Использование визуальных моделей для разведочного анализа слабоструктурированных текстовых данных

Е.А. Макарова [1], Д.Г. Лагерев [1]

[1] *Брянский государственный технический университет, бульвар 50-летия Октября, д.7, Брянск, 241035, Россия*

Аннотация

Обработка слабоструктурированных текстовых данных с целью дальнейшего использования в моделях ИАД – трудоемкий процесс, который, помимо материальных затрат, может увеличить время, которое требуется на построение модели, и, как следствие, ухудшить оперативность принятия решений. В данной статье представлены визуальные модели слабоструктурированных текстовых данных и методы их обработки на этапе разведочного анализа. Разведочный анализ позволит сократить время на выбор значимых переменных на начальном этапе исследования и, в дальнейшем, избежать обработки излишних или незначительных. Использование визуализации поможет включить в модель ИАД и обработать только те данные, которые повысят её качество. Описан процесс использование визуализации текстовых данных в процессе разведочного анализа и построения двух типов визуальных моделей – интерактивная «количественная» визуализация и визуализация связей между словами и другими переменными в исследуемых данных. Описана апробация разработанных моделей на примере анализа рынка труда. Представлены примеры визуализации содержимого поля «гибкие навыки» из резюме соискателей и вакансий, отображающие как наиболее часто упоминаемые соискателями из различных профессиональных областей навыки, так и влияние упоминания этих навыков на приглашения соискателей на собеседования. Проведенный эксперимент показал, что использование разработанных визуальных моделей позволяет определить, нужно ли включать текстовую переменную в модель ИАД на этапе разведочного анализа.

Ключевые слова

Обработка естественного языка, визуализация данных, разведочный анализ данных, коэффициент корреляции, анализ рынка труда.

Using Visual Models for Exploratory Analysis of Semi-structured Text Data

E.A. Makarova[1], D.G. Lagerev[1]

[1] *Bryansk State Technical University, boulevard of the 50th anniversary of October, 7, Bryansk, 241035, Russia*

Abstract

The processing of semi-structured textual data for further use in DM models is a labor-intensive process, which, in addition to material costs, can increase the time required to build a model, and, as a result, worsen the efficiency of decision-making. This article presents visual models of semi-structured text data and methods for their processing at the stage of exploratory analysis. Exploratory analysis will reduce the time to select significant variables at the initial stage of the study and, in the future, avoid the processing of redundant or insignificant variables. The use of visualization will help to include in DM model and process only data that will improve DM model quality. The process of using visualization of textual data in the process of exploratory analysis and the construction of two types of visual models is described - interactive "quantitative" visualization

ГрафиКон 2022: 32-я Международная конференция по компьютерной графике и машинному зрению, 19-22 сентября 2022 г., Рязанский государственный радиотехнический университет им. В.Ф. Уткина, Рязань, Россия
EMAIL: m4karova.e@yandex.ru (Е.А. Макарова); lagerevdg@yandex.ru (Д.Г. Лагерев)
ORCID: 0000-0002-5410-5890 (Е.А. Макарова); 0000-0002-2702-6492 (Д.Г. Лагерев)

and visualization of relationships between words and other variables in the data under study. Approbation of the developed models is described on the example of labor market analysis. Examples of visualization of the content of the "soft skills" field from the CV and vacancies are presented, displaying both the skills most often mentioned by applicants from various professional fields, and the impact of mentioning these skills on inviting applicants for interviews. The experiment showed that the use of the developed visual models makes it possible to determine whether it is necessary to include a text variable in the DM model at the stage of exploratory analysis.

Keywords
Natural language processing, data visualization, exploratory data analysis, correlation coefficient, labor market analysis.

1. Введение

Управленцы в различных социально-экономических областях всё чаще используют в своей деятельности методы интеллектуального анализа данных (далее – ИАД). Обычно в их основе лежит использование структурированных реляционных данных, однако, при изучении ряда объектов социально-экономических систем целесообразно учитывать также данные из слабоструктурированных текстовых источников. К подобным данным относятся данные из новостных статей, комментариев, посты в социальных сетях при оценке репутации юридического или физического лица, содержание резюме и вакансий при оценке состояния рынка труда и т.д.

Но не всегда добавление текстовых данных в модель ИАД приводит к улучшению качества. Исследования, посвященные вопросам эффективного использования текстовых данных в моделях прогнозирование финансовых событий, показывают, что прирост достигается только при условии правильного отбора и предобработки текстовых данных [1]. Использование этих данных в модели ИАД требует больших трудозатрат на их сбор и обработку [2]. Кроме того, не все текстовые данные могут быть однозначно интерпретированы и обработаны без привлечения эксперта в предметной области [3], что увеличивает затраты ещё на этапе обработки и разметки данных.

Отсеивание данных, которые не будут способствовать повышению качества модели ИАД, целесообразно провести на этапе разведочного анализа. Однако, большая часть методов и инструментов для разведочного анализа работает только с числовыми или категориальными данными, включая методы, использующие визуализацию [4]. Использование визуальных моделей в процессе разведочного анализа позволяет существенно увеличить скорость исследования данных. В то же время, создание визуальных моделей для представления слабоструктурированных данных во многих случаях приводит к необходимости решения самостоятельной задачи поиска наиболее информативного способа представления таких данных [5].

