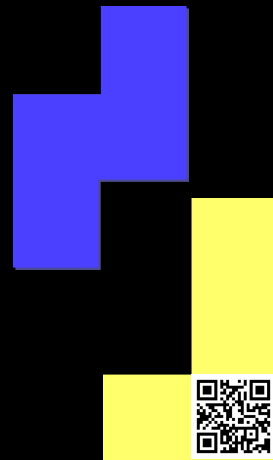




ЦЕНТР НЕПРЕРЫВНОГО ОБРАЗОВАНИЯ
ФАКУЛЬТЕТ КОМПЬЮТЕРНЫХ НАУК

Сопоставление названий товаров из ассортимента аптек

Новицкая Мария, 2025 год



cs.hse.ru/dpo

Введение



ЦЕНТР НЕПРЕРЫВНОГО ОБРАЗОВАНИЯ
ФАКУЛЬТЕТ КОМПЬЮТЕРНЫХ НАУК

Проблема:

- Ручное сопоставление товаров в аптеках приводит к ошибкам и временным затратам (минимум 30% рабочего времени)
- Необходимость автоматизации для повышения эффективности

Цель работы:

Разработка алгоритма для автоматического сопоставления названий товаров.

Актуальность и преимущества автоматизации:

- Сокращение времени на 80-90%
- Уменьшение ошибок.

Постановка задачи и Методы решения

Входные данные:

- Краткий каталог товаров аптеки
- Полный каталог товаров аптеки

Предобработка данных:

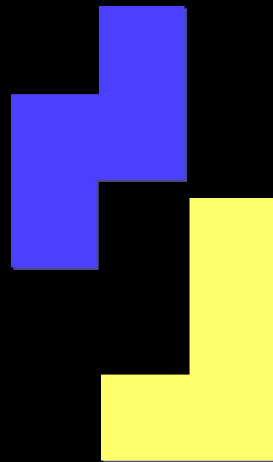
- Очистка текста, токенизация, лемматизация.

Векторизация:

- TF-IDF, CountVectorizer.

Модели машинного обучения:

- Логистическая регрессия,
- Случайный лес,
- CatBoost,
- Нейронные сети

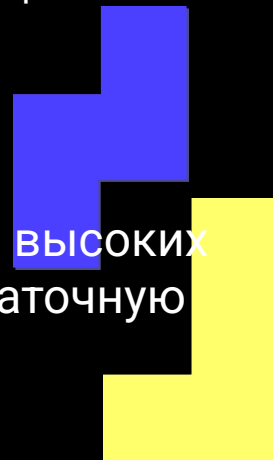


Что мы могли еще использовать и почему не использовали:

Альтернативы векторизации: BERT vs. TF-IDF

- **Вычислительная сложность:**
 - BERT требует GPU и значительного времени для обработки даже небольших датасетов
 - TF-IDF обучился за минуты на CPU.
- **Объем данных:**
 - BERT эффективен на больших текстах (например, статьи), а названия товаров — это короткие строки (3-5 слов)
 - TF-IDF в нашем случае отлично выявляет ключевые слова в названиях

BERT не использован из-за избыточности для коротких текстов и высоких требований к ресурсам. TF-IDF + классические ML показали достаточную эффективность.





Обоснование выбора метрики (Accuracy)

- Простота интерпретации - Accuracy показывает долю верно классифицированных товаров, что легко понять.
- Сбалансированные данные - После предобработки распределение классов стало близким к равномерному, поэтому Accuracy корректно отражает качество модели.
- Сравнение с другими метриками - Precision и Recall важны для задач с дисбалансом, но здесь их F1-score (среднее гармоническое) также близок к Accuracy (разница $\leq 2\%$)

Предобработка данных



ЦЕНТР НЕПРЕРЫВНОГО ОБРАЗОВАНИЯ
ФАКУЛЬТЕТ КОМПЬЮТЕРНЫХ НАУК

Этапы:

- Удаление спецсимволов и цифр.
- Токенизация и приведение к нижнему регистру.
- Лемматизация (например, "таблетки" → "таблетка").

