БАЗА ДАННЫХ. СТАНДАРТИЗАЦИЯ И НОРМАЛИЗАЦИЯ ТЕКСТОВ ДЛЯ ИЗВЛЕЧЕНИЯ ИНФОРМАЦИИ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

ОГЛАВЛЕНИЕ

ПОДГОТОВКА ТЕКСТОВ К АНАЛИЗУ ДАННЫХ И ЗАПОЛНЕНИЕ ПРОПУСКОВ МЕТОДАМИ NATURAL LANGUAGE PROCESSING (NLP)

Каждый специалист в оценочной деятельности сталкивался с проблемой структурирования информации, получаемой из разных источников в текстовом формате. Именно в тексте содержится необходимая информация об объектах недвижимости, которая представляет ценность для формирования итоговой оценки стоимости данных объектов. Проблема состоит в том, что данная информация находится в тексте в неструктурированном произвольном виде, и для того, чтобы извлечь данную информацию из текста, оценщик тратит определенное количество времени. Для обработки огромного массива данных ему потребуется очень много времени на извлечение и структурирование полезной информации из текста, поэтому имеет смысл автоматизировать данный процесс. При грамотной обработке текстовых описаний объектов, можно формировать табличное представление, которое будет содержать всю полезную информацию об объектах, которая может быть извлечена из текстов. Данную информацию можно использовать для последующих исследований.

ФОРМИРОВАНИЕ ТРЕБОВАНИЙ К ДАННЫМ И БАЗЕ ДАННЫХ ДЛЯ ХРАНЕНИЯ ИНФОРМАЦИИ

В первую очередь, для дальнейшей работы необходимо сформировать требования к данным, базе данных и первичной обработке текстов, а также проанализировать полученный результат для перехода к дальнейшим этапам.

Данный этап подразумевает следующее:

1) Данные. На данном этапе формируются требования к данным: достоверные ли данные из выбранных источников информации, актуальная ли содержится информация и т.д.
2) База данных. На данном этапе формируются требования к базе данных: как лучше структурировать и хранить полученную информацию, как организовать доступ к данной информации и т.д.
3) Методы первичной обработки данных. На данном этапе формируются требования к методам первичной обработки данных: должны использоваться как проверенные, так и новейшие методы исследования в данной конкретной области, все результаты должны быть воспроизводимы и интерпретируемы и т.д.
4) Анализ требований. На данном этапе идет поиск путей удовлетворения требований на уровне концепции (архитектура, функции, программно-техническая платформа, используемые модули и т.п.), а также рассмотрение альтернативных вариантов концепции, их анализ и выбор лучшей

НОРМАЛИЗАЦИЯ ТЕКСТА И ПОДГОТОВКА ЕГО ДЛЯ ПОСЛЕДУЮЩЕЙ ОБРАБОТКИ

Для дальнейшей обработки и анализа текстовой информации необходимо привести её к определенному виду, другими словами, нормализовать тексты. Под нормализацией мы понимаем очистку данных от ненужных символов, замена различных сокращений на полный формат и т.п. Для обработки данных мы используем специально разработанные

функции, которые помогают обрабатывать данные равными частями размером 1000 текстов параллельно. Сами функции включают в себя:

1) Удаление url-адресов в объявлениях
2) Удаление специальных идеограмм, пиктограмм и смайликов
3) Транслитерация латинских символов в кириллицу по определенному стандарту (ГОСТ 7.79-2000, ГОСТ Р 52290-2004, ГОСТ Р 7.0.34-2014, ICAO DOC 9303, стандарт Википедии основанный на BGN/PCGN). Также есть возможность замены символов, у которых очень схожее написание. Например, латинская буква «c» и «с» написанная на кириллице
4) Замена знаков восклицания и вопросительного на знак «.»
5) Замена схожих по написанию знаков тире на единый формат «-»
6) Нормализация нестандартных символов
7) Функция для замены несколько подряд идущих символов пробела, знаков табуляции на один знак пробела
8) Преобразование чисел формата «сто пятьдесят два» на числовое значение «152»
9) Приведение текстов объявлений к нижнему регистру
10) Функция для замены орфографических ошибок и опечаток слов с дополнительными преобразованиями для анализа

