Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ

КАФЕДРА СИСТЕМЫ ОБРАБОТКИ ИФОРМАЦИИ И УПРАВЛЕНИЯ

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

НА ТЕМУ:

Поиск и анализ данных о наборе персонала на основе Python и Selenium

Студент ИУ5И-35М _____ Лю Фэнлинь
(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель _____ Ю.Е. Гапанюк
(Подпись, дата) (И.О.Фамилия)

Консультант _____ _____
(Подпись, дата) (И.О.Фамилия)

2024 г.

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

УТВЕРЖДАЮ

Заведующий кафедрой ___ИУ5__
(Индекс)
_____ _В.И. Терехов__
(И.О.Фамилия)
« _____ » _____ 20 ____ г.

З А Д А Н И Е
на выполнение научно-исследовательской работы

по теме __*Поиск и анализ данных о наборе персонала на основе Python и Selenium*

Студент группы ___ИУ5И-35М___

_____Лю Фэнлинь _____
(Фамилия, имя, отчество)

Направленность НИР (учебная, исследовательская, практическая, производственная, др.)

_____учебная_____

Источник тематики (кафедра, предприятие, НИР) _____

График выполнения НИР: 25% к _5_ нед., 50% к _9_ нед., 75% к _13_ нед., 100% к _17_ нед.

Техническое задание Проблема заключается в том, как точно собрать эффективную ин формацию из разнообразных данных и как извлечь важную ценность из массивных данн ых . В этой статье мы разработали автоматизированный веб-краулер на базе Python Selenium для получения большого количества данных о найме на работу с веб-сайта Wiseleaf, предварительной обработки и визуализации данных для предоставления справо чной информации для соискателей работы, а также для изучения большего количества т а л а н т о в с р е д и м н о г и х с о и с к а т е л е й.

Оформление научно-исследовательской работы:

Расчетно-пояснительная записка на __20___ листах формата А4.
Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.)

_____

_____

Дата выдачи задания «09» __сентября__ 2024 г.

Руководитель НИР _____ ___Ю.Е. Гапанюк____
(Подпись, дата) (И.О.Фамилия)

Студент _____ ___Лю Фэнлинь____
(Подпись, дата) (И.О.Фамилия)

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

Оглавление

Введение

Быстрое развитие информационных технологий в Интернете привело к беспрецедентному росту типа и масштаба данных в человеческом обществе, и все больше людей начинают обращать внимание на огромную ценность, скрытую за этими данными [1]. 2024 год - это год, когда число выпускников университетов относительно велико, а ситуация с трудоустройством очень тяжелая, что привело к быстрому развитию онлайн-рекрутинга в Китае. Однако количество и сложность информации, размещенной на существующих сайтах по подбору персонала, затрудняют поиск желаемой работы для соискателей, а для компаний - поиск подходящих кандидатов среди множества претендентов. Проблема заключается в том, как точно собрать эффективную информацию из разнообразных данных и как извлечь важную ценность из массивных данных [2]. В этой статье мы разработали автоматизированный веб-краулер на базе Python Selenium для получения большого количества данных о найме на работу с веб-сайта Wiseleaf, предварительной обработки и визуализации данных для предоставления справочной информации для соискателей работы, а также для изучения большего количества талантов среди многих соискателей [3].

2. Анализ смежных технологий

2.1 Введение и преимущества языка Python

Python - это бесплатный язык программирования с открытым исходным кодом, который стал очень популярным в последние годы. Это простой, легкий в изучении, гибкий и универсальный язык, который поддерживает вызовы большого количества сторонних библиотек, способных обрабатывать большие объемы данных, таких как библиотека Selenium, которая является универсальным инструментом автоматизации, Numpy, высокопроизводительная библиотека обработки данных, и Pyecharts, быстрая библиотека визуализации данных с отличным интерактивным дизайном.

2.2 Принципы поиска информации в Интернете

Веб-краулеры, как и пауки, ползающие по сети, - это программы, которые могут автоматически просматривать Всемирную паутину в поисках информации в соответствии с заданными правилами [4]. В эпоху больших данных веб-краулеры становятся все более сложными и используются на различных веб-сайтах, где они играют важную роль в анализе данных. Краулеры используются для извлечения данных, получения точной информации, необходимой в большом объеме данных, анализа и визуализации данных, а также выявления скрытой ценности, скрытой за данными.

