**Аннотация**

В данном проекте была проведена работа по созданию веб-приложения с помощью языка программирования Python и платформы для сбора и обработки данных ELK (ElasticSearch, LogStash, Kibana). Целью данной работы является сбор и анализ данных с сайта HH.ru по IT-вакансиям. В результате парсинга данных было собрано 352272 вакансии сферы IT по РФ и странам СНГ. Поиск осуществлялся по ключевым словам за период с 10.04.2024 по 11.05.2024.

**Статья**

Сфера IT стала неотъемлемой частью современного мира. Стремительное развитие и расширение объёма данный формирует необходимость наличия специалистов, квалификация которых будет соответствовать требованиям рынка. В связи с этим, исследования, проводимые по данной тематике, становятся все более актуальными.

Благодаря развитию IT- сферы на рынке появились и продолжают появляться новые вакансии. При это перечень квалификаций у разных работодателей на схожих должностях может отличаться. В связи с чем возникает необходимость анализа и структурирования ключевых компетенций для конкретных должностей.

Целью данной работы является выявление основных компетенций для специалистов сферы IT в Санкт-Петербурге.

В работе собранны данные об IT- вакансиях за период с 10.04.2024 по 11.05.2024, по которым проведен анализ требуемых навыков. Сбор данных осуществлен на базе крупнейшей компании интернет-рекрутмента – HH.ru.

Выделим основные задачи, которые были поставлены в данной работе:

1. Анализ литературы и выделение ключевых слов для поиска вакансий
2. Сбор данных по выбранным вакансиям сферы IT в РФ и СНГ.
3. Сбор информации о количестве вакансий за выбранный временной промежуток.
4. Сбор информации о компетенциях для выбранных вакансий.
5. Анализ компетенций и выделение основных навыков, языков программирования.
6. Анализ частоты публикаций вакансий
7. Структурирование результатов работы и построение графиков, диаграмм

Результатом работы является структурированная информация об основных требованиях к соискателям в сфере IT на базе платформы HH.ru. Результаты исследования могут быть полезны для учебных заведений (так как помогут сформировать и откорректировать учебный план, в целях актуализации), для соискателей, работодателей.

**Обзор литературы.**

Множество исследований были посвящены теме исследования рынка IT вакансий. В том числе одно из исследований было сделано платформой HH.ru [1]. В данной работе были поставлены следующие задачи:

1) Проанализировать какие зарплаты предлагает бизнес ИТ-специалистам.

2) Как сильно отличаются ставки опытного и начинающего айтишника, компании из каких регионов готовы платить больше всего.

3)Насколько высокая конкуренция среди соискателей сложилась в 2023 году.

В результате построения регрессионной модели и анализу данных, аворы пришли к следующим выводам:

1. Заработная плата руководителя, специалиста и новичка значительно отличается: начального уровня специалист может рассчитывать на не более 45 000, эксперт - 175 000, в то время как средняя зарплата руководителя составляет около 286 000 в Российской Федерации.

2. Наибольший интерес к IT-сфере в качестве нового рабочего места (или перехода из другой отрасли) проявляют кандидаты из Приволжского федерального округа (45% опрошенных), затем идет Санкт-Петербург (44%), на третьем месте - Сибирь (42%).

3.В Российской Федерации наблюдается наибольшая конкуренция на рынке труда среди гейм-дизайнеров, где на одну вакансию приходится более 46 резюме. В северо-западном федеральном округе самая жёсткая борьба за место работы ведётся среди тестировщиков (72 резюме) и аналитиков данных (67 резюме), в то время как в Москве наибольшая конкуренция наблюдается среди руководителей аналитических ИТ-отделов (более 26 резюме) и также гейм-дизайнеров (25 резюме). В Санкт-Петербурге в 2023 году особенно сложно найти работу в сфере дата-сайенса из-за большого количества претендентов (более 30 резюме).

4. Анализ рынка труда показывает, что в первом полугодии 2023 года чаще всего выходили на рынок труда и боролись за работу разработчики JavaScript (более 54 резюме на вакансию). На втором месте оказались разработчики и программисты, работающие на React (более 36 резюме на вакансию по стране). На третьем месте по уровню конкуренции — разработчики TypeScript (почти 33 резюме).

Многие исследования были посвящены анализу вакансий рынка IT и тенденции востребованности IT-специалистов. Так, согласно статье сайта habr.com в которой был проанализирован спрос на IT-специалистов по России (за 2023–2024 гг.) одни из самых востребованных вакансий: системные и бизнес-аналитики, менеджеры проектов и продуктов, а также Java-разработчики. Авторы пришли к следующим выводам:

«1)Спрос на ИТ-специалистов за год вырос: в сентябре 2022 года на портале hh.ru было открыто 64 тыс. вакансий, а в сентябре 2023 года компании разместили 76 тыс. вакансий. Годовой прирост составил 18%, согласно статистике hh.ru.

2)Активнее других бизнес начал искать аналитиков (системных и бизнесовых), а также менеджеров проектов. За год число размещенных вакансий для этих позиций увеличилось практически в 2 раза.

3)В среднем каждый месяц компании размещают более 9 тысяч вакансий для разработчиков. Практически каждая шестая вакансия — Java-разработчик. Их доля от общего числа вакансий для разработчиков составляет 16%. В первой тройке востребованных стеков также расположились PHP и Python разработчики — 12% и 11% соответственно.