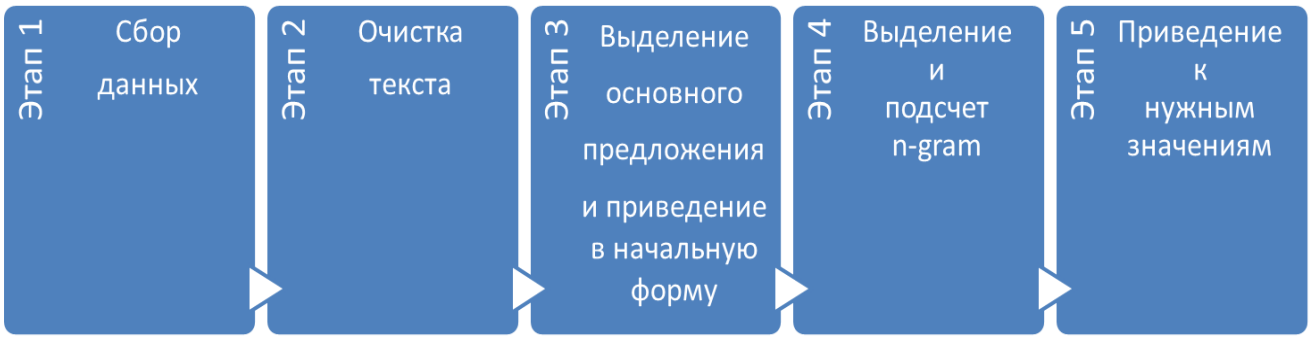
В случае возникновения ошибок различного рода на каждом этапе обработки предусмотрена возможность сохранения типа ошибки и этапа, где она возникает, для дальнейшей обработки.

Таблица 1. Пример преобразования начального варианта текста (слева) в конечный формат (справа)

ID:9791Предлагаются в аренду отапливаемые склады «B» класса. Общей площадью 3 208 м2 + 2 777 м2 в разных помещениях на 2 этаже. Рабочая высота 4,3 м. Ворота докового типа - 6, грузовые лифты г/п 3-5 тонн. Полы бетон с антипыльевым покрытием. Режим работы 24/7. Охраняемая парковка, телефония и интернет, система видеонаблюдения. Эксплуатационные платежи включены в ставку.	ид : 9791 предлагаются в аренду отапливаемые склады « в » класса . общей площадью 3208 кв . м . плюс 2777 кв . м . в разных помещениях на 2 этаже . рабочая высота 4 , 3 м . ворота докового типа - 6 , грузовые лифты г / п 3-5 тонн . полы бетон с антипылевым покрытием . режим работы 24/7 . охраняемая парковка , телефония и интернет , система видеонаблюдения . эксплуатационные платежи включены в ставку .
ОФИСНЫЕ ПОМЕЩЕНИЯ ОТ СОБСТВЕННИКА!При проектировании офисного комплекса мы тщательно исследовали все детали, которые важны арендаторам и устранили все мелочи, которые могут вызвать малейший дискамфорт. В этом офисе вам будет максимально комфортно.✔Большая парковка с ул. Уральской и на территории, свободные места гарантируем✔2 этаж, охрана, видеонаблюдение✔Вся необходимая для комфортной работы инфраструктура✔Звукоизолированные помещения✔Комфортный микроклимат в помещении благодаря вентиляции и кондиционированию✔Общая комната отдыха с кухней на этаже✔Бесплатный wi-fi+в здании отличный LTE-сигнал✔Возможность расширения помещения✔Рядом с СБС, очень интенсивный трафикОфис состоит из двух сквозных комнат по 22,4 кв.м. №10/2 и 10/3 на планеЗвоните!	офисные помещения от собственника . при проектировании офисного комплекса мы тщательно исследовали все детали , которые важны арендаторам и устранили все мелочи , которые могут вызвать малейший дискомфорт . в этом офисе вам будет максимально комфортно . большая парковка с ул . уральской и на территории , свободные места гарантируем 2 этаж , охрана , видеонаблюдение вся необходимая для комфортной работы инфраструктура звуко изолированные помещения комфортный микроклимат в помещении благодаря вентиляции и кондиционированию общая комната отдыха с кухней на этаже бесплатный ви-фи плюс в здании отличный лте-сигнал возможность расширения помещения рядом с сбс , очень интенсивный трафик офис состоит из 2 сквозных комнат по 22 , 4 кв . м . no 10/2 и 10/3 на плане звоните .