2.3 Каркас Selenium

Selenium - это ранняя основа для автоматизации веб-тестирования, техника автоматизации, которая имитирует реальные действия пользователя в веб-краулере [5]. В отличие от метода Get в библиотеке Requests, ползание с помощью фреймворка Selenium

легко реализовать, он прост в использовании, и вероятность его обнаружения сервером значительно ниже [6].

3. Поиск данных о наборе персонала

3.1 Проектирование Selenium crawler

В этой статье мы используем фреймворк Selenium для сбора данных с сайта по подбору персонала Wiseleaf. Сначала мы открываем целевой URL и входим в систему через Selenium, находим элемент Html, содержащий информацию о задании, через метод Xpath и извлекаем целевые данные, переходим на следующую страницу через Selenium после сбора одной страницы данных, затем начинаем проходить цикл, и, наконец, записываем и сохраняем полученные данные в CSV-файл после сбора заранее заданного количества страниц. Схема проектирования показана на рисунке 1.

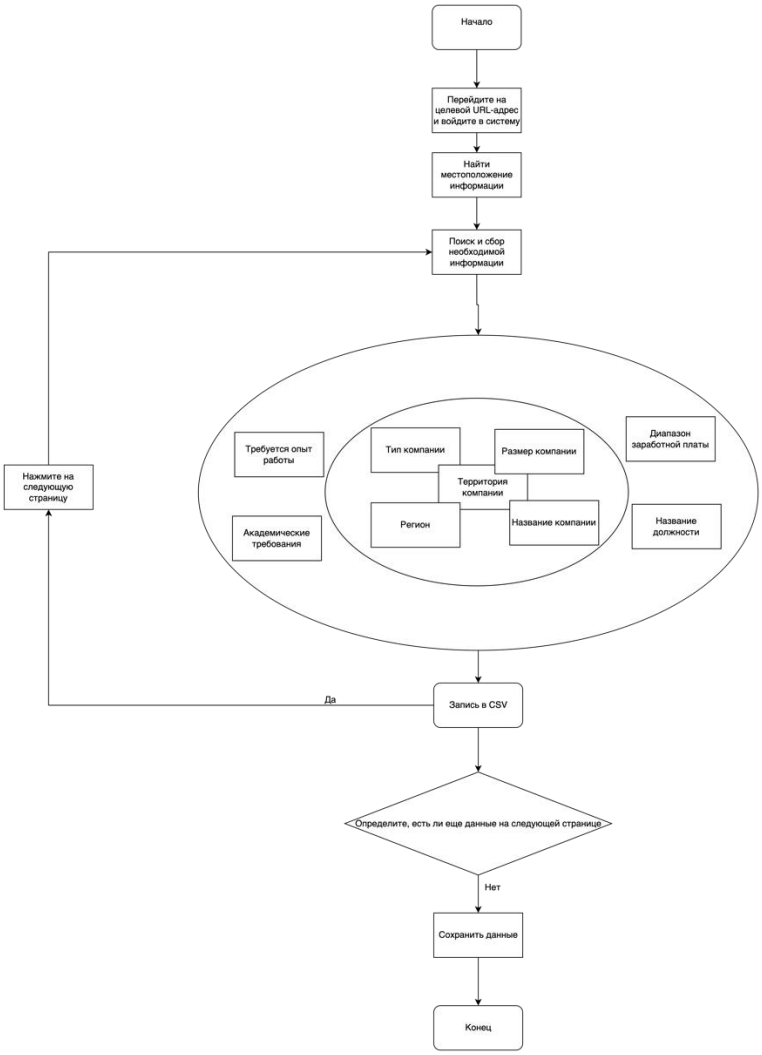

Рис. 1 Блок-схема программы

3.2 Шаги

3.2.1 Создание экземпляра и доступ к URL-адресу

Инстанцируйте драйвер, запустите программу Chromedriver через Webdriver для прямого доступа к целевому веб-сайту и добавьте неявное ожидание стабилизации элементов страницы перед переходом к следующему шагу.

