

一、可行性研究

1. 引言

1.1 项目背景

- a. 软件名称：真假职位信息检测；
- b. 软件介绍：本软件根据提供的职位信息来判断这个职位是否真实存在。

2. 可行性研究的前提

2.1 要求

- a. 功能：检测职位信息是否真实存在；
- b. 输出：图表；
- c. 输入：csv 文件输入职位信息；
- d. 基本的处理流程：
首先算法处理，通过 csv 文件内的职位信息得到结果和参数；
然后网页将结果制作成图表展现出来。

二、概要设计说明

1. 任务概述

1.1 目标

检测所输入的职位信息是否真实存在。

1.2 运行环境

Python;
浏览器。

1.3 条件与限制

支持运行环境即可。

2. 数据描述

2.1 静态数据

- a. 根据 csv 文件内的职位信息得到的参数；
- b. 属性名：
['location','industry','title','description_wordsLen','requirements_wordsLen','department','company_profile_wordsLen','function','benefits_wordsLen','description_tfidf46','requirements_tfidf8','salary_range_start','description_tfidf5','required_experience','company_profile_tfidf7','description_tfidf13','required_education','description_tfidf18','requirements_tfidf16','benefits_tfidf1','title_wordsLen','company_profile_tfidf16','requirements_tfidf13','description_tfidf17',

description_tfidf42','description_tfidf44','description_tfidf47','description_tfidf25','description_tfidf36','description_tfidf23','requirements_tfidf19','requirements_tfidf2','description_tfidf0','company_profile_tfidf9','salary_range_end','employment_type','has_company_logo','description_tfidf6','description_tfidf22','benefits_tfidf2','description_tfidf37','description_tfidf38','description_tfidf8','description_tfidf1','description_tfidf19','description_tfidf24','description_tfidf32','company_profile_tfidf2','description_tfidf2','description_tfidf3','requirements_tfidf17','description_tfidf35','description_tfidf7','description_tfidf11','company_profile_tfidf12','description_tfidf31','company_profile_tfidf3','company_profile_tfidf5','description_tfidf33','requirements_tfidf0','requirements_tfidf11','company_profile_tfidf4','description_tfidf34','requirements_tfidf7','description_tfidf41','description_tfidf16','company_profile_tfidf17','requirements_tfidf9','description_tfidf43','has_questions','description_tfidf27','description_tfidf4','description_tfidf9','company_profile_tfidf15','description_tfidf12','requirements_tfidf14','description_tfidf26','company_profile_tfidf14','benefits_tfidf4','description_tfidf15','description_tfidf39','requirements_tfidf10','requirements_tfidf5','description_tfidf29','requirements_tfidf12','requirements_tfidf18','benefits_tfidf7','requirements_tfidf4','benefits_tfidf11','description_tfidf14','description_tfidf20','description_tfidf21','company_profile_tfidf18','benefits_tfidf10','requirements_tfidf15','benefits_tfidf8','requirements_tfidf6','description_tfidf10','company_profile_tfidf13','company_profile_tfidf1','company_profile_tfidf0','benefits_tfidf3','company_profile_tfidf20','company_profile_tfidf23','company_profile_tfidf21','description_tfidf30','description_tfidf40','requirements_tfidf1','requirements_tfidf3','description_tfidf45','benefits_tfidf0','company_profile_tfidf6','benefits_tfidf6','company_profile_tfidf19','description_tfidf28','company_profile_tfidf11','benefits_tfidf5','company_profile_tfidf10','benefits_tfidf9','company_profile_tfidf22','telecommuting','company_profile_tfidf8'];

c. 属性的重要性排序:

['383.6','348.4','276.8','242.4','171.0','151.8','151.6','143.4','110.4','103.8','88.8','83.0','81.8','81.4','75.8','74.4','71.0','70.6','70.2','68.0','65.0','64.8','62.2','61.6','61.2','60.6','59.0','56.4','56.0','54.4','54.0','53.0','50.4','48.8','46.6','44.2','43.0','42.6','39.6','39.6','36.4','35.8','35.4','35.4','34.8','34.2','34.0','33.6','33.0','32.8','32.0','31.6','31.2','30.8','29.6','29.0','27.6','26.6','25.4','25.0','24.6','24.4','23.8','23.6','23.6','23.6','23.4','23.4','23.0','22.6','22.4','22.4','22.2','22.2','21.8','21.8','21.0','20.8','20.8','20.6','20.4','18.8','18.4','18.4','18.0','17.6','17.4','17.4','17.2','17.0','16.8','16.4','16.4','16.4','14.8','13.8','13.2','13.0','12.8','12.8','12.8','11.4','10.6','9.6','9.2','9.0','8.6','8.4','8.0','8.0','7.6','7.4','6.8','6.6','6.6','6.6','5.8','4.2','3.8','3.4','3.2','2.6','2.4'];

d. 各个图表的位置、标题、标签、数值。

2.2 动态数据

输入数据: csv 文件内的职位信息;

输出数据: 根据 csv 文件得到的结果和参数。

3. 界面设计

3.1 用户界面

- a. 屏幕格式：标题、表格组成一个图表，一共由若干个图表组成；
- b. 报表格式：echarts 处理后的 canvas；
- c. 输出：html 文件。

4. 算法设计

先对 csv 文件进行预处理：

- 把字符串类型转换为小写；
- 处理缺失的数据；
- 对具有明显分类的数据添加标签；
- 通过 TF-IDF 处理文本数据；

然后进行模型训练；

- 查看元素数据各个元素的重要性；
- 获得 lgb 模型训练后的得分；
- 对预测结果进行处理。

5. 出错处理设计

5.1 出错处理对策

数据预处理时将非数（nan）打上一个标记，若有非数存在，模型得到的结果不准确。

三、详细设计说明

1. 程序描述

1.1 网页

1.1.1 功能

输出结果（以图表形式）；

1.1.2 输入项目

由 js 文件传入的参数；

1.1.3 输出项目

图表。

1.2 算法

1.2.1 功能

通过模型训练，由职位信息得到参数、真假职位信息的结果；

1.2.2 输入项目

内含职位信息的 csv 文件；

1.2.3 输出项目

参数；

职位信息是否存在的结果（存放在 submissions.csv 文件中）。

四、操作手册

1. 使用说明

1.1 输入

1.1.1 数据格式

a. 顺序

按照 excel 文件的列以 title, location, department, salary_range, company_profile, description, requirements, benefits, telecommuting, has_company_logo, has_questions, employment_type, required_experience, required_education, industry, function, fraudulent 的顺序输入进 csv 文件。

1.1.2 输入举例

title: Office Manager

location: GR, I, Athens

department: Administration

salary_range: 0-130000

company_profile : Workable is a venture-backed startup making cloud-based recruitment software for fast-growing companies around the world. We're looking for people who want to change the way companies and people meet each other. Life at Workable At Workable we are creating an environment that has all the excitement and intellectual stimulation of a startup, minus the fads and pretension. We don't work 80-hour weeks, but we do work in an efficient and disciplined manner. We don't have ninjas and rock stars, we have people who are outstanding at what they do. We don't think it's old fashioned to have a sensible business model and we enjoy working with smart people.> learn more about Workable and our employee benefits

description: We are looking for an experienced office manager to take charge of facilities and personnel administration tasks. The position is full-time with flexible office hours and the opportunity to telecommute regularly or occasionally. You will be supporting a 10-15 staff office. Your responsibilities will include: Making sure your colleagues have a well-functioning, adequately equipped and supplied workplace at all times Performing basic HR administration tasks such as on-boarding new hires, managing employee benefits and staff record-keeping Co-ordinating with our accountants for the prompt and accurate handling of expenses, invoices and other financial records Providing basic reception services and occasionally assisting in customer support tasks Managing appointments and travel for company executives