ФОРМИРОВАНИЕ СТАНДАРТИЗОВАННОЙ БАЗЫ ДАННЫХ

Процесс формирования базы данных включает следующие этапы:

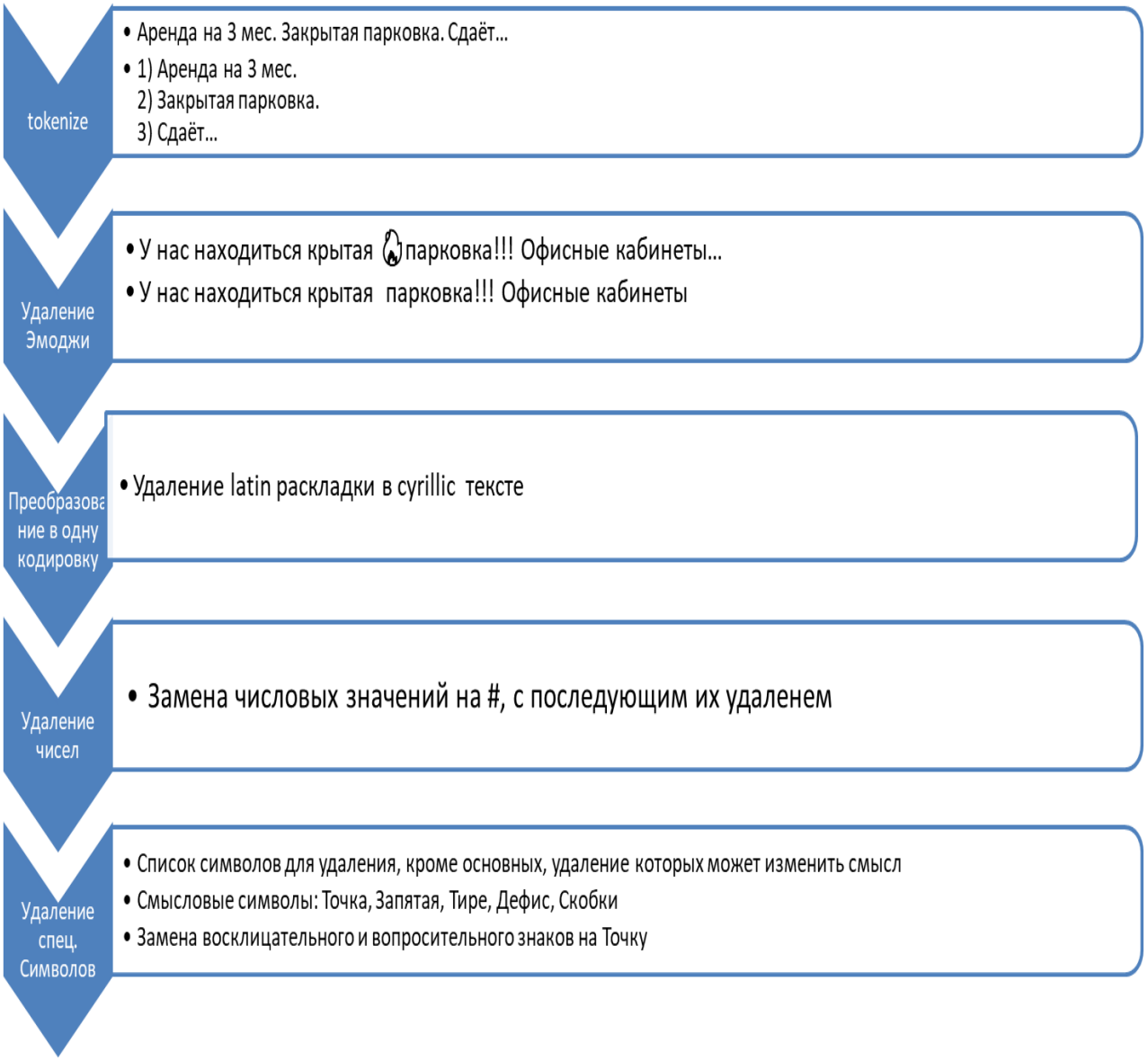


Сбор данных осуществлялся по следующим источникам рыночных данных:

Третий

Domofond	Move	Afy_sale	Afy_rent
• link • description	• link • opisanie = description	• link • description	• link • description

Третий этап (очистка текста) наиболее сложный. Ниже он представлен в виде отдельных подэтаповю



АЛГОРИТМЫ ЗАПОЛНЕНИЯ ПРОПУСКОВ С ПОМОЩЬЮ МЕТОДОВ NATURAL LANGUAGE PROCESSING (NLP)

На данном этапе представлены методы и алгоритмы анализа текстов, полученных после нормализации, и алгоритмы заполнения пропущенных значений, характеристик объектов недвижимости из текстовых описаний.

Поиск синонимических групп

Для полного понимания, какими значениями выражается определенная характеристика объектов, необходимо составить полный перечень возможных слов или словосочетаний, которые могут дать подробную информацию по характеристике. Такими словами могут выступать так называемые синонимические группы – это группы слов, состоящие из разных частей речи, способные описывать характеристику исследуемого

объекта недвижимости путем употребления их в тексте вместе со словами, которые имеют оценочное значение рассматриваемой характеристики.