4)Самые активные наниматели айтишников — компании из сферы «Информационные технологии». Но при этом, в первой тройке нанимателей Сбер, Тинькофф и Иннотех — компании из сферы «финансы».

5)Специалистов грейда Junior стали искать на 101% чаще, но в абсолютных значениях бизнесу интересны опытные специалисты уровня Senior — на момент сентября этого года 2,7 тыс. вакансий для Senior против 0,6 тыс. для Junior.

6)По медианным зарплатам лидируют разработчики Ruby on rails — компании готовы платить им 252 000 рублей в месяц. За ними следуют Golang и Java разработчики.

7)Свердловская область в тройке среди регионов нанимателей. Компании из Екатеринбурга и области расположились после Москвы и Санкт-Петербурга.

8)Почти каждая вторая вакансия для ИТ-специалистов предполагает формат удаленной работы: с 25% в 2021 году до 40% в 2023 году.»[2].

Авторы Волкодаева А.В. и другие так же рассмотрели в своей работе тенденции востребованности ИТ-специалистов. В результате проведения статистического анализа авторы пришли к выводу о том, что «в ближайшие несколько лет сохранится тенденция роста востребованности IT-специалистов на рынке труда. Область применения результатов: полученные результаты целесообразно применять социально-экономическими системам, осуществляющим мониторинг на рынке труда и в системе образования»[3].

Рассмотрим работы авторов, посвящённые анализу компетенций ИТ-специалистов. Константинова Л. А. и Крамаренко И. В. В своей работе провели анализ спроса на компетенции в области информационных технологий от лидирующих российских компаний. Авторы использовали эмпирический метод в своей работе. В результате проделанной работы были сделаны следующие выводы: «значимым требованием к соискателям в ИТ-сфере стало знание языков программирования (Python, C++, Java), которое присутствует в 38% вакансий, а также теоретическая подготовка (знания алгоритмов и структур данных, методов математической статистики и теории вероятностей, принципов машинного обучения), фигурирующая в 23% вакансий. При этом наибольшим спросом пользуются аналитики (36% вакансий) и фулстек-разработчики, включая фронтенд-разработчиков и бэкенд-разработчиков (27% вакансий) »[4].

Терников А.А. и Александрова Е.А. так же посвятили свою работу анализу спроса на навыки ИТ-специалистов. «В статье представлен алгоритм, позволяющий определить ключевые комбинации навыков, которые необходимы компаниям в сфере информационных технологий в зависимости от профессиональных групп. Использованные в статье подходы, TF-IDF и n-граммы, позволили извлечь и структурировать знания, умения и навыки, полученные из неструктурированной базы данных интернет-вакансий на рынке труда»[5]. Данный алгоритм позволяет определить и стандартизировать ключевые навыки, которые могут быть использованы для создания системы российских классификаторов по профессиям и навыкам.

Некоторые авторы посвятили свои работы разработке технологии анализа актуальных вакансий с целью формирования новых профессиональных программ. Так, авторы Шухман А.Е., Гришина Л.С. и другие [6] в результате проведенного анализа профессий «программист» на базе крупнейшей компании интернет-рекрутмента – HH.ru выделили основные компетенции: навыки командной работы, работа с базами данных и системами контроля версий, владение английским языком. Авторы отмечают, что самый востребованный язык программирования среди данных вакансий - JavaScript.

В результате проведения литературного обзора был сформирован список ключевых слов, который будет использоваться для дальнейшего парсинга данных с сайта HH.ru: ВІ-аналитик, аналитик данных, DevOps-инженер, гейм-дизайнер, дата-сайентист, маркетолог-аналитик, интернет-маркетолог, программист (разработчик), системный администратор, системный аналитик, специалист по информационной безопасности, тестировщик, фронтенд, веб-программист, разработчик мобильных приложений, Java-программист, 1С-программист, Бакенд-программист (Python, C++), Product Manager, Project managers, бизнес-аналитик, системный аналитик, разработчики (frontend, backend и mobile), Java-разработчиков, PHP и Python разработчики, Golang-разработчик, C++ разработчики, Java-разработчиков, PHP разработчики, Python разработчики, C++ разработчики, аналитик данных, специалист службы поддержки, продуктовый аналитик, разработчик на платформе 1C, Kotlin разработчик, системный администратор, системный аналитик, специалист по информационной безопасности, системный аналитик, разработчик мобильных приложений, DevOps, администратор баз данных, AI/ML спецалисты, дата-сайентисты, ИТ-специалисты высокого уровня квалификации, ИТ-специалисты низкого уровня квалификации, инженеры, разработчики ПО, Веб и мультимедиа разработчики, системные администраторы и администраторы баз данных.

**Характеристика датасета и описание модели данных**

База датасета: [api.hh.ru](https://api.hh.ru/)

Сбор данных с помощью языка программирования Python

Географические ограничения: РФ и страны СНГ

Период сбора информации: 10.04.2024-11.05.2024 гг.

Фильтр поиска: по ключевым словам

Количество данных: 352272 вакансии

Формат файла датасета: JSON File (.json)

Размер файла датасета: 905 МБ

Описание модели данных:

- Идентификатор вакансии: Уникальный идентификатор вакансии.

- Название вакансии: Название IT-вакансии.

- Местоположение: Город или регион, где расположена компания.

- Заработная плата: Уровень заработной платы, указанный в объявлении.

- Описание вакансии: Полное описание требований, обязанностей и условий работы.