3.2.2 Поиск и извлечение данных о наборе персонала

Метод Xpath используется для поиска и получения данных о найме, включая диапазоны зарплат, требования к образованию и опыту, а также информацию о конкретной компании. Поскольку ошибки могут возникать из-за скорости сети, структура try...except

используется для предотвращения ошибок во время операций захвата и перелистывания страниц. Структуры .except используются для предотвращения возникновения ошибок.

3.2.3 Сохранение данных

Итоговая коллекция данных из 4891 элемента, содержащая 20 городов, была сохранена в CSV и уничтожена браузером. Некоторые из этих данных показаны на рисунке 2.

Поскольку программное обеспечение для подбора персонала китайское, информация в наборе данных, включая названия компаний, названия должностей и т.д., на китайском языке, и я переведу необходимую информацию следующим образом.

В следующей таблице указаны:Название компании, должность, диапазон заработной платы, адрес компании, требуемый опыт, требуемое образование, размер компании, характер компании, сфера деятельности компании.

	公司名称	职位名称	薪酬范围	公司地区	经验要求	学历要求	公司规模	公司性质	公司领域
0	北京迈捷映博科技有限公司	大数据开发工程师	8千-1.5万	北京-海淀区	1年以下	大专	100-299人	其它	大数据开发\n数据挖掘\nOracle\nDB2\nSQL
1	深圳市讯方技术股份有限公司	大数据工程师	1.5万-3万	北京-海淀区	不限	大专	1000-9999人	股份制企业	Spark\nKafka\nHadoop\nHbase\nYarn\nHDFS\nHIVE
2	浩鲸科技	大数据开发工程师	1.2万-2万	北京-丰台区	1-3年	本科	1000-9999人	合资	数据治理\n数仓开发\nHadoop\nSpark\nHive
3	北京蓝飞信息科技有限公司	大数据分析师	8千-1.2万	北京-东城区	不限	大专	20-99人	民营	数据分析\n大数据分析\n数据采集\n大数据平台
4	赛意信息	大数据开发工程师	1.8万-2.8万	北京-海淀区	5-10年	本科	1000-9999人	上市公司	数仓开发\nShell\nHadoop\nHive\nSpark\nKafka\nFlink...
...	...	...	...	...	...	...	...	...	...
4886	山东产业技术研究院	软件开发工程师2(农业院)	面议	济南-历城区	3-5年	硕士	100-299人	事业单位	大数据开发\nETL\n数据建模\n运维
4887	水发智慧产业集团有限公司	ETL工程师	9千-1.2万	济南-历城区	3-5年	本科	300-499人	国企	ETL
4888	济南微生态生物医学省实验室	大数据与人工智能研究专业技术人员	6千-1万	济南-槐荫区	不限	硕士	1000-9999人	事业单位	科研
4889	浪潮集团	技术专家-数据方向	2万-3万	济南-历下区	不限	博士	10000人以上	国企	数据专家
4890	浪潮集团	大数据平台架构师	1.8万-3万	济南	3-5年	硕士	10000人以上	国企	大数据架构\n云计算架构\n平台架构

4891 rows × 9 columns

Рисунок 2 Отобранные графики данных

4 Предварительная обработка данных

4.1 Очистка данных

Прежде чем анализировать данные, их необходимо сначала очистить. Данные могут содержать дубликаты, недостающие значения или выбросы, которые могут повлиять на последующую визуализацию. Обработка данных состоит из следующих этапов:

4.1.1 Дедупликация данных

Используйте функцию drop_duplicates для удаления дубликатов записей:

df.drop_duplicates(inplace=True).