Для поиска синонимичных групп используется модуль обработки естественного языка genism. Из данного модуля нам необходима модель word2vec, которая подразумевает обучение нейронной сети определять по контексту слова близкие по значению. У данной модели реализована функция show similar (), которая позволяет получить первые N слов, которые очень часто встречались рядом с нашим словом, поданным на вход функции. Набор слов может содержать как синонимы к данному слову и образовывать синонимичную группу, так и обратно противоположные по значению – антонимы, либо нейтральные – те слова, которые могут образовывать с искомым словом устойчивое словосочетание, которое также может выражать характеристику исследуемого объекта недвижимости. Из данного перечня слов необходимо выбрать все слова, которые могут образовывать синонимическую группу. На изображении ниже представлено векторное пространство объектов обученной модели word2vec на корпусе текстовых описаний объектов недвижимости за период с февраля 2019 года по февраль 2021 года, где показан пример поиска похожих по значению слов к искомому «Парковка» (красный цвет) и дополнительные слова для сравнения их схожести (зеленый цвет). Данные поиск осуществляется по заданному входящему слову и выводит первые N объектов векторного пространства, которые очень часто встречались в похожем контексте и способны описывать похожие характеристики объекта. Чем ближе к нашему искомому слову находятся другие слова, тем больше вероятность того, что данные слова составляют синонимическую группу.