- Требуемые навыки: Список навыков, необходимых для данной вакансии (дополнительный анализ текста)

- Уровень образования: Уровень образования, необходимый для -кандидатов. (дополнительный анализ текста)

- Испытательный срок/вариант занятости (например, стажировка, полный рабочий день и т.д.)

**Программная реализация**

В ходе выполнения задания было использовано несколько программных компонентов:

1. Наименование - PyCharm

Тип: Кроссплатформенная интегрированная среда разработки для языка программирования Python.

Описание: Предназначена для создания парсера для сбора и выгрузки вакансий и их описания с сайта hh.ru.

2.Наименование - Logstash

Тип: Инструмент обработки и агрегации данных.

Описание: Используется для централизованного сбора, фильтрации и трансформации данных перед их передачей в Elasticsearch.

3.Наименование - Elasticsearch

Тип: Система управления и поиска данными обеспечивая быстрый доступ к информации.

Описание: Предоставляет возможность хранения, поиска и анализа больших объемов данных в режиме реального времени.

4. Наименование - Kibana

Тип: Инструмент визуализации данных.

Описание: Позволяет создавать интерактивные визуализации данных из Elasticsearch, для удобного исследования и анализа информации.

**Описание ограничений по работе с источником данных**

1. За один запрос можно выгрузить только 2000 последних вакансий.

2. В среднем 10 запросов/сек с одного IP-адреса к API HeadHunter.

3. На данный момент можно получить информацию по более чем 98 миллионам вакансий, хранящихся на API.

4. В руководстве к пользованию API HeadHunter сталкиваемся с необходимостью регистрации приложения. Можно выгрузить без регистрации, однако количество запросов с одного IP будет ограничено (например, выгрузка информации только за последний месяц).

**Описание и использование подходов горизонтального масштабирования**

Масштабирование — это процесс добавления ресурсов, направленный на возможность информационной системы справляться с растущими объемами данных и их обработкой.

В целом под горизонтальным масштабированием понимают разбиение системы на более мелкие структурные компоненты и разнесение их по отдельным физическим машинам (или их группам), и (или) увеличение количества серверов, параллельно выполняющих одну и ту же функцию. То есть, путем такого масштабирования увеличивается общая производительность.

В процессе работы программы был получен итоговый файл с 352272 вакансиями размером 905 Мб, для выгрузки такого количества данных потребовалось 5 часов 28 минут, что достаточно немного.

Поэтому использование 1 сервера было достаточным для поставленной задачи и горизонтальное масштабирование не потребовалось.

**Временные характеристики (аналитические задачи)**

Подбор и анализ литературы – 3 дня

Написание парсера – 4-5 дней

Выгрузка данных с HH.ru – 5,5 часов

Среднее время реализации кода в Python – 4 секунды

Выгрузка данных в Elasticsearch для визуализации – 1 день

Среднее время на подготовку графиков – 2-3 секунды

Описание результатов и составление отчета – 2-4 дня

**Результаты исследования**

Рассмотрим в процентном соотношении необходимые навыки для соискателей (рисунок 1).

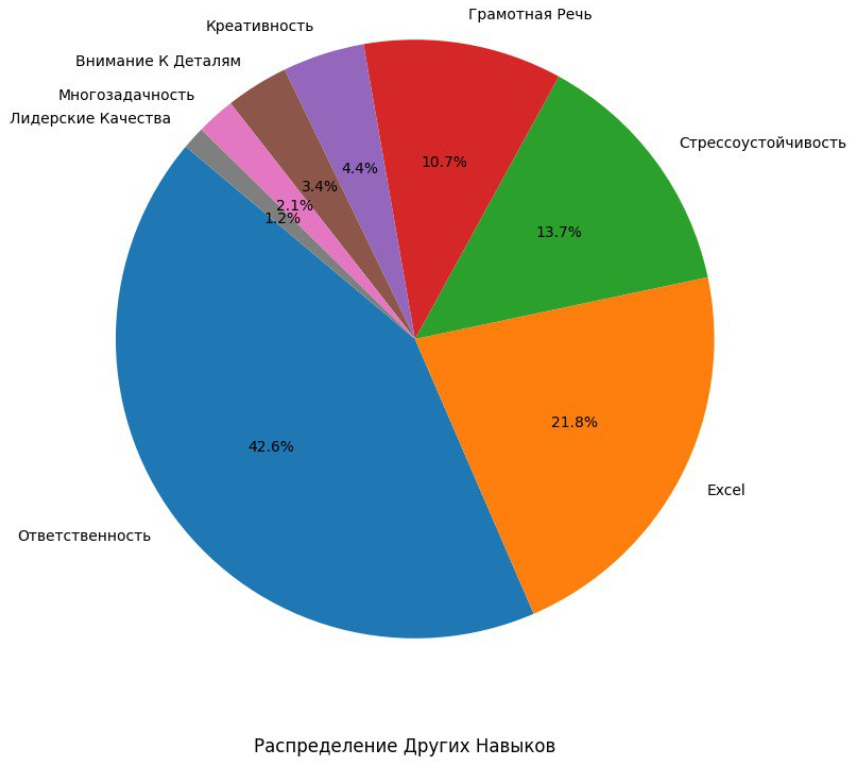


Рисунок 1 – необходимые навыки специалистов IT

Исходя из графика на рисунке 1 можно сделать следующий вывод: 46,2% работодателей считают, что «ответственность»-один из основных навыков, которым должен обладать соискатель. Второе место по необходимым навыкам занимает знание Excel-21,8% и третье место-стрессоустойчивость13,7%.