Таким образом, необходимо разработать метод обработки слабоструктурированных текстовых данных и визуальные модели, которые применимы на этапе разведочного анализа, и которые позволят сократить время аналитика или инженера данных на выбор значимых переменных на начальном этапе исследования. В дальнейшем использование такого метода позволит избежать обработки излишних данных путем включения в модель и последующей обработки только значимых данных.

2. Использование обработки естественного языка и визуальных моделей в процессе разведочного анализа

Целью разведочного анализа данных является «погружение» в исследуемые данные, нахождение корреляций, аномалий и общих закономерностей [6]. На основе его результатов можно отбросить или принять изначально выдвигаемые гипотезы, и даже составить новые, которые без разведочного анализа данных не рассматривались. Кроме того, разведочный анализ

позволяет произвести отбор переменных, отбросив те из них, которые излишни или незначительны в контексте решаемой задачи [7].

Общая схема работы исследователя с моделями ИАД, поддерживающими слабоструктурированные текстовые данные, включая этап разведочного анализа, изображена на рисунке 1, где ХД – хранилище данных.

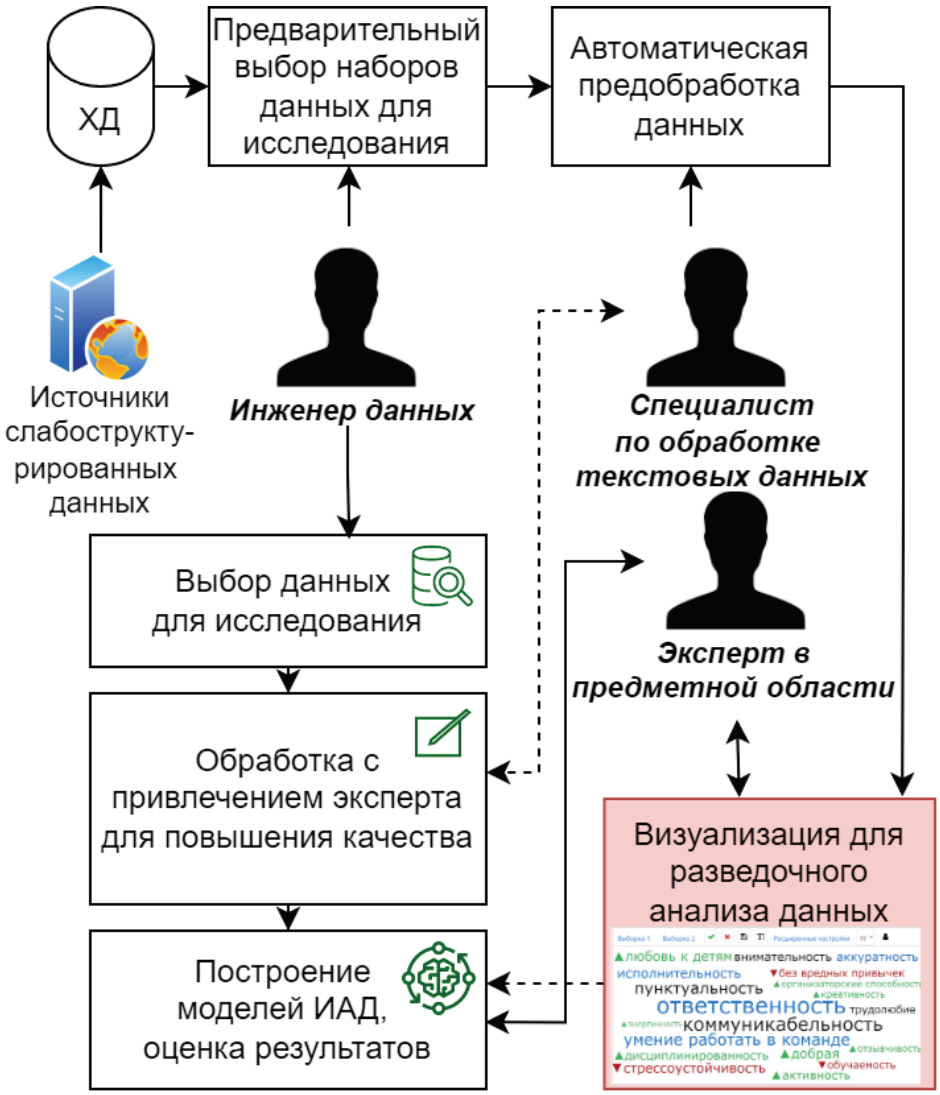

Рисунок 1 – Схема работы исследователя с моделями ИАД, включающими в себя слабоструктурированные текстовые данные

Базовая автоматическая предобработка данных для построения визуальной модели, достаточной для разведочного анализа, включает в себя следующие этапы:

- удаление шумов, HTML-тегов, по заранее заданным регулярным выражениям;
- удаление стоп-слов;
- токенизация;
- стемминг.

Расширенная предобработка данных, которая позволит улучшить качество анализа, будет включать следующие этапы:

- возможность настройки регулярных выражений через пользовательский интерфейс, облегчающих работу с ними без предварительной подготовки [8];
- использование обученных на исследуемой выборке моделей Word2Vec для объединения семантически близких слов и словосочетаний в группы, настройка порога объединения, как общего, так и для конкретных слов [9];

- замена стемминга на лемматизацию при достаточном уровне аппаратных ресурсов.