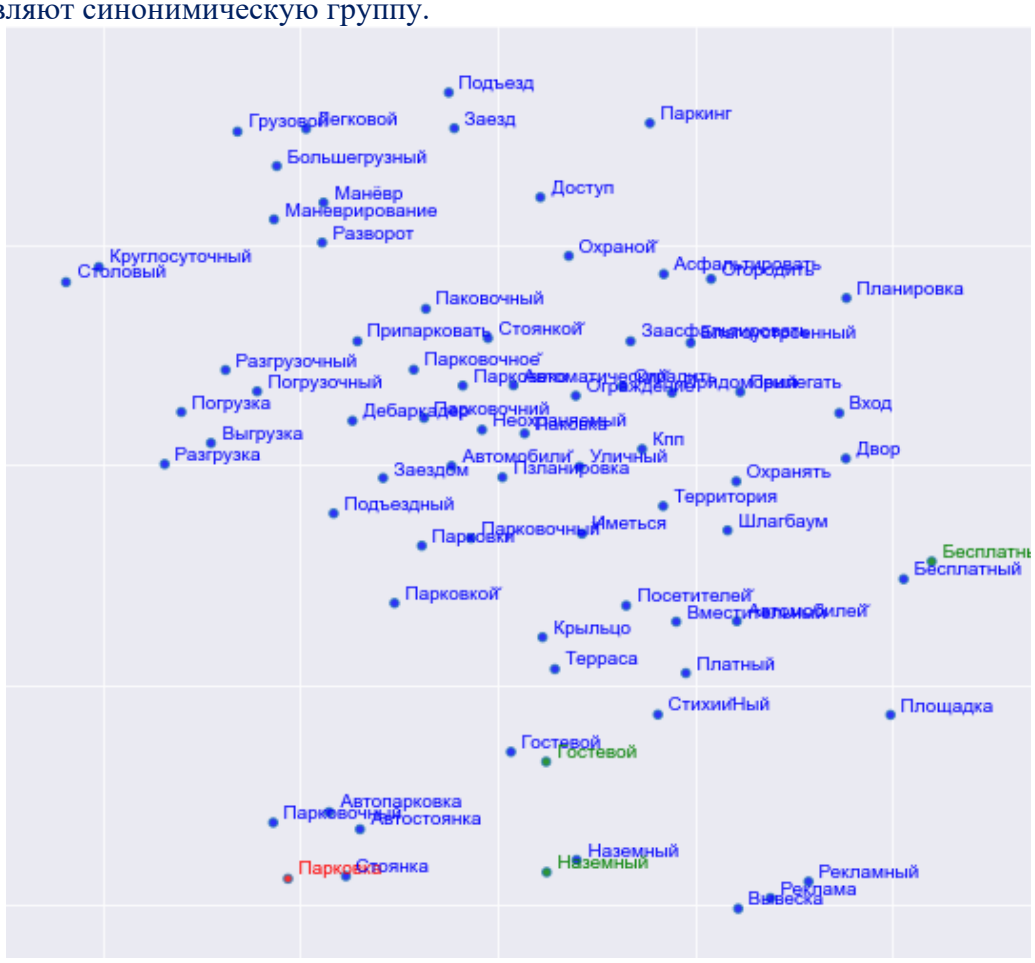


Рис. 1 Представление слов в векторном пространстве

ЗАПОЛНЕНИЕ ПРОПУСКОВ

Для заполнения пропусков были разработаны несколько алгоритмов, способных достаточно точно и правильно находить пропущенные значения характеристик исследуемых объектов недвижимости.

Алгоритм №1

Сформировав конечный список слов синонимичной группы, нам необходимо определить принадлежность каждого из слов к определенному значению. Например, для парковки этот список выглядит следующим образом (Значения из справочника) :

1) Для грузового транспорта
2) Организованная подземная
3) Организованная наземная крытая
4) Стихийная наземная
5) Организованная наземная открытая

Для определения принадлежности парковки к определенному значению из справочника, необходимо:

- Сформировать массив значений синонимичной группы для автоматической генерации регулярного выражения по поиску данных значений и слов, находящихся слева и справа от искомого.
- Взять текстовое описание объектов недвижимости и привести все слова из данных описаний в начальную форму, это необходимо для уменьшения размерности так называемого мешка слов (Bag-of-Words). Данная функция выполняется с помощью модуля pymystem3.
- Далее идет поиск всех слов из синонимичной группы и формированием биграмм слов, состоящих из искомого значения и слов, которые стоят слева и справа от искомого.
- На данном шаге мы задаем словари с преобразованием биграмм к значениям из справочника с использованием правил замены.
- На последнем шаге идет формирование конечных значений из справочника и запись в базу данных.