4.1.2 Обработка нуля

Используйте функцию fillna, чтобы заменить Nan символом, например, "none", или просто удалите строку, содержащую Nan:

```
df['адрес'] = df['адрес'].fillna("['none']")
df.dropna(axis=0, how='any')
```

4.1.3 Обработка данных исключений

Данные могут содержать аномальные данные, поэтому данные сортируются по квартилям с использованием метода верхнего квартиля

Метод верхнего квартиля используется для сортировки данных по квартилям от наименьшего к наибольшему для обнаружения и удаления выбросов. Часть кода показана ниже:

```
sal_low = df['avg_salary'].quantile(q=0.25)
sal_high = df['avg_salary'].quantile(q=0.75)
```

```
sal_interval = sal_low-sal_high
```

4.2 Нормализация данных

Данные могут содержать китайские иероглифы или другие символы, которые могут повлиять на визуализацию данных. Чтобы обеспечить нормализацию данных для эффективной визуализации и анализа, данные необходимо разделить и оптимизировать. Например, данные о зарплате содержат такие символы, как "тысяча", "миллион" и "-", которые необходимо унифицировать в арабские цифры, а также подсчитать среднюю, самую высокую и самую низкую зарплату. Код показан ниже:

```
or i in range(0, df.shape[0]):
t = df.loc[[i],[' 薪酬范围 Диапазон заработной
платы']].values.tolist()[0][0] if re.search('(.*)-(.*)',t):
a = re.search('(.*)-(.*)', t).group(1)
if a[-1] == ' 千 Тысячи ':
a = eval(a[0:-1]) * 1000
elif a[-1] == ' 万 10Тысячи. ':
a = eval(a[0:-1]) * 10000
b = re.search('(.*)-(.*)', t).group(2)
if b[-1] == ' 千 Тысячи ':
b = eval(b[0:-1]) * 1000
elif b[-1] == ' 万 10Тысячи. ':
b = eval(b[0:-1]) * 10000
avg_salary = (a + b) / 2
min_salary = a
```

```
max_salary = b
df.loc[[i], [' 平均薪资 Средняя заработная плата ']] = avg_salary

df.loc[[i], [' 最低薪资 минимальная_зарплата']] = min_salary

df.loc[[i], [' 最高薪资 Самая высокая зарплата']] = max_salary
```

5. Визуализация данных

5.1 Анализ городов первого уровня

Анализ конкретного распределения набора персонала в четырех городах верхнего уровня - Северном, Шанхае, Гуанчжоу и Шэньчжэне - представлен на рисунке 3.

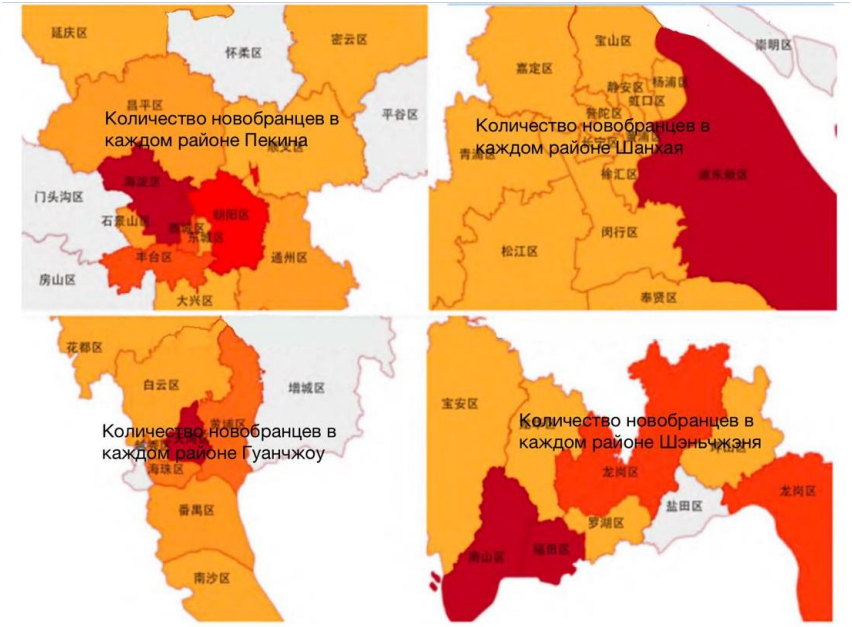

Рисунок 3 Распределение количества городов первого уровня

(1) Пекин. В Пекине находится самое большое количество предприятий по обработке больших данных в Китае. В 2021 году в Пекине насчитывается 3531 предприятие по обработке больших данных, в основном сосредоточенных в районах Хайдянь, Чаоян и Сичэн, среди которых в районе Хайдянь насчитывается 1640 высококачественных предприятий, что составляет почти половину высококачественных предприятий в Пекине. Это связано с тем, что большинство выдающихся пекинских интернет-компаний и университетов сосредоточены в районе Хайдянь, который имеет относительно прочную техническую базу и академическую атмосферу. Это связано с тем, что большинство лучших интернет-компаний и институтов Пекина сосредоточены в районе Хайдянь,

который имеет солидную техническую базу и академическую атмосферу.