При наличии временного ресурса и необходимости улучшить качество данных уже на этапе разведочного анализа, любой из этих этапов может быть включен в процесс автоматической обработки. Параметры обработки, которые выбраны пользователем, будут использованы для повторного разведочного анализа или дальнейших манипуляций с данными. Полученный обработанный текст будет использован для построения визуальных моделей, которые описаны далее.

2.1. Визуальные модели трендов и связей в данных, включая текстовые

Для использования визуализации в процессе разведочного анализа в данной работе предлагается разработать две визуальные модели:

1. Интерактивная «количественная» визуализация.
2. Визуализация связей между словами и другими показателями исследуемых объектов.

Обе модели будут строиться на основе визуальной модели представления текстов «облако слов», отображающей распространенность тех или иных слов в исследуемом тексте или наборе текстов. Тренды и связи между отображаемыми языковыми единицами (словами, словосочетаниями) будут визуализироваться с помощью изменения размеров и цвета. Перед построением визуальных моделей необходимо совершить расчеты характеристик отдельных языковых единиц для повышения информативности представления таких данных.

2.1.1. Построение интерактивной «количественной» визуализации

Интерактивная «количественная» визуализация отображает наиболее встречающиеся в исследуемых выборках слабоструктурированных текстовых данных слова и словосочетания и виде «облака слов» с возможностью изменения пользователем как выборки, так и её визуальной модели, используя интерфейс визуализации.

Первая цель интерактивной визуализации – проверка корректности выборки и редактирование запроса. Интерактивная «количественная» визуализация также может помочь пользователю разобраться со следующими характеристиками текстовых данных:

- самые упоминаемые в выборке слова, кроме общеупотребительных (стоп-слов);
- разница упоминания слов между двумя выборками.

Если на этапе предварительной обработки был применен стемминг, то для отображения будут выбрано оригинальное слово к стемме, исходя из наиболее частого употребления в выборке. Пример приведён в таблице 1. В случае, если были применены другие методы обработки, поддерживающие обработку словосочетаний и объединение их в группы, то в качестве отображаемого словосочетания будет выступать наиболее часто упоминаемое.

Таблица 1 – Пример поиска стемм и отображения их в визуализации

Описание	Пример
Слова, упомянутые в выборке (с количеством упоминаний)	ответственный (52), ответственность (106), ответственна (30)
Стемма, используемая при подсчете	ответствен
Слово, отображаемое в визуализации	ответственность

На этапе настройки создания визуализации, производится подсчет уникальных стемм, присутствующих в выборке Количество слов и словосочетаний, которое будет отображено на визуализации, выбирается пользователем, по умолчанию их 20. Веса для отображения в визуализации для каждой языковой единицы будет рассчитываться по формуле:

$$t_i = \sum_j^n st_j \quad (1)$$

$$s_i = (f_{max} - f_{min}) \cdot \frac{t_i - t_{min}}{t_{max} - t_{min}} f_{min} \quad (2)$$

где t_i – общее количество упоминаний для отображаемой языковой единицы i;
n – множество слов и словосочетаний, входящих в группу, объединенную по семантической близости или, в случае использования стемм – слов, ассоциированных со стеммой;
st_j – количество повторений каждого элемента множества n;
s_i – шрифт языковой единицы i.
f_{max} – максимальный размер шрифта;
f_{min} – минимальный размер шрифта;
t_{min} – наименьшее количество упоминаний их отображаемых языковых единиц;
t_{max} – наиболее упоминаемая из отображаемых языковых единиц.

Разница упоминания языковых единиц между двумя выборками может быть полезна, например, для выявления трендов от одного временного периода к другому. Так же оценка может быть проведена по регионам и любым другим свойствам исследуемых объектов.

Для подсчета, является ли разница между упоминаниями значительной, используется формула:

$$\begin{gathered} \text{если } \frac{t_2}{t_1} \geq 1{,}5, \text{то изменение положительное,} \\ \frac{t_2}{t_1} \leq 0.66, \text{то изменение отрицательное,} \\ 0{,}66 < \frac{t_2}{t_1} < 1{,}5, \text{то изменения незначительное,} \\ \text{при условии, что } |t_1 - t_2| \geq 10 \end{gathered} \quad (3)$$

где t_1, t_2 – общее количество упоминаний для языковой единицы из первой и второй выборки соответственно.

После построения и отображения, у пользователя, помимо возможности визуального анализа, есть несколько вариантов взаимодействия с языковыми единицами в визуальной модели:

- корректировка запроса, путём отметки слов или словосочетаний, объекты с которыми должны быть исключены выборки;
- редактирование визуализации: удаление языковых единиц из визуализации (но не из выборки), настройка параметров группировки по семантической близости;
- возможность перейти в корреляционную визуализацию, связанную с упоминанием этой языковой единицы и остальных.

2.1.2. Построение визуализации связей между словами и другими показателями исследуемых объектов

Анализ корреляций между переменными в данных может нести следующие цели:

- поиск зависимостей, которые необходимо учесть при детальном исследовании и построении гипотез;
- отбор переменных для использования в моделях машинного обучения [10].

