



Поиск

совпадений  
для унификации названий  
спортивных школ



# Подготавливаем эталонные данные



Очистили названия от лишних пробелов до и после текста

Нашли потенциальные дубликаты в регионах

Удалили опечатки в названиях регионов

Нашли потенциальные дубликаты в названиях !



Заменяли school\_id у дубликатов в сырых данных

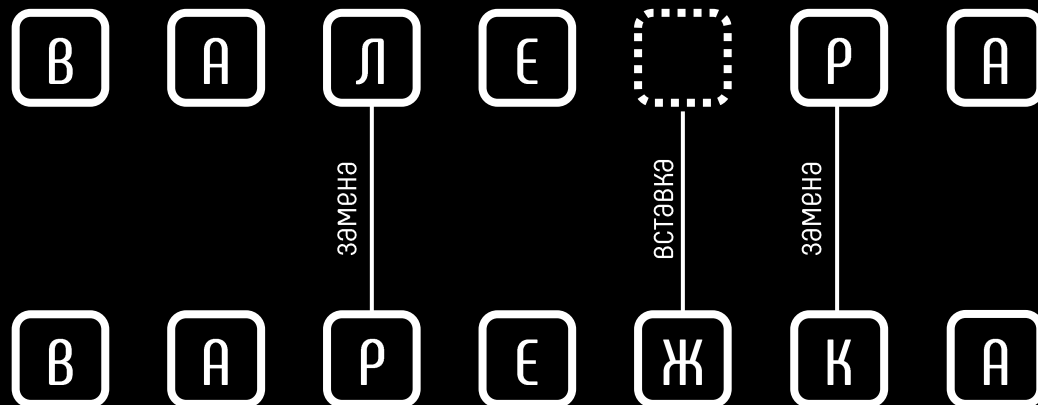
3

дубликата в эталонных данных

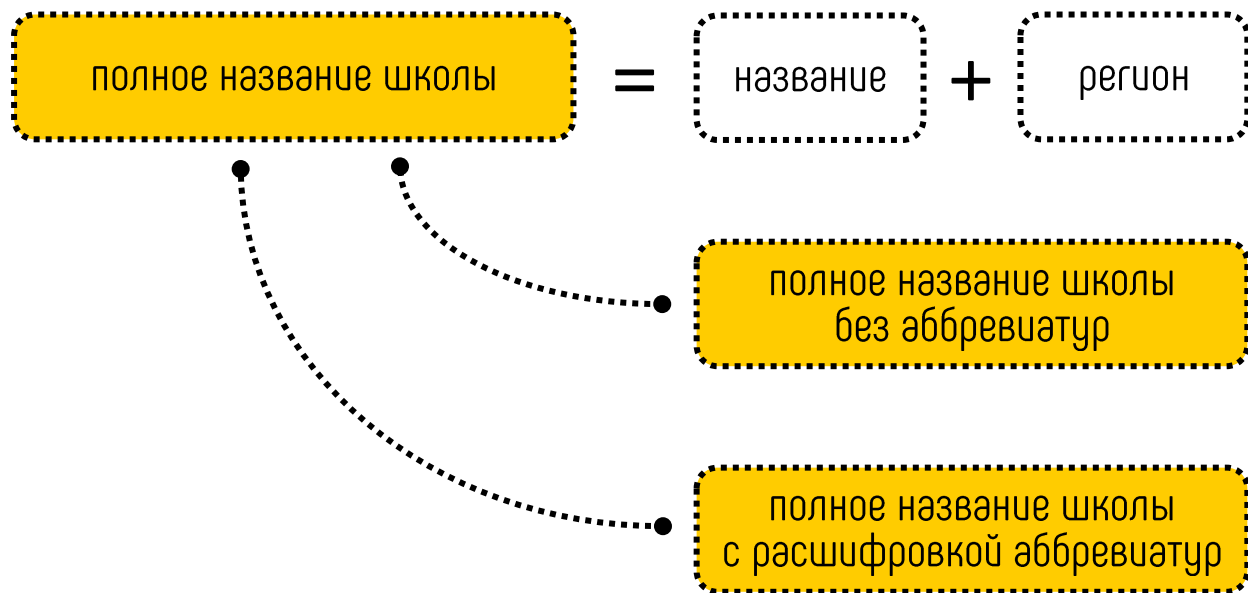
Для поиска дубликатов/опечаток в названиях и регионах использовали расчёт расстояния Левенштейна.

Это расчёт минимального количества корректировок, необходимых, чтобы из одной строки получить другую.

Чем расстояние меньше, тем более похожи строки. Строки с опечатками имеют очень близкие расстояния.