(2) Шанхай. Шанхай является ведущей провинцией в развитии индустрии больших данных в Китае. В 2021 году в Шанхае насчитывалось 1651 предприятие по обработке больших данных, расположенных в основном в новом районе Пудун, районе Янпу и районе Миньхан, из которых 417 были высококачественными предприятиями в новом районе Пудун. Данные показывают, что спрос на таланты в Шанхае в основном сосредоточен в новом районе Пудун, в то время как спрос в других районах относительно средний. Поэтому при планировании будущей карьеры выпускников факультета больших данных можно ориентироваться на новый район Пудун в Шанхае.

(3) Шэньчжэнь.В 2021 году в Шэньчжэне будет насчитываться в общей сложности 1446 предприятий с качеством больших данных, что уступает только Шанхаю, и будет являться вершиной развития цифровой экономики Китая. В 2021 году добавленная стоимость основной отрасли цифровой экономики Шэньчжэня составит около 30% ВВП, а масштаб и качество цифровой экономики Шэньчжэня займут первое место среди крупных и средних городов Китая. Результаты анализа данных показывают, что наибольшее количество рекрутов набирается в районе Лунган, за ним следуют районы Наньшань и Футянь, из которых район Лунган является комплексной пилотной зоной больших данных в провинции Гуандун и сильной зоной цифровой экономики в Шэньчжэне. Поэтому выпускники факультета больших данных могут рассматривать район Лонгган в Шэньчжэне для выбора будущей карьеры.

(4) Гуанчжоу. Гуанчжоу прокладывает "новый путь" для искусственного интеллекта и цифровой экономики и стремится стать демонстрационной зоной цифровой экономики мирового класса. В 2021 году в городе Гуанчжоу будет 1051 высококачественное предприятие в области больших данных, уступая только городу Шэньчжэнь. И город Гуанчжоу, и город Шэньчжэнь имеют хорошую основу для развития интернета, что привлекает большое количество превосходных предприятий в области больших данных. Результаты анализа данных показывают, что район Тяньхэ имеет самый высокий спрос на таланты, и 10-й съезд четко заявил, что район Тяньхэ должен стать районом искусственного интеллекта и цифровой экономики, а также моделью глубокой интеграции между промышленностью и городом. Поэтому выпускники факультета больших данных могут выбрать для своего развития район Тяньхэ в Гуанчжоу.

5.2 Анализ других крупных городов

Топ-10 городов (за исключением четырех крупных городов Севера, Шанхая, Гуанчжоу и Шэньчжэня) показаны на рисунке 4. В тройку лидеров по спросу на рекрутинг вошли Ханчжоу, Чэнду и Нанкин, за ними следуют Сиань и Ухань. Видно, что Ханчжоу, Чэнду и Нанкин имеют относительно высокий спрос на специалистов, связанных с большими данными, и в тяжелой ситуации с трудоустройством эти три города можно рассматривать для поиска работы.