В зависимости от количества отображаемых слов, выбранных на предыдущем этапе, будут рассчитаны дихотомические переменные, исходя из факта упоминания стеммы в текстовой записи. Также будет добавлена дихотомическая переменная наличия описания, если в выборке присутствуют пустые поля. Выбор коэффициента корреляции будет зависеть от типа переменных, связь слова или словосочетания с которой изучается, список их представлен в таблице 2.

Для оценки тесноты связи в визуализации используется шкала Чеддока [13], представленная в таблице 3. Пользователь сам выбирает уровень меры тесноты связи, которые будут выделены

в визуализации. На практике, по итогам окончательных расчетов, исследователю нужно использовать классические методы для поиска корреляций и определения их силы [14].

Таблица 2 – Меры связи

Тип переменной, связь с упоминанием которой рассматривается	Мера связи
Дихотомическая	Коэффициент φ[11]
Интервальная	Бисериальный коэффициент [12]
Ранговая	Рангово-бисериальный коэффициент [12]

Таблица 3 – Шкала Чеддока

Количественная мера тесноты связи	Мера связи
-0.1:0.1	Отсутствует
0.1:0.3 (-0.1:-0.3)	Слабая
0.3:0.5 (-0.3:-0.5)	Умеренная
0.5:0.7 (-0.5:-0.7)	Заметная
0.7:0.9 (-0.7:-0.9)	Высокая
0.9:0.99 (-0.9:-0.99)	Весьма высокая

Рассчитанные таким образом корреляции не являются окончательными, на их основе нельзя принимать решений о наличии влияния упоминания определенных слов на другие показатели. Однако, предварительные расчеты позволяют решить, углублять ли обработку данного текста для использования его в дальнейшем исследовании, чтобы получить более точные выводы по наличию зависимостей.

2.2. Экспериментальная часть

Для апробации разработанных визуальных моделей использовались данные с сайта «Работа в России»[15]. Объектом разведочного анализа являются обезличенные резюме, которые создали на портале пользователи из разных регионов России и стран СНГ. Вопрос, который стоит перед потенциальным исследователем на первом этапе: стоит ли направить усилия на обработку поля «гибкие навыки» для использования результатов в модели ИАД. Анализ слабоструктурированных данных, которые заполняются вручную, требуют внедрения автоматизированной обработки, которую авторы описывали в своих предыдущих работах. Например, объединение синонимичных сочетаний («легко обучаем», «готов к обучению», «быстро обучаюсь») в одно и т.д. Примеры подобных данных представлены в таблице 4 в том виде, в котором они хранятся в базе данных, включая пользовательские ошибки.

В зависимости от задачи исследования, подобные слабоструктурированные текстовые данные могут быть значимыми для модели, а могут оказаться излишними. Перед началом исследования, необходимо выяснить, много ли пропусков в этих данных и сколько уникальных стемм в них присутствует. Эти данные будут отображены на этапе первоначальной настройки визуализации и влияют на оценку потенциальной пользы данных и выбор пользователем размера визуализации. В выборке резюме, опубликованных в 2021 году, от соискателей, претендующих на должность «Программист», изучаемые данные заполнены у 37% пользователей и содержат 980 стемм.

При стандартном размере визуализации (20) получен результат, отображенный на рисунке 2. При наведении курсора на языковую единицу указывается количество упоминаний. Более крупный шрифт языковой единицы обозначает более частое её упоминание.

Таблица 4 – Примеры заполнения поля «гибкие навыки»

Желаемая должность	Гибкие навыки
Начальник цеха	«Ответственность, целеустремленность, коммуникабельность, стрессоустойчивость. Хобби – спорт, фотография»
Бухгалтер	«ОТВЕТСТВЕННОСТЬ, ИСПОЛНИТЕЛЬНОСТЬ, ДИСЦИПЛИНИПРВОАННОСТЬ»
Дворник	«Алкоголь не употребляю»
Инженер электросвязи	«Ответсвенность, коммуникабельность, дисциплинированность»
Курьер	«- презентабельный внешний вид| – выносливость и мобильность| – ответственность и порядочность| – пунктуальность и аккуратность| – коммуникабельность, обаяние| – отсутствие вредных привычек»

При стандартном размере визуализации (20) получен результат, отображенный на рисунке 2. При наведении курсора на языковую единицу указывается количество упоминаний. Более крупный шрифт языковой единицы обозначает более частое её упоминание. Если на визуализации одна выборка, то цвета – чёрный и синий, рассчитаны как наиболее контрастные к друг другу и белому фону – не несут смысловой нагрузки, служат целью визуального отделения одних единиц от других при отображении их на одном горизонтальном уровне. Если добавлена «Выборка 2» – выборка для сравнения – то самые употребляемые единицы также рассчитываются по «Выборке 1», а разница их упоминания обозначается цветами (зелёный – упоминание выросло, красный – упоминание упало) и треугольниками, направленными, соответственно, вверх и вниз. Треугольники добавлены, в том числе, для поддержки работы с визуализацией лицами с нарушениями цветовосприятия.

