6. РАСПОЗНАВАНИЕ ТЕКСТОВ ОБЪЯВЛЕНИЙ И СТРУКТУРИ-РОВАНИЕ РЫНОЧНОЙ ИНФОРМАЦИИ

- Одной из наиболее сложных задач, с которой сталкивается специалист в оценочной деятельности, является структурирование информации, получаемой из текста объявления. Дело в том, что наиболее значимая информация о стоимости объекта оценки заключена в содержании объявления, которое обычно составляется в произвольной форме в соответствии с личными предпочтениями автора объявления (продавца, риелтора и т.п.).
- Для того, чтобы иметь возможность обрабатывать огромные массивы такой информации в автоматическом режиме, ее следует представить в виде таблиц. При грамотной обработке описание объекта, содержащееся в тексте объявления, можно использовать для формирования нужных признаков и представления их в виде таблиц, на основе которых выполняется последующая обработка.
- В процессе выполнения данного исследования выделялись различные релевантные признаки. К релевантным данном исследовании отнесены признаки, которые в наибольшей степени влияют на стоимость и ликвидность. Это позволило представить описание объекта в структурированном виде и обеспечить дальнейшую обработку этой информации.
- В таблицах ниже представлен перечень ценообразующих факторов, а также значений, которые они принимают, которые в совокупности представляют собой основу базы данных, которая подвергается дальнейшей обработке методами машинного обучения.

Офисная недвижимость

Группа факторов	Ценообразующий фактор		Значения ценообразующих факторов
Общие признаки	Ссылка		
	Источник		
	Телефон		
	Ссылка на скриншот		
	Заголовок объявления		
	Описание		
	Сегмент		Офисная недвижимость
	Тип сделки		Продажа
			Аренда
Признаки местоположения	Город		
	Точный адрес		
	Координаты		
	Название административного района		
	Территориальная зона (маленькая)		
	Кластер (географический)		
	Кластер (функционально-географический)		
	Территориальная зона (большая)		Исторический центр города
			Центры деловой активности
			Многоквартирная жилая застройка
			ИЖС
			Промзоны
			Окраины
			Зона автомагистралей
			Зеленая зона
			Многофункциональные зоны
			Объекты соц.назначения
			Зона путей морского сообщения
			Прибрежно-курортная зона
	Красная линия		Красная линия
			Внутриквартально
	Точки интереса	Ближайшая точка (наименование)	Школа
			Детский сад
			Станция метро
		Расстояние до точки, напрямую в метрах	Остановка общественного транспорта
			Поликлиника/Больница
			Центр города (Культурный и политический)
		Расстояние до точки, пешком в метрах	Парки/скверы/лесные массивы
			Море (для курортных регионов)
			МКАД (для Московской области)
		Расстояние до точки, пешком в мин.	КАД (для Ленинградской области)
			Кладбище

Офисная недвижимость

Группа факторов	Ценообразующий фактор	Значения ценообразующих факторов
Тип объекта	Тип объекта	Комплекс зданий
		Отдельностоящее здание
		Встроенное помещение
Технические характеристики здания	Общая площадь офисного здания, кв.м	
	Общая площадь складского комплекса, кв.м	
	Год постройки	
	Класс	A
		B (B+,B-)
		A+B (если объект высококлассный, но мы не можем отнести ни к A, ни к B)
		C
		объекты вне классификации
	Тип здания	Офисный/Бизнес центр
		Административное здание
		Офисно-торговый комплекс
		Офисно-складской комплекс
		Офисно-жилой комплекс
		Многофункциональный комплекс
		Особняк
		Жилой дом
	Бизнес центр (наименование)	
	Этажность	
	Материал стен	Кирпичные
		Железобетонные (панельные, блочные)
		Монолитные
		Сэндвич-панели
		Прочие
	Доступ	Свободный
		Закрытый (пропускная система)
	Охрана	Круглосуточная
		Видеонаблюдение
		Нет
	Парковка	Подземная
		Наземная крытая
		Наземная открытая (организованная, для легкового транспорта)
		Стихийная
		Для грузового транспорта
		Отсутствует