# Создаём новые признаки



Новые признаки нужны, чтобы расширить возможности поиска, ведь мы не знаем заранее, в каком формате пользователь введёт текст

Препроцессинг нужен, чтобы текст был готов к применению модели машинного обучения



Приведение всех слов к нижнему регистру



Очистка текста от спецсимволов (кавычки, запятые)

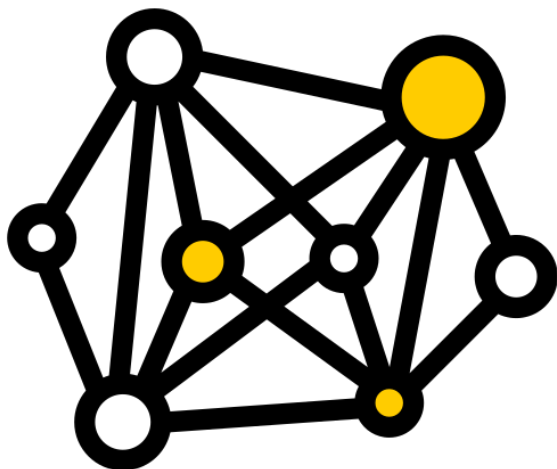


Разбиение текста на токены (составные части)

# Семантический анализ

– это основа нашего решения

Семантика – это дисциплина, которая изучает связь между словами, определяет зависимость значения слова от контекста фразы



Семантическая модель включает:

- слово
- его определение
- сочетания с другими словами
- составление из него фраз и предложений

Для решения нашей задачи мы будем искать связь между признаками, созданными из эталонного датасета, и признаками, созданными из пользовательского ввода

# Компьютер не понимает человеческий язык

Компании как Яндекс, Сбербанк, Google и другие разрабатывают большие языковые модели, обученные на всём пространстве интернета

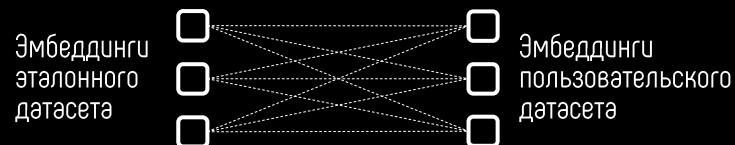
Тарелка фруктов



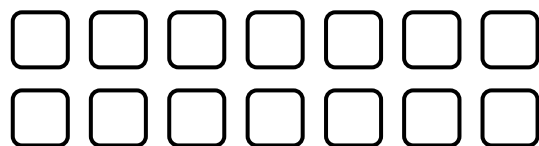
А мы используем эти модели для создания **эмбеддингов** – числового представления слов и предложений

# Ищем совпадения

Перекры́стный  
семантический  
анализ



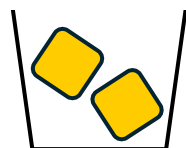
Все результаты  
семантического  
анализа



Отбор лучших результатов  
(параметр `max_results`)

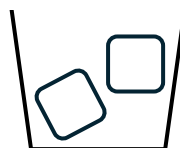


Оценка степени  
сходства



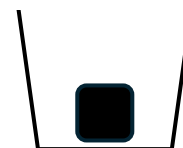
**Высокая**  
score  $\geq 0.8$

Всегда показываем  
эти результаты



**Средняя**  
score  $0.6-0.8$

Показываем, если  
нет ничего лучше




**Низкая**  
score  $< 0.6$

Говорим, что  
ничего не найдено

# Оцениваем качество моделей

**Accuracy** – метрика качества, показывающая процент правильных предсказаний модели

Чем выше параметр `max_results`, тем выше метрика, так как выше шанс, что хоть один кандидат является корректным совпадением

	Максимум 1 кандидат	Максимум 5 кандидатов
Модель LaBSE (Google)	82%	92% 
Модель SBERT (Сбербанк)	67%	86%

# Качество данных на входе определяет качество данных на выходе



## Рекомендации для улучшения предсказаний моделей ML

- Добавить внешний уникальный id – ОГРН/ОГРНИП
- Избавиться от аббревиатур в эталонных данных, особенно если:
  - не является общеизвестной/общепринятой (например, РСЯ)
  - имеет больше одного значения (МО, АО)

## Вариант развития проекта

Создать полноценный пайплайн для хранения модели и данных, обработки поступающих новых эталонных данных и дообучения модели на их основе



Хочешь посмотреть тетрадку  
с исследованием или эту  
презентацию?

Заходи на GitHub!



<https://github.com/A-Yordanova>

Остались вопросы  
или хочешь поделиться  
своим мнением?

Пиши в Телеграм!



[@a\\_yordanova](https://t.me/a_yordanova)

Have a nice day!