Рассмотрим требования к образованию для соискателей (рисунок 2).

Изображение выглядит как диаграмма, снимок экрана, текст, круг

Автоматически созданное описание

Рисунок 2 – необходимое образование специалистов IT

Исходя из результатов анализа требуемого образования можно сделать вывод о том, что большинство работодателей (63,3%) рассматривают специалистов с высшим образованием.

Рассмотрим результаты анализа процента упоминаний в вакансиях требований к знанию у соискателей языков программирования (рисунок 3).

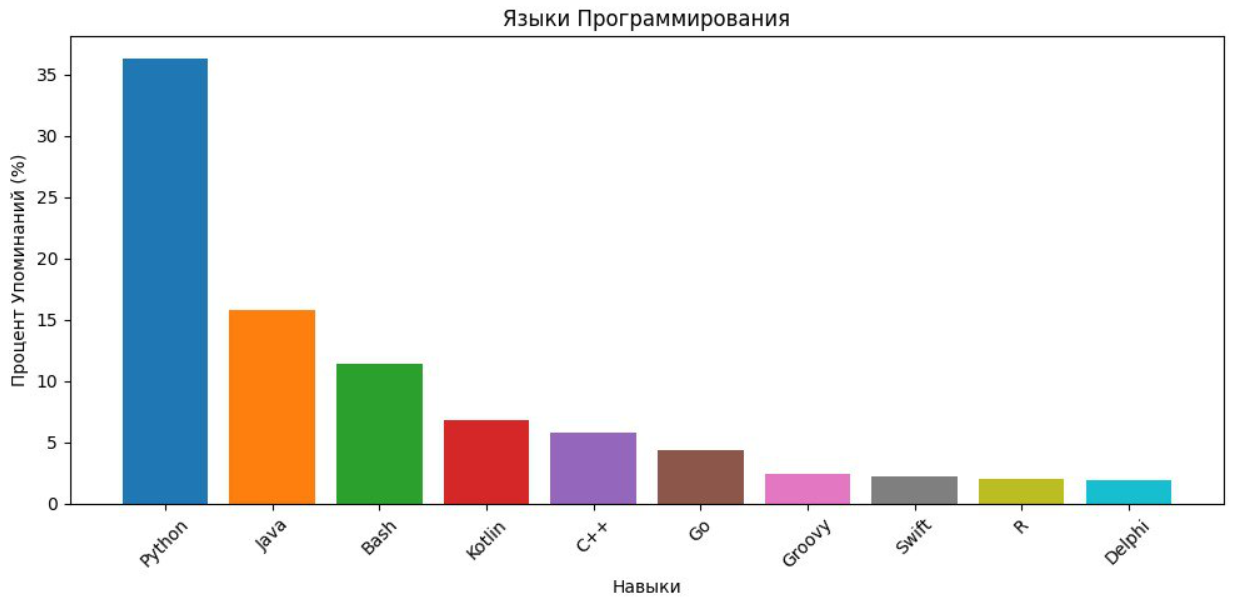


Рисунок 3 – требования к знанию языков програмирования

Более, чем в 35% вакансий упоминается необходимость владения языком программирования Python, более 15% - Java и примерно 13%- Bash.

Рассмотрим требования к знаниям языков среди соискателей (рисунок 4).

Изображение выглядит как текст, снимок экрана, График, диаграмма

Автоматически созданное описание

Рисунок 4 – требования к знанию языков

Выделим 3 основных языка, которые наиболее часто упоминаются в требованиях к соискателям: русский язык (33%), английский язык (26%) и узбекский язык (18%).

Проанализируем результаты частоты публикаций вакансий (время суток), которые представлены на рисунке 5.

Изображение выглядит как текст, График, линия, диаграмма

Автоматически созданное описание

Рисунок 5 – частота публикации вакансий (время суток)

В результате анализа было выявлено, что наибольшее количество публикаций приходится с 9:00 по 16:00, причем в 10:00 происходит публикация более 36000 вакансий.

Проведем анализ и выявим в какие дни недели наблюдается наибольшее количество публикаций вакансий (рисунок 6).