Офисная недвижимость

Группа факторов	Ценообразующий фактор		Значения ценообразующих факторов
Технические признаки помещения	Общая площадь, кв.м		
	Назначение		Офис
			Прочее
	Этаж		Первый
			2 и выше
			Цоколь
			Подвал
			Мансарда
	Состояние отделки	Первичный рынок	Без отделки
			Под чистовую отделку
			Муниципальный ремонт
			Современный ремонт
			Дизайнерский ремонт
		Вторичный рынок	Требуется капитальный ремонт
			Требуется косметический ремонт
			Эконом
			Современный ремонт
			Дизайнерский ремонт
	Высота потолков, м		
	Планировка		Open-space
			Смешанная (open-space и коридорная)
			Коридорная
	Отдельный вход		Есть
			Нет

Офисная недвижимость

Группа факторов	Ценообразующий фактор	Значения ценообразующих факторов
Ценовые признаки	Цена предложения, руб.	
	Удельная цена, руб./кв.м (первоначальная)	
	Удельная цена, руб./кв.м (последняя)	
	Эксплуатационные расходы	
	Коммунальные в цене (для аренды)	Не включено
		Включено частично
		Все включено
	Залог, руб.	
	Налоговые платежи	
Временные признаки	Объявление участвует в ликвидности	True
		False
	Дата создания	
	Дата парсинга	
	Статус объявления	'white' - когда у объявления отсутствует отметка о том что оно снято или продано на сайте
		‘gray’ - когда у объявления присутствует отметка о том что оно снято или продано на сайте и дата последнего парсинга менее 28 дней (4 недель)
		'black' - когда у объявления присутствует отметка о том что оно снято или продано на сайте и дата последнего парсинга более 28 дней (4 недель)
	Дата снятия	
	Количество просмотров в неделю	

Торговая недвижимость

Группа факторов	Ценообразующий фактор		Значения ценообразующих факторов
Общие признаки	Ссылка		
	Источник		
	Телефон		
	Ссылка на скриншот		
	Заголовок объявления		
	Описание		
	Сегмент		Торговая недвижимость
	Тип сделки		Продажа
			Аренда
Признаки местоположения	Город		
	Точный адрес		
	Координаты		
	Название административного района		
	Территориальная зона (маленькая)		
	Кластер (географический)		
	Кластер (функционально-географический)		
	Территориальная зона (большая)		Исторический центр города
			Центры деловой активности
			Многоквартирная жилая застройка
			ИЖС
			Промзоны
			Окраины
			Зона автомагистралей
			Зеленая зона
			Многофункциональные зоны
			Объекты соц.назначения
			Зона путей морского сообщения
			Прибрежно-курортная зона
	Красная линия		Красная линия
			Внутриквартально
	Точки интереса	Ближайшая точка (наименование)	Школа
			Детский сад
			Станция метро
		Расстояние до точки, напрямую в метрах	Остановка общественного транспорта
			Поликлиника/Больница
			Центр города (Культурный и политический)
		Расстояние до точки, пешком в метрах	Парки/скверы/лесные массивы
			Море (для курортных регионов)
			МКАД (для Московской области)
		Расстояние до точки, пешком в мин.	КАД (для Ленинградской области)
			Кладбище

Торговая недвижимость

Группа факторов	Ценообразующий фактор	Значения ценообразующих факторов
Тип объекта	Тип объекта	Комплекс зданий
		Отдельностоящее здание
		Встроенное помещение
Технические характеристики здания	Общая площадь торгового комплекса, кв.м	
	Общая площадь офисного здания, кв.м	
	Торговая площадь (GLA), кв.м	
	Год постройки	
	Класс	торгово-развлекательные центры и торговые центры;
		районный/микрорайонный торговый центр;
		стрит-ретейл
		торговые площади
	Тип здания	Многофункциональный комплекс
		Офисно-торговый комплекс
		Торговый комплекс
		Жилой дом
		Отдельностоящее здание магазина
	Наименование ТЦ	
	Этажность	
	Якорные арендаторы	
	Материал стен	Кирпичные
		Железобетонные (панельные, блочные)
		Монолитные
		Сэндвич-панели
		Прочие
	Охрана	Круглосуточная
		Видеонаблюдение
		Нет
	Парковка	Подземная
		Наземная крытая
		Наземная открытая (организованная, для легкового транспорта)
		Стихийная
		Для грузового транспорта
		Отсутствует