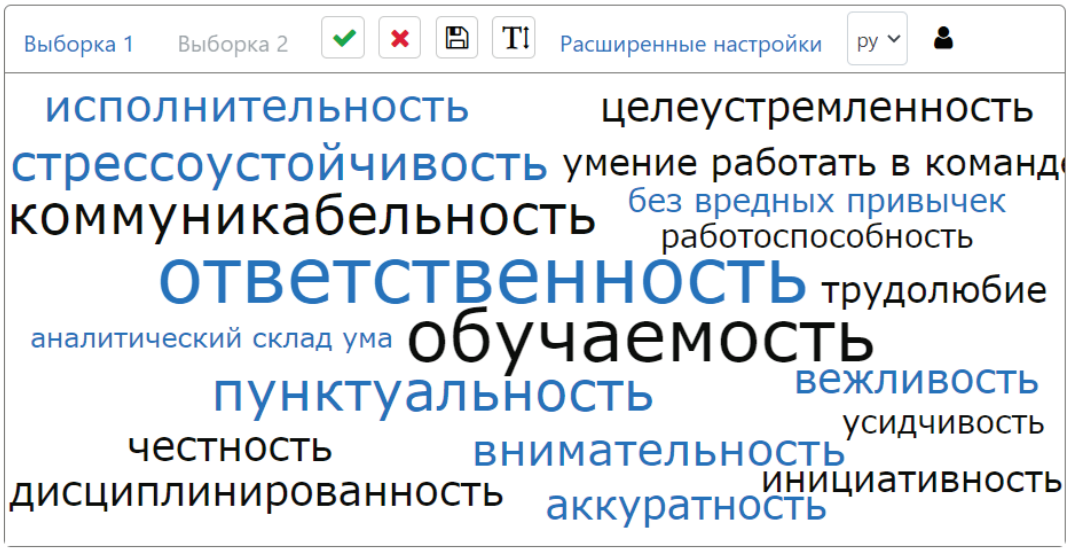

Рисунок 2 – Количественная визуализация по профессии «Программист»

Количественные визуализации позволяют отследить разницу трендов упоминания навыков в резюме от года к году, например, если в «Выборку 2» добавить тот же запрос, но изменить год на 2020ый. На рисунке 3 видно, что по сравнению с предыдущим годом упало упоминание таких качеств, как «обучаемость» и «целеустремленность».

Кроме того, разницу упоминания «гибких навыков» между соискателями по профессиям «Вожатый» («Выборка 1») и «Программист» («Выборка 2») можно увидеть, настроив соответствующие запросы. Наиболее заметными отличиями являются рост упоминания таких качеств, как «любовь к детям», «доброта» и «активность» и падение упоминания таких качеств как «стрессоустойчивость» и «отсутствие вредных привычек» (рисунок 4).

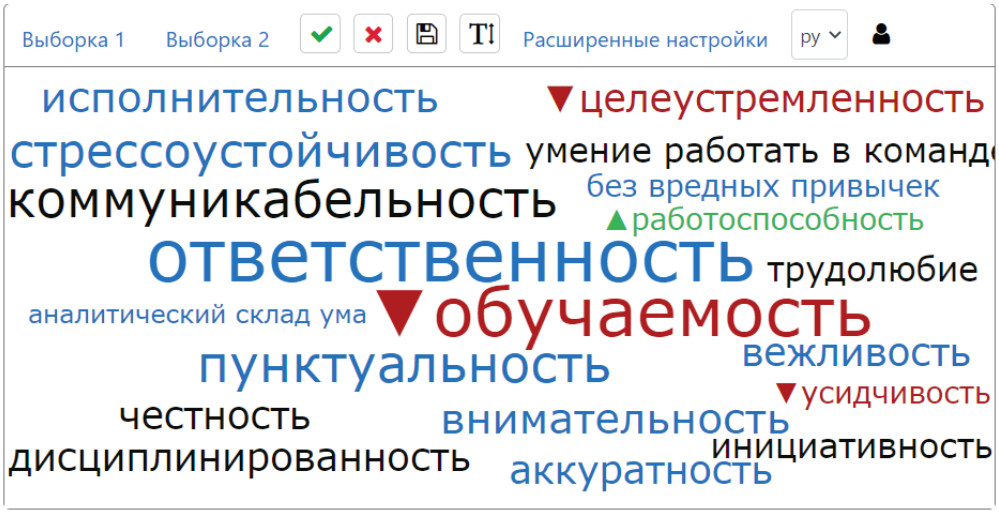

Рисунок 3 – Визуализация разницы между выборками резюме по профессии «Программист» в 2020 и 2021 году

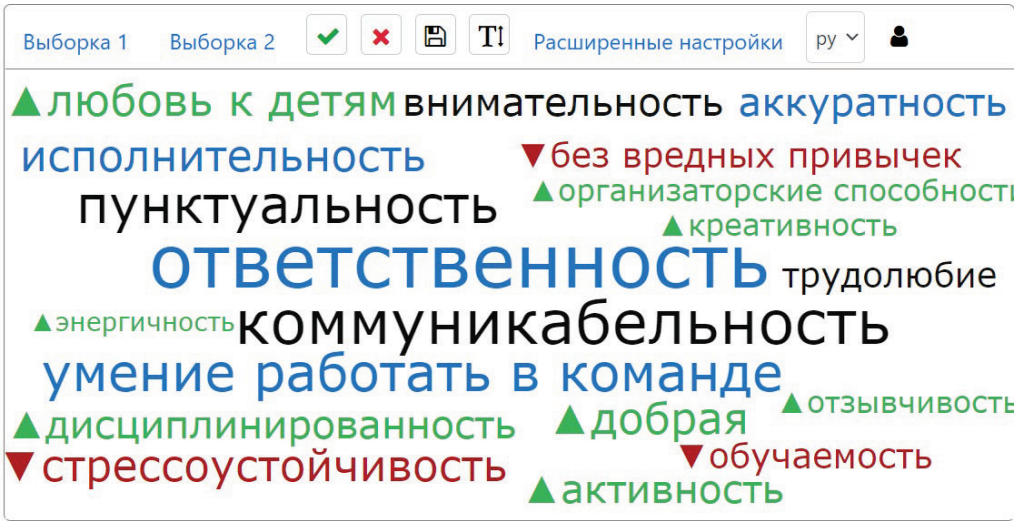

Рисунок 4 – Визуализация разницы между двумя профессиями «Вожатый» и «Программист»