Алгоритм №2

Данный алгоритм основан на анализе синтаксического дерева разбора предложения, где содержится искомое слово из синонимичной группы.

Необходимо выполнить следующие шаги:

- Для каждого текстового описания идет поиск всех предложений, где содержится любое из слов синонимичной группы.

- Для данных предложений строится дерево синтаксического разбора, где вершинами графа будут сами слова из предложений, а ребра обозначают связи между словами в предложении.
- На данном этапе с помощью специального алгоритма подсчитывается граф (зеленый) заданной глубины, где считается, что искомое слово – это первый уровень, а последующие связи второй, третий и т.д. Пример:

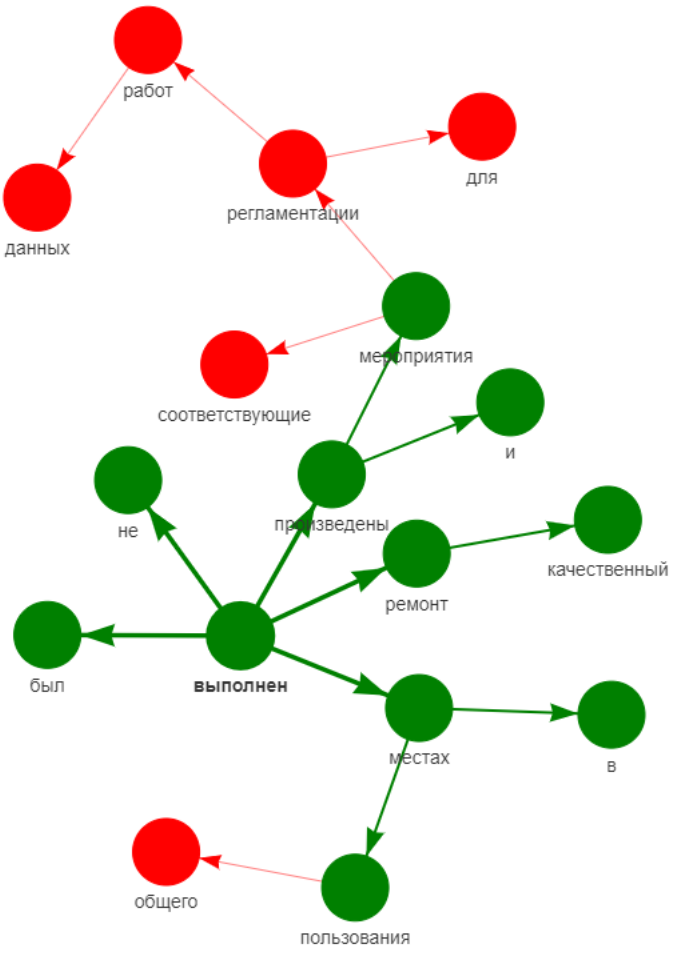


Рис. 2 Синтаксическое дерево разбора предложения «Для регламентации данных работ соответствующие мероприятия произведены и не был выполнен качественный ремонт в местах общего пользования»

В данном случае искомым словом было “ремонт”, так как в предложении есть сложные структуры, которые описывают не только состояние ремонта, но и тот факт, что ремонт может быть выполнен не в полном объеме, не выполнен вообще и т.п.

- Далее ведется поиск определенной структуры в графе, которая определяется определенной последовательностью частей речи, имеющие определенные связи в данной структуре. Если есть совпадение, то присваивается определенное значение искомого параметра, в соответствие со значениями в справочнике.