Торговая недвижимость

Группа факторов	Ценообразующий фактор		Значения ценообразующих факторов
Технические признаки помещения	Общая площадь, кв.м		
	Назначение		
	Этаж		Первый
			2 и выше
			Цоколь
			Подвал
			Мансарда
	Состояние отделки	Первичный рынок	Без отделки
			Под чистовую отделку
			Муниципальный ремонт
			Современный ремонт
			Дизайнерский ремонт
		Вторичный рынок	Требуется капитальный ремонт
			Требуется косметический ремонт
			Эконом
			Современный ремонт
			Дизайнерский ремонт
	Высота потолков, м		
	Отдельный вход		Есть
			Нет
	Витринные окна		Есть
			Нет

Торговая недвижимость

Группа факторов	Ценообразующий фактор	Значения ценообразующих факторов
Ценовые признаки	Цена предложения, руб.	
	Удельная цена, руб./кв.м (первоначальная)	
	Удельная цена, руб./кв.м (последняя)	
	Эксплуатационные расходы	
	Коммунальные в цене (для аренды)	Не включено
		Включено частично
		Все включено
	Залог, руб.	
	Налоговые платежи	
Временные признаки	Объявление участвует в ликвидности	True
		False
	Дата создания	
	Дата парсинга	
	Статус объявления	'white' - когда у объявления отсутствует отметка о том что оно снято или продано на сайте
		‘gray’ - когда у объявления присутствует отметка о том что оно снято или продано на сайте и дата последнего парсинга менее 28 дней (4 недель)
		'black' - когда у объявления присутствует отметка о том что оно снято или продано на сайте и дата последнего парсинга более 28 дней (4 недель)
	Дата снятия	
	Количество просмотров в неделю	

Производственно-складская недвижимость

Группа факторов	Ценообразующий фактор		Значения ценообразующих факторов
Общие признаки	Ссылка		
	Источник		
	Телефон		
	Ссылка на скриншот		
	Заголовок объявления		
	Описание		
	Сегмент		Производственно-складская недвижимость
	Тип сделки		Продажа
			Аренда
Признаки местоположения	Город		
	Точный адрес		
	Координаты		
	Название административного района		
	Территориальная зона (маленькая)		
	Кластер (географический)		
	Кластер (функционально-географический)		
	Территориальная зона (большая)		Исторический центр города
			Центры деловой активности
			Многоквартирная жилая застройка
			ИЖС
			Промзоны
			Окраины
			Зона автомагистралей
			Зеленая зона
			Многофункциональные зоны
			Объекты соц.назначения
			Зона путей морского сообщения
			Прибрежно-курортная зона
	Красная линия		В непосредственной близости от автомагистрали
			На удалении от магистрали
	Точки интереса	Ближайшая точка (наименование)	Автомагистраль
		Расстояние до точки, напрямую в метрах	Железнодорожные пути
		Расстояние до точки, пешком в метрах	Центр города
		Расстояние до точки, пешком в мин.	МКАД (для Московской области)
			КАД (для Ленинградской области)

Производственно-складская недвижимость

Группа факторов	Ценообразующий фактор	Значения ценообразующих факторов
Тип объекта	Тип объекта	Комплекс зданий
		Отдельностоящее здание
		Встроенное помещение
Технические характеристики здания	Общая площадь здания, кв.м	
	Общая площадь складского комплекса, кв.м	
	Размер офисной части, кв.м	
	Размер прилегающей территории, га. (Площадь земельного участка, га)	
	Год постройки	
	Физическое состояние здания	Хорошее
		Удовлетворительное
		Неудовлетворительное
	Класс	А
		В (В+, В-)
		А+В (если объект высококлассный, но мы не можем отнести ни к А, ни к В)
		С
		объекты вне классификации
	Тип здания	Складской комплекс
		Приозводственно-складской комплекс
		Офисно-складской комплекс
		Производственный комплекс
		Склад
		Производство
		Многофункциональный комплекс
		Отдельностоящее здание
		Ангар
	Этажность	
	Материал стен	Кирпичные
		Железобетонные (панельные, блочные)
		Монолитные
		Сэндвич-панели
		Металлические
		Прочие
	Огороженная территория (забор)	Есть
		Нет
	Наличие Ж/Д ветки (на участке)	Есть
		Нет
	Доступ	Свободный
		Закрытый