В изучении рынка труда не меньший интерес представляет информация, указанная в вакансиях работодателей. С точки зрения исследования «гибких навыков» в изучаемом наборе данных есть проблема с обработкой вакансий. В описании вакансии, в отличие от резюме, нет отдельного поля для их перечисления. Для извлечения этих данных из вакансий были предварительно обработаны данные из поля «Гибкие навыки» из резюме и автоматически составлен список из 318 гибких навыков, которые достоверно часто (более 50 раз) упоминали пользователи. Ограничения по количеству упоминаний были применены по причине того, что, помимо гибких навыков, часть пользователей указало в данном поле навыки, относящиеся к профессиям («уверенный пользователь Excel»). Основываясь на полученном списке, была получена визуализация гибких навыков, которые чаще всего упоминали работодатели для поиска работников по вакансии «Программист» в 2021 году («Выборка 1») с указанием изменений трендов по сравнению с 2020 («Выборка 2»). Визуализация представлена на рисунке 5.

Как видно из сравнения рисунков 2 и 5, навыки, которые указывают соискатели и которые желают видеть работодатели, отличаются, хотя картины, в целом, похожие. Следующие выборки, по которым будет построена количественная визуализация — гибкие навыки, которыми описывали себя соискатели на позицию «Программист», получившие приглашения на собеседования («Выборка 1») и выборка резюме без указания, было получено приглашение или нет («Выборка 2»). Визуализация представлена на рисунке 6.

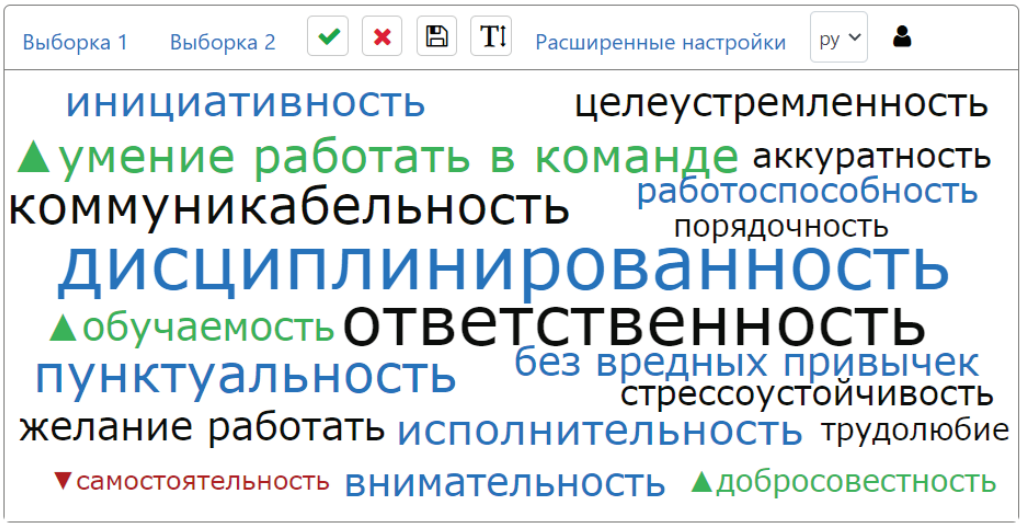

Рисунок 5 – Визуализация разницы между выборками вакансий на должность «Программист» в 2020 и 2021 году

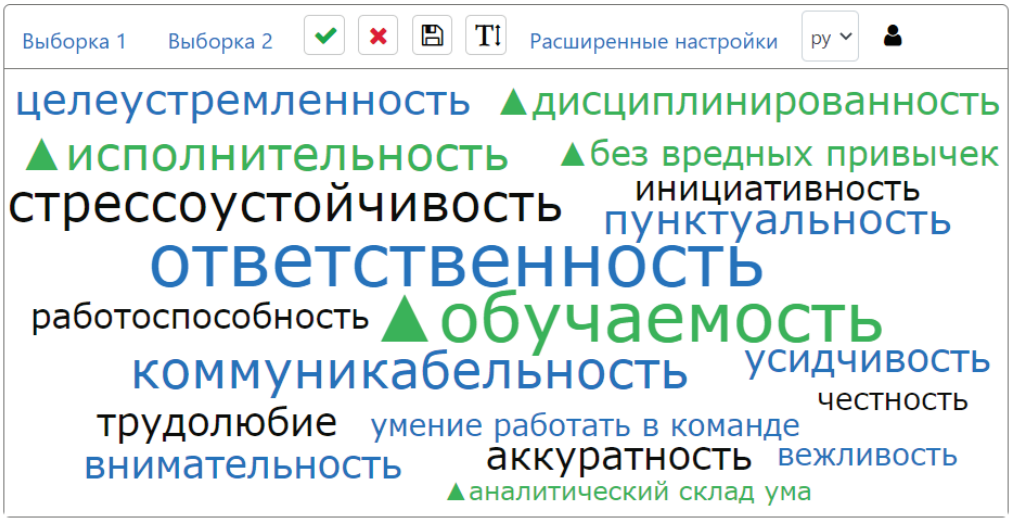

Рисунок 6 – Визуализация разницы между выборками резюме по критерию получения приглашения