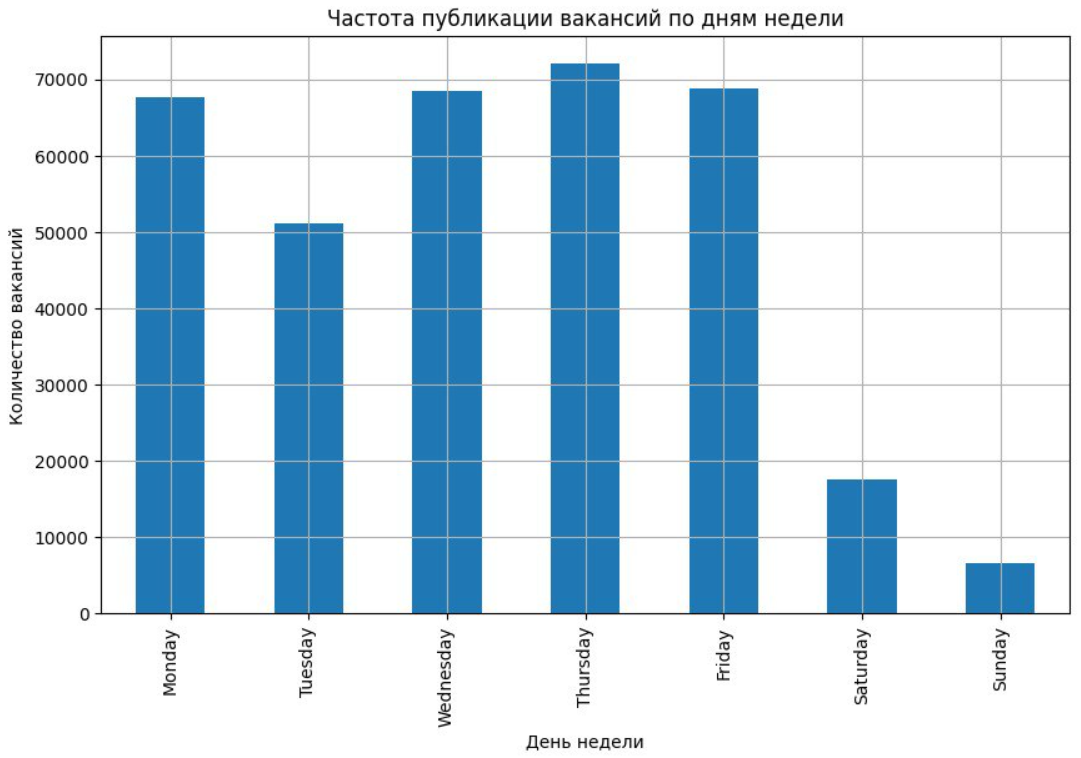


Рисунок 6 – частота публикаций вакансий по дням недели

Можем заметить, что наибольшее количество опубликованных вакансий приходится на четверг (более 70000 вакансий), так же практически 70000 вакансий публикуются в понедельник, среду и пятницу.

Проанализируем и выявим какой опыт работы наиболее часто упоминается работодателями в объявлениях (рисунок 7).

Изображение выглядит как текст, Шрифт, снимок экрана, дизайн

Автоматически созданное описание

Рисунок 7 – требования к опыту работы

Как можно заметить, большинство работодателей упоминаю требуемый опыт работы от 1 до 3 лет (45,62%). Так же 34,13% работодателей рассматривают соискателей с опытом работы от 3 до 6 лет.

Рассмотрим частоту публикаций объявлений в час в период с 10.04.2024 по 11.05.2024 гг. (рисунок 8).

Изображение выглядит как снимок экрана, текст, линия, График

Автоматически созданное описание

Рисунок 8 – частота публикации объявлений в час в период с 10.04.2024 по 11.05.2024 гг.

В результате анализа было выявлено, с 8 мая по 10 мая наблюдается резкий скачок частоты публикации объявлений в час. Аналогичная картина наблюдается и с частотой публикации объявлений в день (рисунок 9).

Изображение выглядит как График, линия, снимок экрана, диаграмма

Автоматически созданное описание

Рисунок 9 - частота публикации объявлений в день в период с 10.04.2024 по 11.05.2024 гг.

Рассмотри топ 10 городов, где работодатели публикуют наибольшее количество объявлений (рисунок 10).

Изображение выглядит как текст, График, диаграмма, линия

Автоматически созданное описание

Рисунок 10 - топ 10 городов, где работодатели публикуют наибольшее количество объявлений

С наибольшим отрывом лидирует Москва (более 60000 публикаций), в Санкт-Петербурге опубликовалось в рассматриваемый период примерно 18000 объявлений, и примерно одинаковое количество публикаций наблюдается в таких городах, как: Екатеринбург, Алматы и другие.

Представим на рисунке 11 количество вакансии, представленные на сайте HH.ru. Чем больше шрифт вакансии, тем больше объявлений на сайте.