Производственно-складская недвижимость

Группа факторов	Ценообразующий фактор	Значения ценообразующих факторов
Технические характеристики здания	Охрана	Круглосуточная
		Видеонаблюдение
	Автоматические ворота	Есть
		Нет
Технические признаки помещения	Общая площадь, кв.м	
	Назначение	
	Этаж	Первый
		2 и выше
		Цоколь
		Подвал
		Мансарда
	Состояние отделки	Хорошая
		Удовлетворительная
		Неудовлетворительная
	Высота потолков, м	
	Материал пола	Антипылевое покрытие
		Бетон
		Дерево
		Прочее
	Материал перекрытий	Бетон
		Кирпич
		Смешанный
	Отопление	Есть
		Нет
	Мощность электросети, кВт	
	Водоснабжение	Есть
		Нет
	Количество мокрых точек, шт.	
	Грузоподъемной механизм	Есть
		Нет
	Холодильная камера	Есть
		Нет

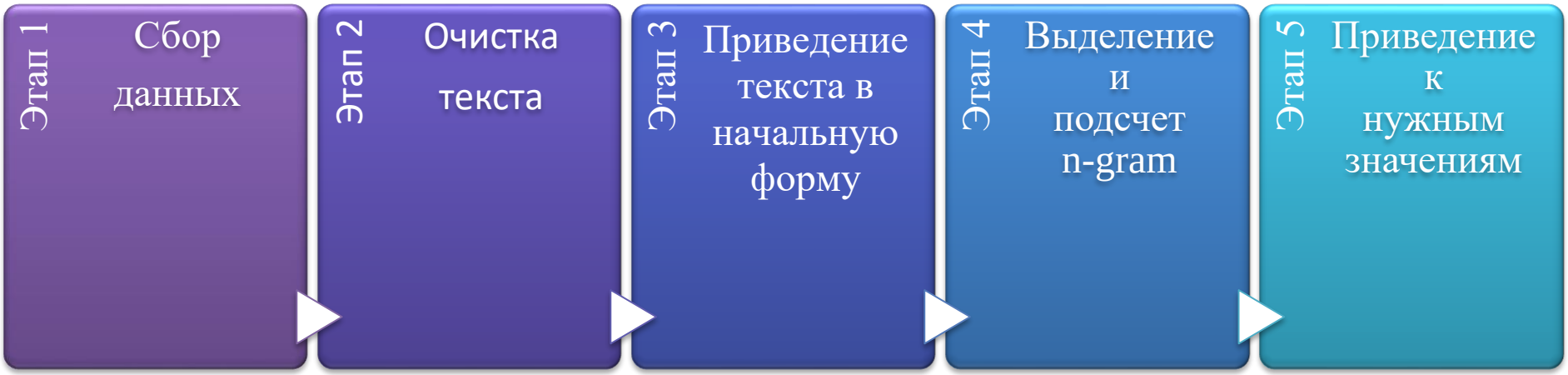
Производственно-складская недвижимость

Группа факторов	Ценообразующий фактор	Значения ценообразующих факторов
Ценовые признаки	Цена предложения, руб.	
	Удельная цена, руб./кв.м (первоначальная)	
	Удельная цена, руб./кв.м (последняя)	
	Эксплуатационные расходы	
	Коммунальные в цене (для аренды)	Не включено
		Включено частично
		Все включено
	Залог, руб.	
	Налоговые платежи	
Временные признаки	Объявление участвует в ликвидности	True
		False
	Дата создания	
	Дата парсинга	
	Статус объявления	'white' - когда у объявления отсутствует отметка о том что оно снято или продано на сайте
		‘gray’ - когда у объявления присутствует отметка о том что оно снято или продано на сайте и дата последнего парсинга менее 28 дней (4 недель)
		'black' - когда у объявления присутствует отметка о том что оно снято или продано на сайте и дата последнего парсинга более 28 дней (4 недель)
	Дата снятия	
	Количество просмотров в неделю	

АЛГОРИТМ ОБРАБОТКИ НЕЧИСЛОВЫХ ПАРАМЕТРОВ

Гипотеза распределения. Слова, которые встречаются рядом друг с другом и используются в одних и тех же контекстах, имеют тенденцию иметь схожие значения.

