任何条件等等等等，这里的关键词，请各位在进行判断的时候遇到一个记下一个 14. 职责描述河任职资格中是否关于职责和资格的部分被一笔带过，更加强调福利和薪资 15. 福利和薪资部分是否过于夸张了，或者描述是否太短了

company

属性名	简要说明	异常点
positionTotal	发布职位数，主要用于判别此公司是不是属于广撒网的那种类型	结合其他属性分析
companyName	无	无
companyCharacter	公司性质，公司的营业信息是可以被伪造的，但是对于一些国企类型的公司还是可以放心的	16. 国企企业可以认为是真实的
feedbackRation	和之前的简历查看百分比类似，这里也有误区： 1. 反馈率是对于一个公司收到的所有简历而言的，所以反馈率是作为一个评价公司的属性，而不是用来评价招聘职位的 2. 公司可以在不查看简历的情况下对简历进行标记（反馈），在数据库介绍里提到的查看百分比一般大于等于反馈率，也只是说的一般，万万不可将两者之间的联系看的过于紧密，直白点，两者相差非常大也是有可能的。	17. 当公司发布的所有职位收到了一定数目的简历，那么假如反馈率过小，可以认为这个公司对于投递人员并不感兴趣，很大可能对简历更加感兴趣，有收集简历的可能； 18. 反馈率过大时，说明这个公司对投递人员更加感兴趣，有忽悠人员入职的可能，这也是一种虚假；
companySize	公司规模和发布职位数及职位提供岗位数的关系非常密切。	19. 公司所有职位需求人数加在一起 > companySize x ，其中 x 是一个阈值，代表这家公司的招人极限，取值范围为（0,1），大于 companySize x 说明 公司发布的所有职位需求的人数超出了招人极限，这显然是不现实的，举个例子，一家 50 人的小型公司，在 58 同城仅仅一个招聘网站上就需求 20 人，你觉得可能吗？

companyAddr	无	20. 可以简单对比一下与发布职位的地址的相似度。
companyIntro	有的公司没有介绍	21. 主要看一下有没有介绍、介绍正不正规以及介绍是不是太短了，还有就是公司介绍是不是主要提及了招聘职位信息
businessScope	除国家规定特殊行业如：烟.酒.药瓶.等等 基本啥都能经营。	22. 对于一些规模较小的公司，假如他的经营范围非常的广，这显然是非常不合理的，尽管这些都符合规定，但是不符合经营一家公司的常理，这样也属于异常 23. 经营范围介绍的过于宽泛，不详细，只有大类。
creditCode	无	24. 没有不行
operatingStatus	无	25. 除了开业和存续或者类似的意思，其他属于异常
orgNumber	无	26. 没有不行
regAddress	无	27. 没有不行
regAuthority	无	无
regCapital	无	无

3. 判别过程

先对招聘信息大体浏览自行判断，假如模棱两可，那么就可以参考此判别规则，假如可以确定自己的结果，就无需参考此结果

设总异常点数目为 N = 27

1. 根据异常点标号的顺序对一条信息就行判断即可
2. 异常点判别完成之后，记录信息异常的数目 count，假如 count >= 9，就将结果设为假（即 0），count < 9，就将结果设为真（即 1）

由于信息是按照公司组织的，所以一般判断公司的一定数目的信息的结果相差不大，那么可以判断公司剩余的招聘信息也和之前的判定结果一样