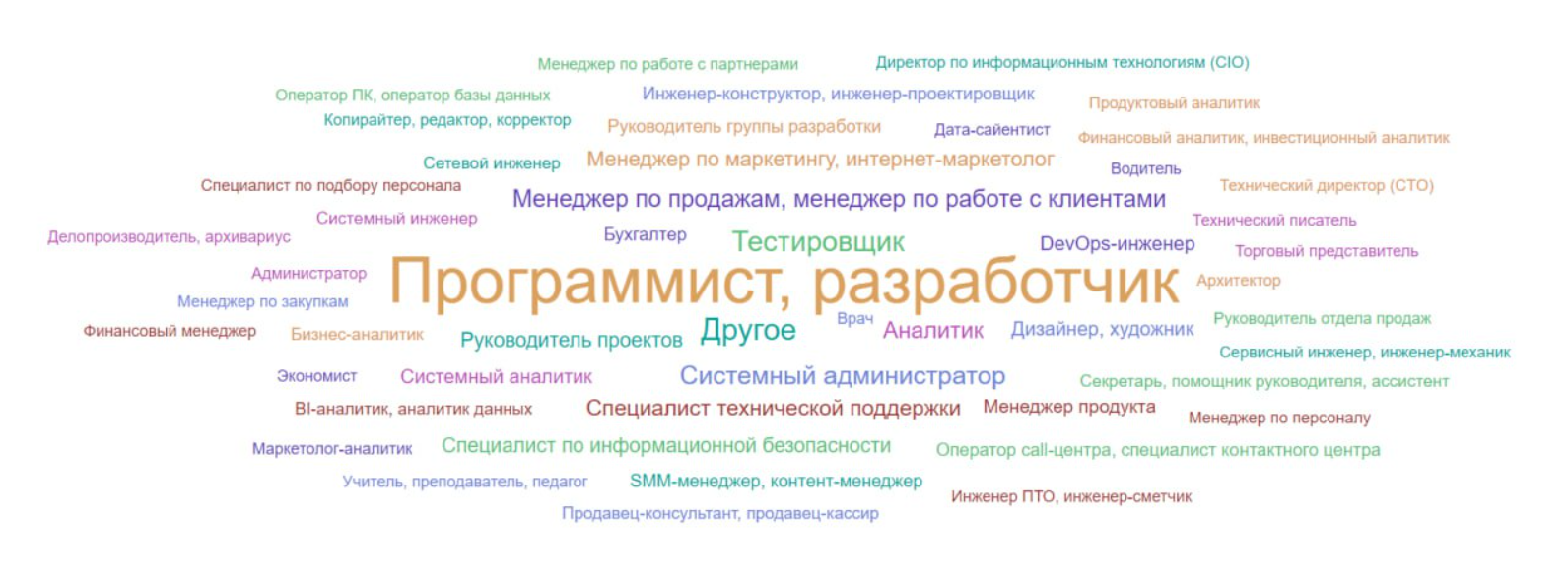


Рисунок 11 – упоминания вакансий на сайте HH.ru

Выделим топ-10 компаний размещающих наибольшее количество вакансий (рисунок 12).

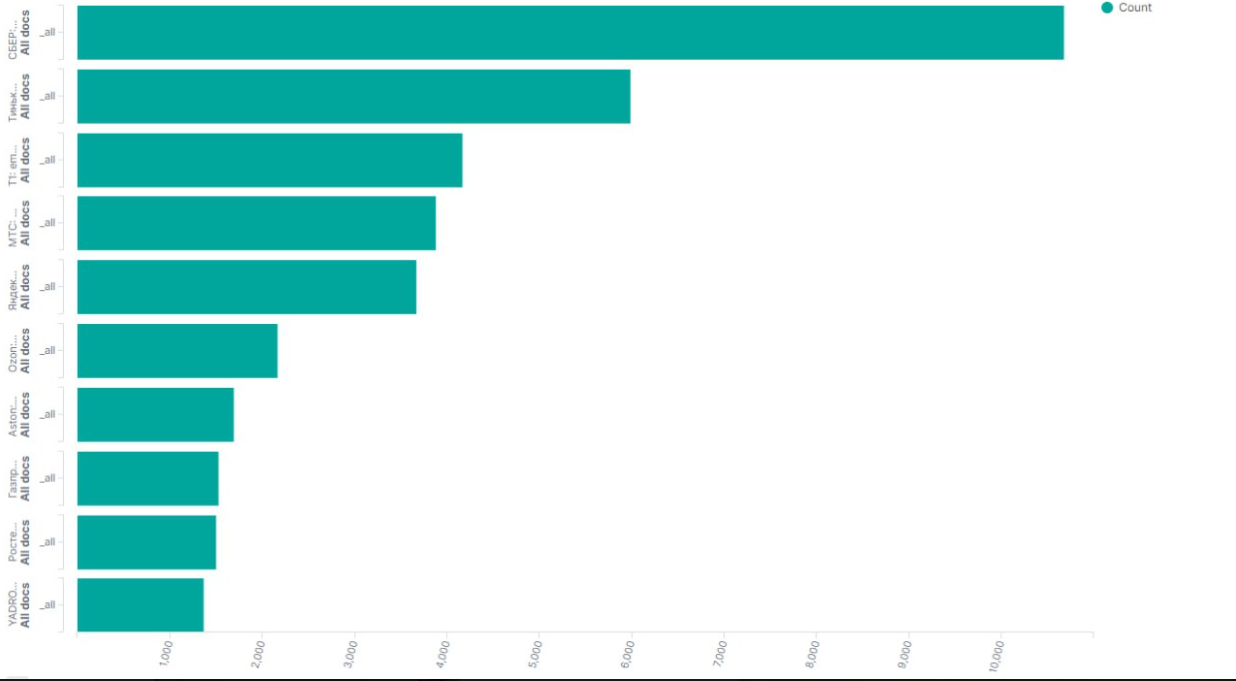


Рисунок 12 - Топ 10 компаний

Лидером среди топа-10 компаний является ПАО «Сбербанк» - более 10000 опубликованных вакансий.

Рассмотрим заработные платы в топ-10 городов, где публикуется наибольшее количество вакансий (рисунок 13).



Рисунок 13 – заработные платы в топ-10 городов по количеству публикаций

Анализируя результаты, приведённые на рисунке 12 можно сделать вывод: в Москве наиболее распространенная заработная плата, упоминаемая в объявлениях – 150000 рублей. В Санкт-Петербурге большинство объявлений с заработной платой – 100000 рублей, в Екатеринбурге – 70000 рублей.

Рассмотрим статистику по занятости среди вакансий (рисунок 14).