ЭТАПЫ ОБРАБОТКИ ПРИЗНАКА



СБОР ДАННЫХ

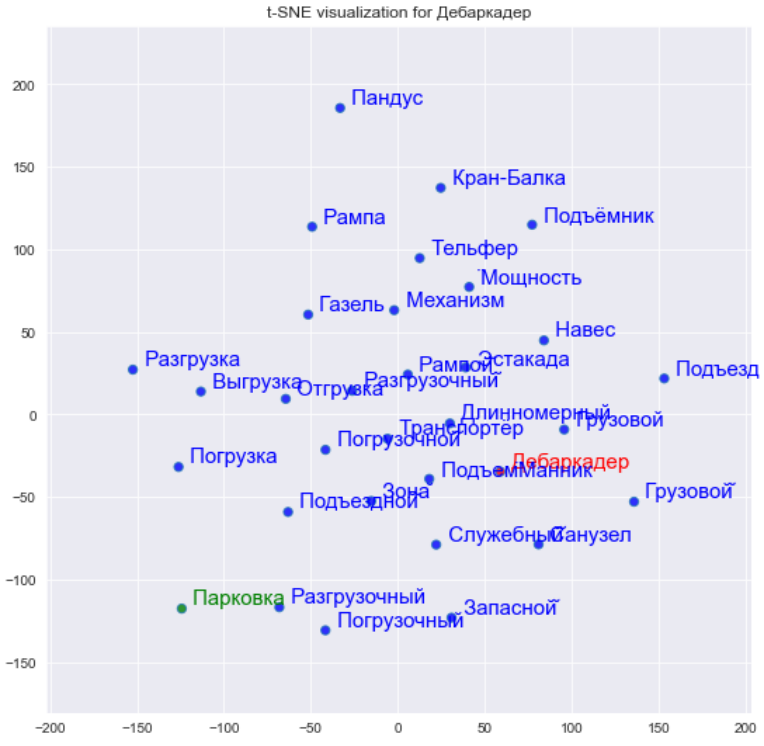
Domofond	Move	Afy	Avito
• link • description	• link • opisanie = description	• link • description	• link • description