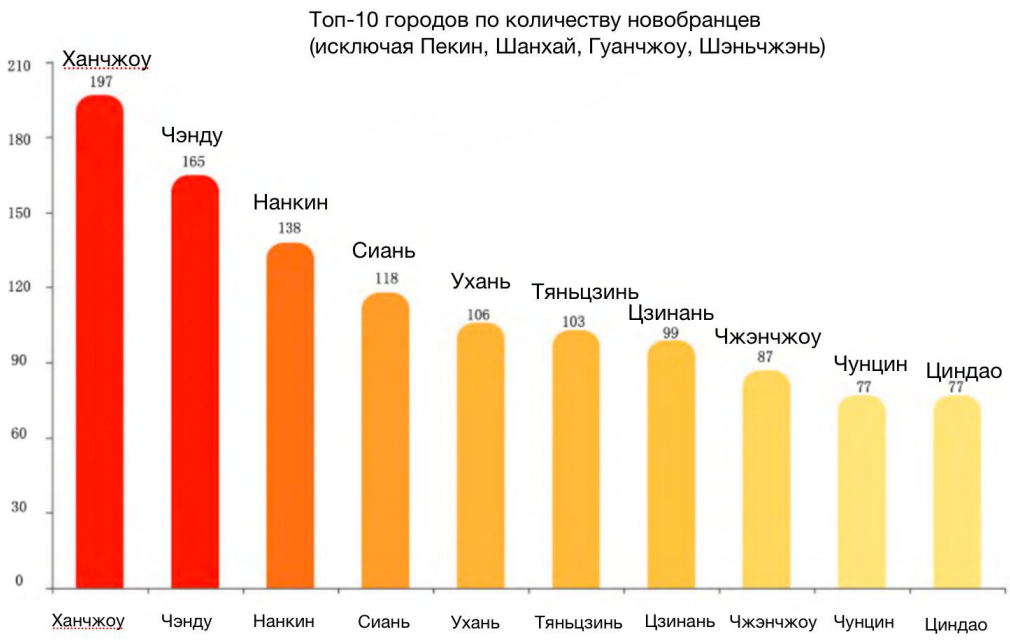

Рисунок 4 Карта количества других городов

5.3 Анализ рабочихмест

Как показано на рисунке 5, наибольшим спросом пользуются инженеры по разработке больших данных - 63,9%, а аналитики больших данных занимают второе место - 18,59%. Спрос на инженеров по разработке больших данных гораздо выше, чем на другие должности, поэтому специалистам, работающим на должностях, связанных с разработкой больших данных, легче найти работу, чем на других должностях.

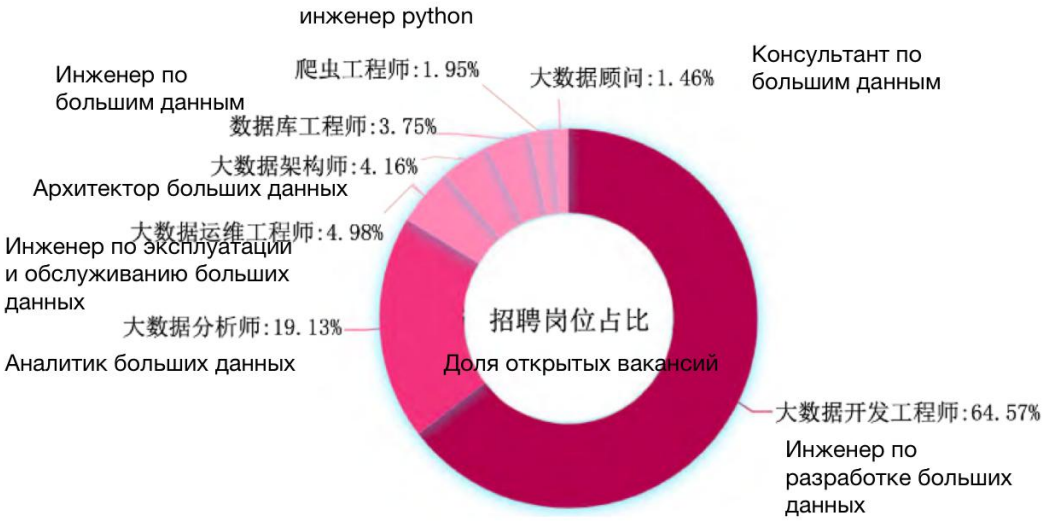

Рисунок 5 Диаграмма распределения рабочих мест

5.4 Эмпирический анализ

Как показано на рисунке 6, большинство компаний требуют 3-5 лет опыта работы (46,66%), за ними следуют компании с опытом работы от 1 до 3 лет (22,64%). Большинство компаний требуют опыт работы от 3 до 5 лет - 46,66%, затем следует опыт работы от 1 до 3 лет - 22,64%.