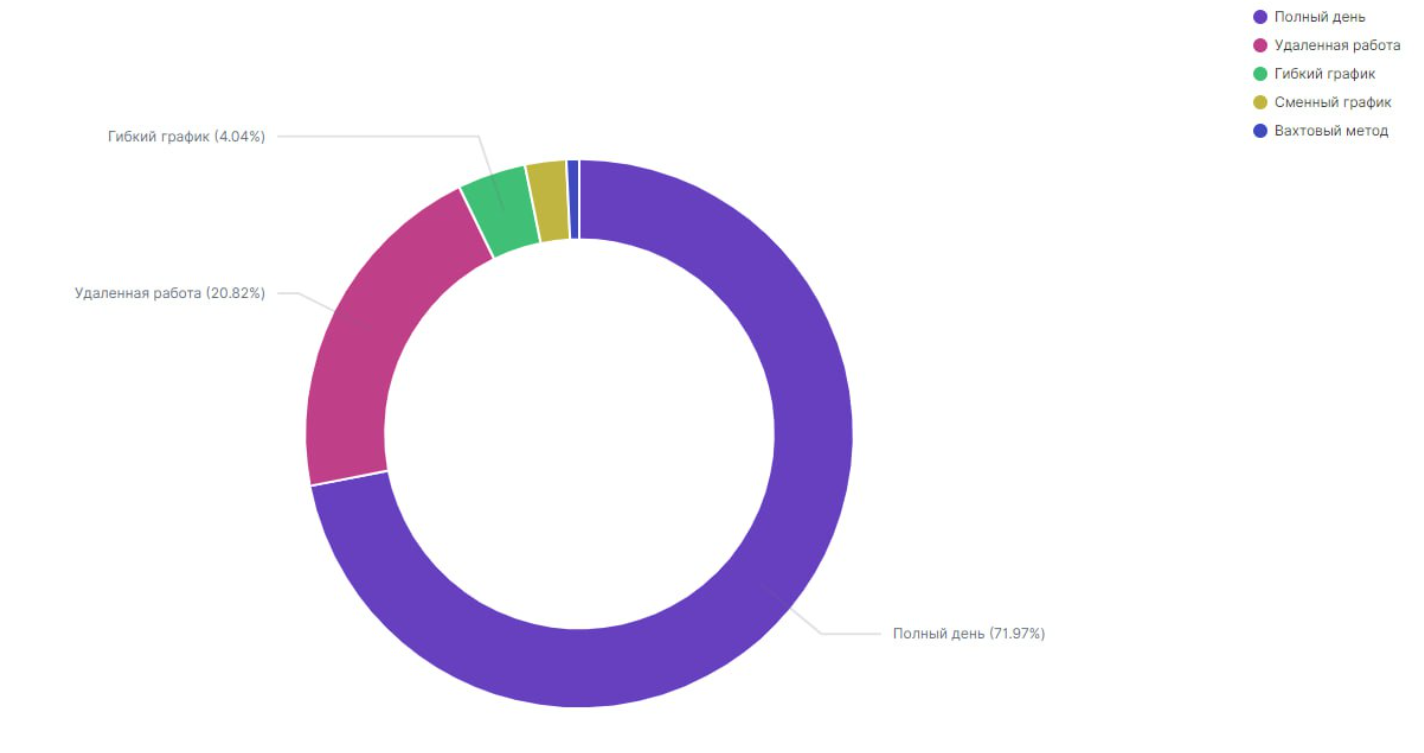


Рисунок 14 – график занятости

Исходя из результатов можем сделать вывод о том, что большинство работодателей ищут специалистов на полный рабочий день (71,97%). Такая же тенденция наблюдается и в других городах среди топ-5 городов (рисунок 15).

Изображение выглядит как снимок экрана, текст

Автоматически созданное описание

Рисунок 15 – тип занятости среди топ-5 городов по количеству публикаций вакансий

Рассматривая требуемый опыт работы в топ-5 городов (рисунок 16), можем заметить, что в большинстве вакансий необходимо наличие опыта работы от 1 до 3 лет.

Изображение выглядит как текст, снимок экрана, диаграмма

Автоматически созданное описание

Рисунок 16 – опыт работы в топ-5 городов

Как можем заметить на рисунке 17, большинство заработных плат указывается в рублях (89,27%).

Изображение выглядит как снимок экрана, текст, диаграмма

Автоматически созданное описание

Рисунок 17 – Валюта заработной платы по странам

Изображение выглядит как текст, программное обеспечение, снимок экрана

Автоматически созданное описание

Рисунок 18 – топ-5 городов станции метро

Изображение выглядит как текст, снимок экрана

Автоматически созданное описание

Рисунок 19 – топ-5 городов и улицы

Изображение выглядит как текст, снимок экрана

Автоматически созданное описание

Рисунок 20 – топ-5 городов и компании

Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание

Рисунок 21 – облако компаний

***Результаты.***

Компании ищут сотрудников со следующими основными навыками:

Ответственный стрессоустойчивый, владеет Excel, имеет высшее образование/техническое образование, знание языков программирования: Python, Java, Bash, знание языков: русский язык, английский язык, опыт работы от 1 до 3 лет, устройство на полный рабочий день.

Резюме компаний и объявлений:

Компании публикуют большую часть объявлений с 9:00 до 16:00. Наибольшее количество публикаций приходится на понедельник, среду, четверг и пятницу. Наименьшее количество вакансий публикуется в воскресенье. Работодатели в большей степени ищут: программистов, разработчиков, тестировщиков, аналитиков. Компании в основном ищут сотрудников на полную занятость (72%) и удаленную работу (21%). Большая часть вакансий нацелена на поиск сотрудников для работы в Москве, Санкт-Петербурге, Екатеринбурге. Большая часть объявлений приходится на «Сбербанк», «Тинькофф», «Яндекс Краут», «МТС», «Yadro» и другие.