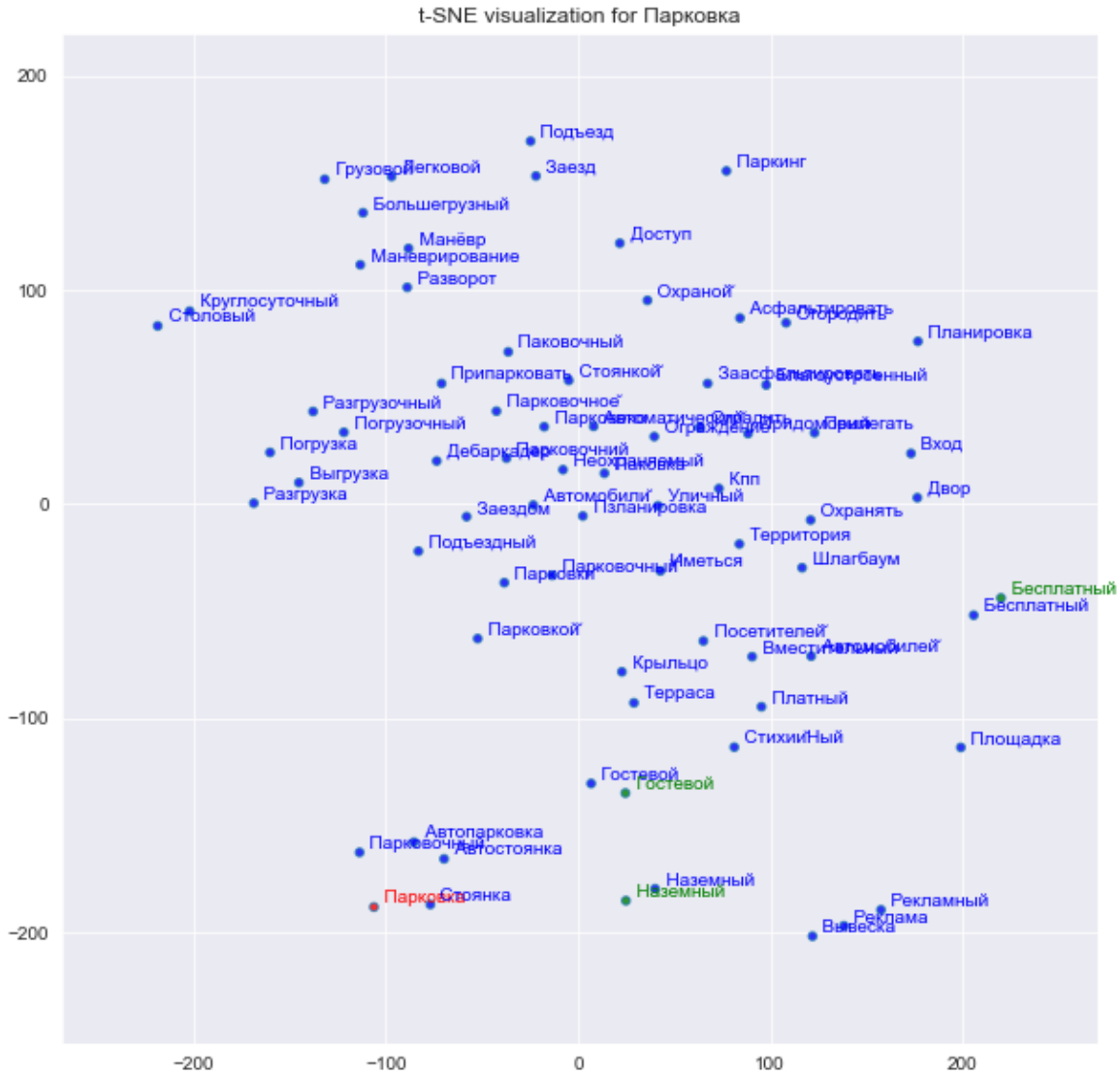
Количество объектов в объединенной выборке: **более 2 млн. записей**

ОЧИСТКА ТЕКСТА

1) Удаление url-адресов
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

Word2Vec


t-SNE visualization for Дебаркадер
Пандус
Кран-Балка
Рампа
Подъёмник
Тельфер
Мощность
Газель
Механизм
Навес
Разгрузка
Эстакада
Выгрузка
Отгрузка
Разгрузочный
Подъезд
Длинномерный
Транспортер
Грузовой
Погрузочной
Погрузка
Дебаркадер
Подъемманник
Зона
Грузовой
Подъездной
Служебный
Санузел
Парковка
Разгрузочный
Запасной
Погрузочный


t-SNE visualization for Парковка
Подъезд
Грузовой
Легковой
Заезд
Паркинг
Большегрузный
Манёвр
Доступ
Маневрирование
Разворот
Охраной
Круглосуточный
Столовый
Планировка
Паковочный
Припарковать
Стоянкой
Разгрузочный
Парковочное
Погрузочный
Погрузка
Вход
Выгрузка
Разгрузка
Кпп
Автомобили
Уличный
Двор
Заездом
Охранять
Территория
Подъездный
Шлагбаум
Парковки
Парковочный
Иметься
Бесплатный
Парковкой
Посетителей
Крыльцо
Терраса
Платный
Площадка
Гостевой
Автопарковка
Автостоянка
Наземный
Парковка
Стоянка
Рекламный
Реклама
Вывеска