Данный метод имеет возможность извлекать нужную информацию, основываясь не на определенных заданных словах, которые характеризуют определенную степень исследуемой характеристики, а на значении слов, которые могут описывать нужную структуру. Это позволяет извлекать информацию исходя из связей в предложении. Пример:

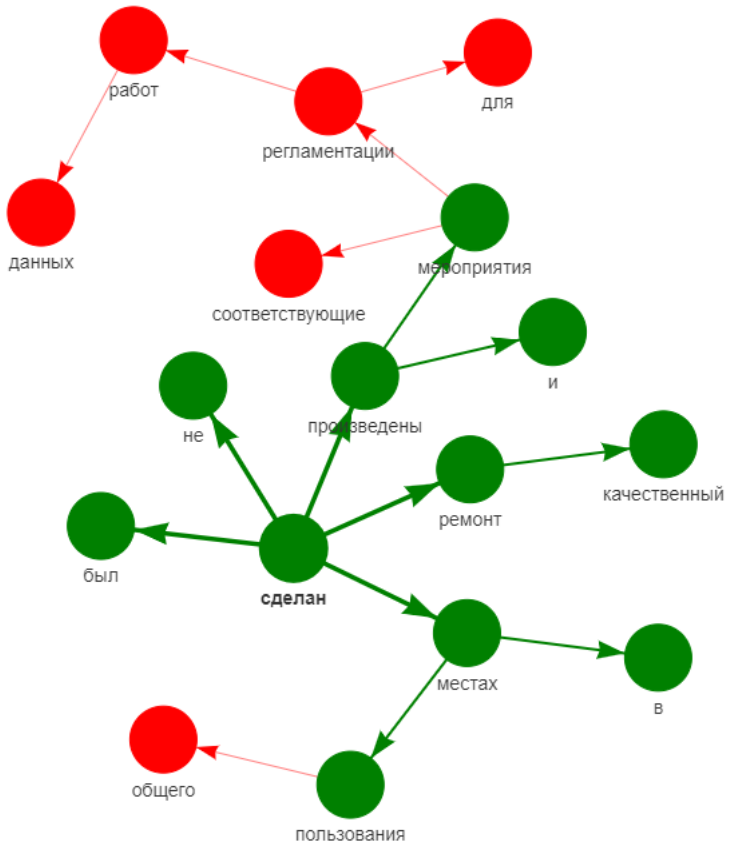


Рис. 3 Синтаксическое дерево разбора предложения «Для регламентации данных работ соответствующие мероприятия произведены и не был сделан качественный ремонт в местах общего пользования»

Пример иллюстрирует, что от замены глагола “выполнен” на “сделан”, дерево разбора осталось таким же, что позволяет применить правила для поиска похожих структур для выражения текущего состояния параметра.

Окончательно сформированная база данных позволяет сделать выгрузку данных в виде электронных таблиц требуемой формы. Для этого в рамках данного исследования была разработана программа.

В программе предусмотрено:

1. Формирование признаков, в соответствии с которыми отбираются (фильтруются) объявления и формируется дата сет.

Выбирайте значения для поиска		Дата поиска начиная с			Сортировка
Параметр	**Значения**	1	Январь	2021	По убыванию
Тип сделки	Продажа				
Город					
Сегмент	Земельный участок				
Класс					
Форма					
Рельеф					
Асфальтовая дорогая					
Охрана					
Ограждение					
Электричество					
Водоснабжение					
Канализация					
Газоснабжение					
Источник					

2. Выполнение выгрузки в виде электронной таблицы.

3. Далее, если это требуется, формирование отчета с использованием удобных средств визуализации, содержащего общую информацию о дата сете. Ниже приведен пример отдельного фрагмента отчета.

Отчет по количеству объектов в сегменте 'Офисная недвижимость'		
Город	**Аренда**	**Продажа**
Всего	125 758	16 329
Итог	142 087	
Волгоград	593	200
Воронеж	1 105	232
Екатеринбург	2 765	1 050
Казань	2 203	492
Москва	96 600	9 628
Московская область	3 551	736
Нижний Новгород	1 776	336
Новосибирск	2 185	639
Пермь	1 111	410

Ростов-на-Дону	1 561	468
Самара	1 621	669
Санкт-Петербург	9 307	1 049
Челябинск	1 380	420