Рисунок 6 Требования к опыту в процентах

5.5 Анализ заработной платы

Топ-10 городов с точки зрения средней заработной платы показаны на рисунке 7, при этом средняя зарплата в Северном, Гуанчжоу и Шэньчжэне занимает самые высокие позиции. В дополнение к четырем крупным городам, Нанкин и Ханчжоу также имеют относительно высокие средние зарплаты. Выпускники, специализирующиеся на больших данных и предъявляющие высокие требования к зарплате, могут сосредоточиться на четырех крупных городах на севере, а Нанкин и Ханчжоу также являются хорошим выбором.

Рисунок 7 Карта 10 лучших городов по средней заработной плате

6. Резюме и предложения

Наступила эра Больших Данных, и перед людьми открываются ценные возможности для точного изучения ценности, содержащейся в данных огромного масштаба и сложности. Развитие больших данных приведет к замене некоторых рабочих мест, но также создаст новые [7]. Первая специальность - менеджмент, которая изучает общий контроль и процесс обработки данных, и выпускники должны овладеть методами анализа больших данных и соответствующими передовыми теоретическими знаниями, соответствующими таким профессиям, как аналитики данных и аналитики бизнес-аналитики. Последняя специальность - компьютерные науки, изучающие лежащие в основе технологии и их реализацию, и выпускники должны обладать способностью исследовать алгоритмы добычи данных и разрабатывать системы больших данных. В этот период быстрого развития индустрии Больших Данных студенты, изучающие Большие Данные, должны идти в ногу с тенденцией и стать теми талантами в области Больших Данных, которые нужны компаниям, чтобы не быть уничтоженными в будущем. Спрос на таланты в области больших данных наиболее сосредоточен в городах первого уровня, в то время как спрос в городах второго и третьего уровней также растет с каждым годом. В то же время, предприятия, как основной орган привлечения талантов на работу, должны соотнести свои потребности с рынком труда и создать модель своих критериев найма для достижения точного найма талантов больших данных [8]. Во время профилактики и борьбы с эпидемиями онлайн-рекрутинг также имеет то преимущество, что он сводит к минимуму перемещение людей и минимизирует риск заражения болезнью для кандидатов и персонала [9]. В этой статье мы используем данные,

полученные в процессе найма, чтобы предоставить информацию для соискателей и компаний для принятия решений в нынешней сложной ситуации с трудоустройством. Несмотря на то, что эти решения надежны и практичны, в анализе данных в целом еще много возможностей для совершенствования. Мы надеемся, что изложение этой статьи послужит некоторым ориентиром для ученых и компаний, работающих с большими массивами данных.

Ссылки

[1] Wang Liming. Data sharing and personal information protection[J]. Modern Jurisprudence,2019,41(1): 45-57.

[2] Meng Xianying,Mao Yingshang. Collection and analysis of commodity information based on Python crawler technology[J]. Software,2021,42(11):128-130.

[3] Long Weiqiu. Re-discussing the path of property rights of enterprise data protection[J]. Oriental Law,2018 (3):50-63.

[4] Pan Xiaoying, Chen Liu, Yu Huimin, et al. A review of topic crawling technology research[J]. Computer Application Research,2020,37(4):961-965+972.

[5] Fan Tao, Zhao Zheng, Liu Minjuan. Analysis and implementation of a Selenium-based web crawler [J]. Computer programming skills and maintenance,2019(9):155-156+170.

[6] Hua Junlin. Implementation of a Selenium-based Python web crawler [J]. Computer Programming Skills and Maintenance,2017(15):30-31+36.

[7] Hu Qiong Yue. 2018 Popular positions and specialties for employment of big data talents [J]. Big Data Times,2018(2):27-31.

[8] Shi Yunsheng, Zong Shengwang. Thinking about building a precise employment service system for college students[J]. Cooperation Economy and Technology,2017(2):136-137.

[9] Wang Weiling,Wu Zhigang. Research on the development of digital economy under the influence of the new crown pneumonia epidemic[J]. Economic Vertical,2020(3):16-22.