ВЫДЕЛЕНИЕ ОСНОВНОГО ИНФОРМАТИВНОГО ПРЕДЛОЖЕНИЯ И ПРИВЕДЕНИЕ В НАЧАЛЬНУЮ ФОРМУ

Функция выделения основного предложения работает на поиске опорных слов в предложении:

```
"(?:\S*ремонт\S*|\S*отделк\S*|\S*реконструкц\S*)|(?:\S*Ремонт\S*|\S*Отделк\S*|\S*Реконструкц\S*)"
```

	description	link	sentences	true_sentence
6	Аренда . Офисные кабинеты в ЖК Олимп . Распол...	https://bashkortostan.afy.ru/ufa/snyat-ofis/60...	[Аренда, Офисные кабинеты ЖК Олимп, Расположен...	Расположены на втором этаже Общии вход этажа П...
10	Описание Бизнесцентра Удобные выезды на Дмитр...	https://afy.ru/moskva/snyat-ofis/60001208784	[Описание Бизнесцентра Удобные выезды на Дмитр...	Отделка офиса предусмотрена индивидуально по з...

Приведение в начальную форму:

1) Проверка и исправление правописания в предложении
2) Удаление повторяющихся слов
3) Приведения слов в начальную форму, исходя из текста: Отделка офиса -> отделка офис
4) Удаление стоп-слов, кроме несущих смысловую нагрузку: частица не, нет и др.

	description	link	sentences	true_sentence	initial_sentence
6	Аренда . Офисные кабинеты в ЖК Олимп . Распол...	https://bashkortostan.afy.ru/ufa/snyat-ofis/60...	[Аренда, Офисные кабинеты ЖК Олимп, Расположен...	Расположены на втором этаже Общии вход этажа П...	расположить второе этаж общий вход этаж план п...
10	Описание Бизнесцентра Удобные выезды на Дмитр...	https://afy.ru/moskva/snyat-ofis/60001208784	[Описание Бизнесцентра Удобные выезды на Дмитр...	Отделка офиса предусмотрена индивидуально по з...	отделка офис предусмотреть индивидуально зака...

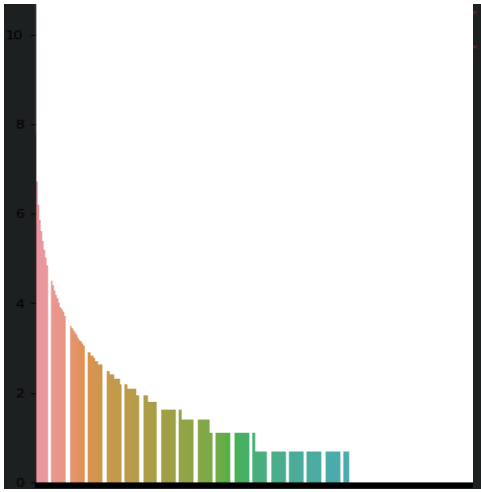
ВЫДЕЛЕНИЕ ОСНОВНОГО ИНФОРМАТИВНОГО ПРЕДЛОЖЕНИЯ И ПРИВЕДЕНИЕ В НАЧАЛЬНУЮ ФОРМУ

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

ID:9791Предлагаются в аренду отапливаемые склады «В» 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 антипыдевым 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 дте-сигнал возможность расширения помещения рядом с сбс , очень интенсивный трафик офис состоит из 2 сквозных комнат по 22 , 4 кв . м . по 10/2 и 10/3 на плане звоните .