**Применение статистических методов.**

В ходе анализа данных были обнаружены некорректные данные, вакансии, которые не относятся к сфере IT. Из данных были исключены выбросы.

Данные неоднородны, так как, например при попытке построения графика з/п для компаний мы можем наблюдать недостаточное количество данных (рисунок 21).

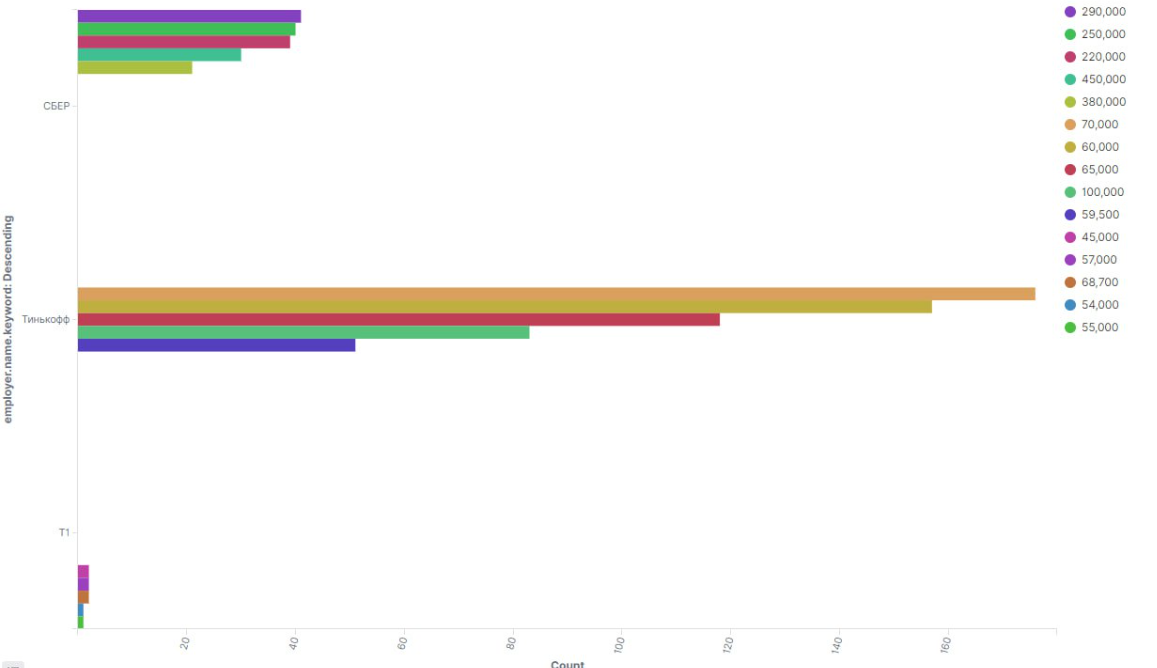


Рисунок 21 – заработные платы в компаниях

**Дистрибутив представлен следующими файлами:**

Код для парсинга данных (Python)

Код для визуализации данных(Python)

Файл для визуализации в Elastic

Описание программной реализации

**Особенности ПО**

Использовались «с нуля установленные» дистрибутивы elasticsearch и kiban

Также выгружены вакансии не по специальности «1» (– ИТ сфера), а, по ключевым словам, для отбора именно вакансий связанных с программирование, DevOps, анализом данных и других вакансий

В рамках дальнейшего исследования можно использовать анализатор текста и маппинг для разделения описания требований к вакансии для более точного результата

Т. к. анализатора текста и маппинг не было предусмотрено в нашей работе, мы часть анализа производили через Python.

**Список литературы**

1. Обзор ИТ-отрасли по итогам первого полугодия: какие зарплаты платят и насколько сложно найти работу // spb.hh.ru URL: https://spb.hh.ru/article/31783?ysclid=lt8ylrej88270300483 (дата обращения: 12.05.2024).

2. Исследование рынка ИТ-вакансий: годовой рост составил 18% // habr.com URL: https://habr.com/ru/articles/770720/ (дата обращения: 10.05.2024).

3. Тенденции востребованности специалистов сферы информационных технологий на рынке труда в России // cyberleninka.ru URL: https://cyberleninka.ru/article/n/tendentsii-vostrebovannosti-spetsialistov-sfery-informatsionnyh-tehnologiy-na-rynke-truda-v-rossii (дата обращения: 10.05.2024).

4. Анализ спроса на компетенции в области информационных технологий от лидирующих российских компаний // cyberleninka.ru URL: https://cyberleninka.ru/article/n/analiz-sprosa-na-kompetentsii-v-oblasti-informatsionnyh-tehnologiy-ot-lidiruyuschih-rossiyskih-kompaniy (дата обращения: 01.05.2024).

5. Спрос на навыки на рынке труда в сфере информационных технологий // cyberleninka.ru URL: https://cyberleninka.ru/article/n/spros-na-navyki-na-rynke-truda-v-sfere-informatsionnyh-tehnologiy (дата обращения: 10.05.2024).

6. Какие компетенции ждёт региональный рынок ИТ-специалистов? Разработка и апробация технологии анализа актуальных вакансий // cyberleninka.ru/article URL: https://cyberleninka.ru/article/n/kakie-kompetentsii-zhdyot-regionalnyy-rynok-it-spetsialistov-razrabotka-i-aprobatsiya-tehnologii-analiza-aktualnyh-vakansiy (дата обращения: 01.05.2024).