Далее построим корреляционную модель с показателем «количество приглашений на собеседования» Данный параметр добавлен инженером данных на основе связной таблицы. Цель расчета корреляций – выяснить, есть ли связь между тем, как пользователь заполнил поле «гибкие навыки» и тем, как часто его приглашали на собеседования. В случае резюме на позицию «Программист» корреляций не было выявлено. Упоминания самых распространенных 20 навыков имели корреляцию с приглашениями на собеседования в диапазоне от -0.096 до 0.077. Подобная гипотеза является маловероятной и после более подробной проработки данных, что понизит её приоритет для проверки. В то же время, корреляционный анализ аналогичной по размеру (3500) выборки резюме соискателей на позицию «Вожатый» выявил некоторые корреляции различной силы связи (рисунок 7). Серым цветом отображаются языковые единицы, корреляция с упоминанием которых находится в диапазоне от -0.1 до 0.1, то есть отсутствует. Размеры шрифта отображают меру связи – чем больше шрифт, тем выше мера тесноты связи.

Таким образом, результаты разведочного анализа демонстрируют, что необходимость обработки и учета поля «гибкие навыки» возникает только при решении определенных задач, таких как оценка влияния указания в резюме «гибких навыков» на востребованность соискателя на вакансиях, связанных с работой с людьми.

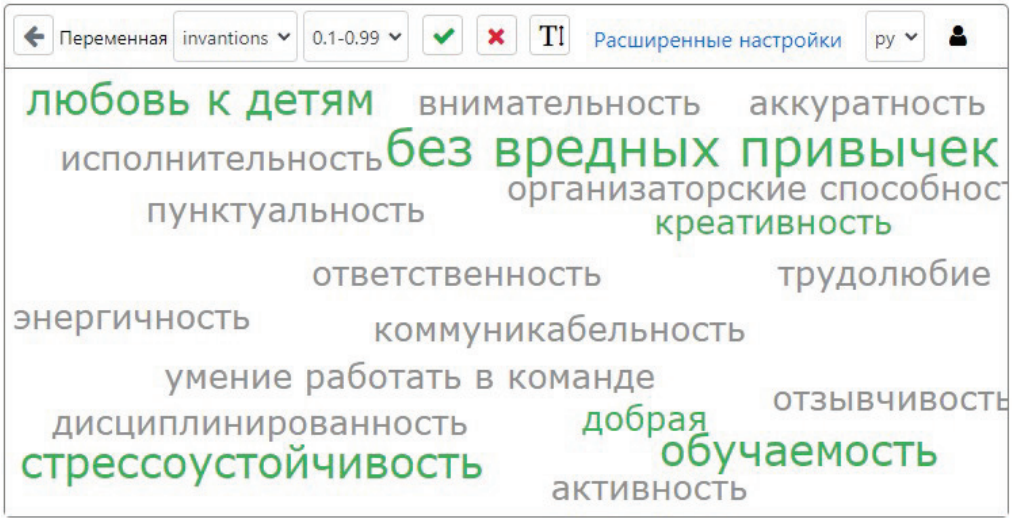

Рисунок 7 – Визуализация корреляций

В то же время, при оценке факторов, влияющих на востребованность технических специалистов, подобная информация может быть незначительной. Конечно, полученный на этом этапе результат еще не является ответом на вопрос, как влияет упоминание определенных слов в резюме на востребованность у потенциальных работодателей. Однако, он позволяет проверить некоторые гипотезы на этапе разведочного анализа и, отбросить их или принять в работу в порядке вероятности достоверности. Таким образом, использование разработанной визуализации позволит уменьшить время до получения первых подтвержденных результатов, а также когнитивный анализ визуализации может помочь исследователю составить новые гипотезы для проверки. Отбор только значимых слабоструктурированных текстовых данных для исследования позволяет улучшить качество моделей ИАД и увеличить скорость анализа данных.

3. Заключение

Перед построением моделей ИАД целесообразно проводить разведочный анализ данных – с целью погружения в данные, проверки или выдвижения обоснованных гипотез, отбора переменных и т.д. Однако, большая часть методов и инструментов для разведочного анализа, в том числе использующих визуализации, работает только с числовыми или категориальными данными.

Для решения данного противоречия были разработаны визуальные модели слабоструктурированных текстовых данных для использования с целью разведочного анализа. Данные модели отражают как количественных показатели текстовых данных и разницы между двумя выборками, так и корреляции между языковыми единицами и другими признаками исследуемых объектов.