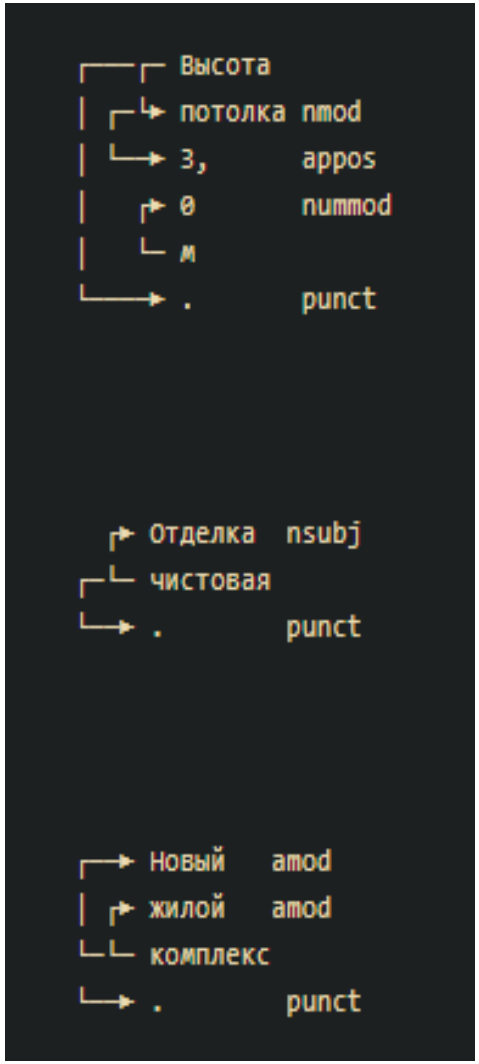
Выделение и подсчет n-gram(n = 2)

Computational Linguistics

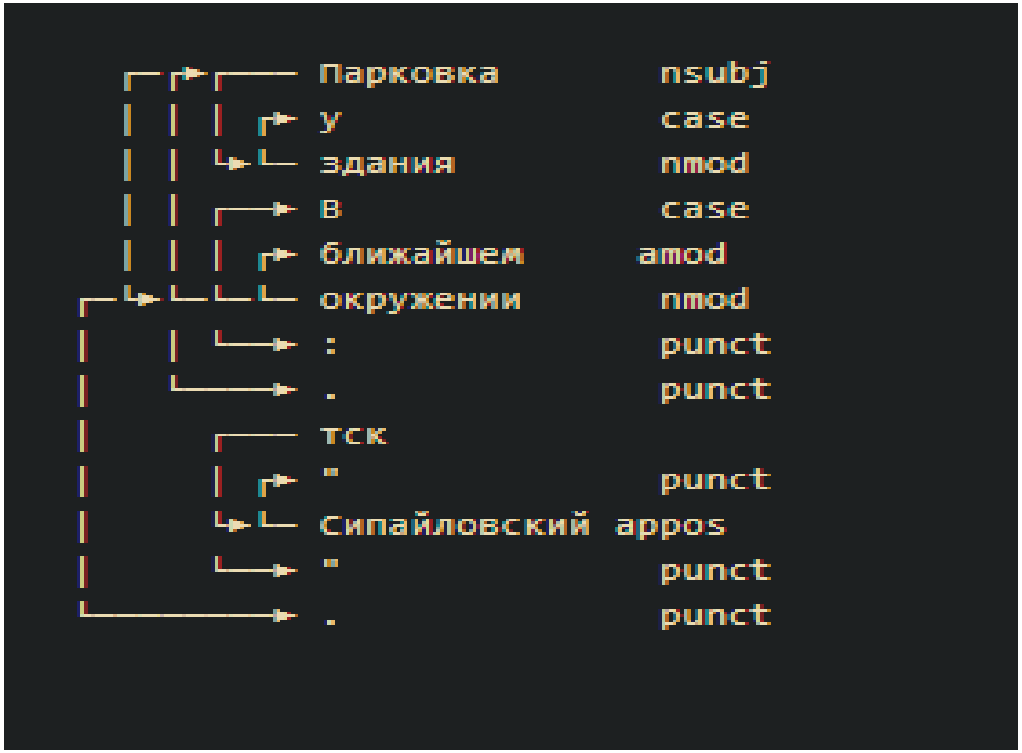


```
(('помещение', 'ремонт'), 24122),
(('хороший', 'ремонт'), 23719),
(('косметический', 'ремонт'), 17116),
(('качественный', 'ремонт'), 16257),
(('чистовой', 'отделка'), 12749),
(('помещение', 'отделка'), 11962),
(('офисный', 'отделка'), 9314),
(('свежий', 'ремонт'), 8985),
```

ВЫДЕЛЕНИЕ НУЖНЫХ СТРУКТУР



С помощью синтаксических связей можно получить все связные слова с опорным словом для дальнейшего подсчета.



ВЫДЕЛЕНИЕ НУЖНЫХ СТРУКТУР