requirements: Excellent use of English, speaking and writing (we mean it, you have to be native speaker or near-perfect) Firm grasp of office

productivity tools (document preparation, spreadsheets, calendar management)Experience with basic human resource administration tasks (employee contracts, benefits, payroll)Experience with basic financial administration tasks (handling expenses, invoices, etc)Experience with facilities management tasks (office upkeep, bill handling, procurement of equipment and supplies)Experience as an executive PA or head of office administration is a plus

benefits: Our goal is to create a company where employees enjoy benefits that make them more productive and contribute directly to the development of their professional skills. We want to be able to attract the best of the best, and make sure they keep getting better. On top of an exciting, vibrant and intellectually challenging workplace and a competitive salary, we are offering:Private healthcare insurance planCorporate cellphoneHigh quality computing equipmentTraining seminars and materialsAnnual profit sharing planFlexible working hours, ability to telecommute on occasion

telecommuting: 0

has_company_logo: 1

has_questions: 1

employment_type: Full-time

required_experience: Mid-Senior level

required_education: Bachelor's Degree

industry: Computer Software

function: Engineering

fraudulent:0

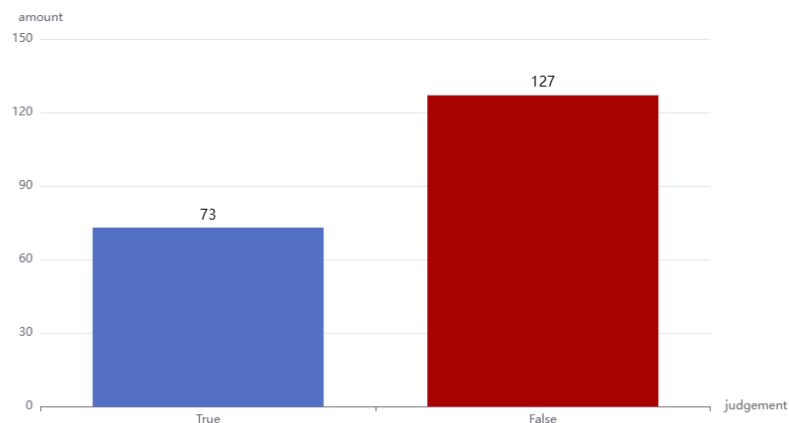
1.2 输出

1.2.1 数据格式

a. 图表;

1.2.2 输出举例

真假职位数量统计



五、测试

使用 doctest:

```
if __name__ == '__main__':  
    app.run()  
  
import doctest  
  
doctest.testmod(verbose=True)
```

结果:

```
2 items had no tests:  
    __main__  
    __main__.index  
0 tests in 2 items.  
0 passed and 0 failed.  
Test passed.
```

使用 pytest-cov 测试代码覆盖率:

```
C:\Users\lhj\PycharmProjects\distinguishJob>python -m pytest ./main.py --cov --cov-report term-missing  
===== test session starts =====  
platform win32 -- Python 3.8.5, pytest-6.2.5, py-1.9.0, pluggy-0.13.1  
rootdir: C:\Users\lhj\PycharmProjects\distinguishJob  
plugins: cov-3.0.0  
collected 0 items  
  
===== warnings summary =====  
D:\Anaconda3\lib\site-packages\pyreadline\py3k_compat.py:8  
  D:\Anaconda3\lib\site-packages\pyreadline\py3k_compat.py:8: DeprecationWarning: Using or importing the ABCs from 'coll  
lections' instead of from 'collections.abc' is deprecated since Python 3.3, and in 3.9 it will stop working  
    return isinstance(x, collections.Callable)  
  
-- Docs: https://docs.pytest.org/en/stable/warnings.html  
  
----- coverage: platform win32, python 3.8.5-final-0 -----  
Name      Stmts  Miss  Cover   Missing  
-----  
main.py       7      2    71%    7, 11  
TOTAL         7      2    71%  
  
===== 1 warning in 1.86s =====
```

覆盖率为 71%