Для демонстрации работы метода был проведен разведочный анализ на данных с сайта «Работа в России». Было выполнено исследование «гибких навыков», которые соискатели по различным направлениям используют в своих анкетах, как это указание меняется от года к году, в разных профессиях, а так же в резюме и вакансиях по одной профессии. Визуализация позволила выявить описания, которые использовали соискатели, получившие приглашения на собеседования. Визуализация связей позволила выяснить, что влияние указанных «гибких навыков» на количество приглашений на собеседование для программистов является незначимым. В то время как указание некоторых «гибких навыков» в резюме у соискателей на другие вакансии, например, вожатых, способствует повышению интереса к данному кандидату у работодателя, и могут иметь достоверную корреляцию с последующим приглашением на собеседование. Таким образом, результаты разведочного анализа демонстрирует, важность использования данных из поля «гибкие навыки» при построении моделей ИАД для решения некоторых задач.

Таким образом, разведочный анализ с использованием визуальных моделей позволяет сгенерировать гипотезы и выполнить первичную их проверку, помочь исследователю выбрать переменные для построения модели ИАД, сэкономив время и увеличив скорость работы модели ИАД, что особенно важно в областях, с динамично изменяющимися данными. Анализ корреляций в текстах может расширить возможности исследователей в большом круге областей, как например, филология, социология, психология и другие.

Дальнейшие исследования планируется направить на увеличение количества используемых методов обработки текстов на естественном языке – обработка отрицаний, поиск антонимов и прочее – с целью расширения работы с семантикой, что позволит улучшить качество моделей ИАД, включающих текстовые данные.

4. Список источников

[1] Description-text related soft information in peer-to-peer lending Evidence from two leading European platforms / Dorfleitner G. [et al] //. Journal of Banking & Finance. 2015. № 64. P. 169-187. DOI: 10.1016/j.jbankfin.2015.11.009.

[2] A Data Preparation Methodology in Data Mining Applied to Mortality Population Databases / Pérez J. [et al] // J Med Syst. 2015. №39. P.1173-1182. DOI: 10.1007/s10916-015-0312-5.

[3] Макарова Е.А., Лагерев Д.Г.. Автоматизация извлечения признаков из слабоструктурированных медицинских данных» // Информационные технологии и математические методы в экономике и управлении: сборник материалов X Международной научно–практическая конференция имени А. И. Китова. 15–16 октября 2020 г. – Москва : ФГБОУ ВО «РЭУ им. Г. В. Плеханова». 2020. С 56–62.

[4] Крылов В. С. Компьютерная лингвистика: разведочный анализ текстов научных публикаций. Информационно-компьютерные технологии в экономике, образовании и социальной сфере. 2022. № 2(36). С. 79-89.

[5] Захарова А. А., Шкляр А. В. Визуальные модели // Проблемы информатики. 2011. № 4. С. 41-47.

[6] Подвесовский А. Г., Лагерев Д. Г., Бабурин А. Н. Автоматизация процессов социологического исследования с использованием методов и программных средств интеллектуального анализа данных // Современные технологии в науке и образовании – СТНО-2017: сборник трудов II Международной научно-технической и научно-методической конференции: в 8 т., Рязань, 01–03 марта 2017 года – Рязань: Рязанский государственный радиотехнический университет. 2017. С. 122–127.

[7] Application of high-dimensional feature selection: evaluation for genomic prediction in man./ Bermingham M. [et al] // Sci Rep. 2015. № 5. DOI: 10.1038/srep10312

[8] Макарова Е.А., Лагерев Д.Г. Автоматизация извлечения признаков из слабоструктурированных медицинских данных» // X Международная научно–практическая конференция имени А. И. Китова «Информационные технологии и математические методы в экономике и управлении» (ИТиММ–2020). 15–16 октября 2020 г.: сборник статей. – Москва : ФГБОУ ВО «РЭУ им. Г. В. Плеханова», 2020. – С 56–62.

[9] Макарова Е. А. Лагерев Д. Г. Оценка семантической близости новостных сообщений на основе анализа заголовков // Вестник компьютерных и информационных технологий. 2021. Т. 18. № 7(205). С. 46–56.

[10] El-Hajj W, Hajj H. An optimal approach for text feature selection // Computer Speech & Language. 2022. №74. P 1-13. DOI: 10.1016/j.csl.2022.101364.

[11] Прокопьев А. В. Использование эконометрического инструментария таблиц сопряженности для оценки эффективности вакцинации // Здоровье – основа человеческого потенциала: проблемы и пути их решения. 2021. Т. 16. № 4. С. 1626–1632.

[12] Шишлянникова Л. Применение корреляционного анализа в психологии // Психологическая наука и образование. 2009. № 1. С. 98–107.

[13] Кравченко К. И., Минеева Т. А. Использование линейного коэффициента корреляции для определения характера связи между переменными // Тенденции развития науки и образования. 2022. № 82-2. С. 26–30.

[14] Герасимов А. Н., Шпитонков М. И. Доверительные границы к коэффициенту корреляции // Исследование операций (модели, системы, решения). 2020. Т. 6. С. 61–69. DOI: 10.14357/ORMSS20200108.

[15] «Работа в России»: обработанные и объединенные сведения о вакансиях, резюме, откликах и приглашениях портала trudvsem.ru [Электронный ресурс] // Роструд; обработка: Бабушкина В.О., Тимошенко А.Ш., Инфраструктура научно-исследовательских данных, АНО «ЦПУР», 2021. URL: http://data-in.ru/data-catalog/datasets/186/. (дата обращения: 22.04.2022).