Итог - Улучшение качества данных для анализа

Векторизация текста и Обучение моделей

Методы:

- TF-IDF: Учитывает важность слов в документе.
- CountVectorizer: Подсчёт частоты слов.

Получаем очищенный текст, подходящий для моделей обучения

Логистическая регрессия:

- Accuracy: 82.9%, время: 1 мин.

Случайный лес:

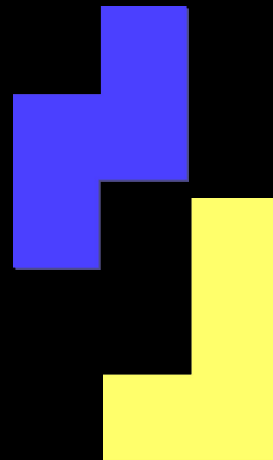
- Accuracy: 85.1%, время: 15 мин.

CatBoost:

- Accuracy: 87.3%, время: 45 мин.

Нейронная сеть:

- Accuracy: 88.1%, время: более 2 ч.



Нейронная сеть.

Архитектура



ЦЕНТР НЕПРЕРЫВНОГО ОБРАЗОВАНИЯ
ФАКУЛЬТЕТ КОМПЬЮТЕРНЫХ НАУК

Структура:

- Тип сети: Полносвязная (FNN — Feedforward Neural Network).
- Слои:
 - Входной: 5000 нейронов (по числу TF-IDF фичей)
 - Скрытые:
 - $8192 \rightarrow 4096 \rightarrow 2048 \rightarrow 1024 \rightarrow 512$ нейронов
 - После каждого слоя — активация ReLU (для нелинейности)
 - Выходной: Число нейронов = количеству категорий товаров.

Почему так много слоев?

Глубокая сеть лучше улавливает сложные зависимости в текстах (например, связь "нимесил" \rightarrow "нимесулид").

Обучение и результаты



Процесс обучения:

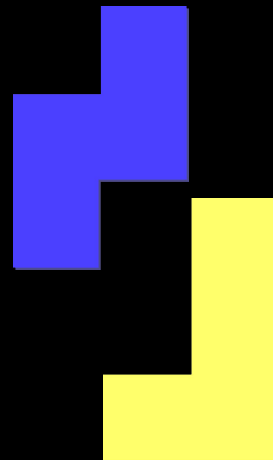
- Данные:
Векторы TF-IDF → преобразованы в тензоры PyTorch и загружаются через DataLoader
- Настройки:
 - Функция потерь: Cross-Entropy (для классификации, Считает loss)
 - Оптимизатор: Adam (скорость обучения = 0.01), Обновляет веса через Adam.
 - Батчи: По 1024 примера (для экономии памяти)

Результаты:

- Ассигасу: 88.1% (лучший показатель среди всех моделей)
- Время обучения: минимум 2 часа

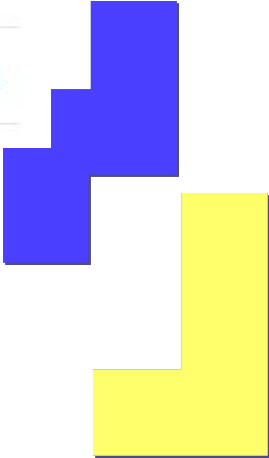
Плюсы и минусы:

-  Точность: На 0.8% лучше CatBoost
-  Ресурсы: Требуется GPU и много времени



Сравнение моделей

Метод	Ассурасу	Время обучения	Примечания
Логистическая регрессия	82.9%	1 мин	Быстро, но менее точно
Случайный лес	85.1%	15 мин	Хорош для небольших данных
CatBoost	87.3%	45 мин	Лучший баланс точности и времени
Нейронная сеть	88.1%	2 ч	Максимальная точность, но ресурсоёмка



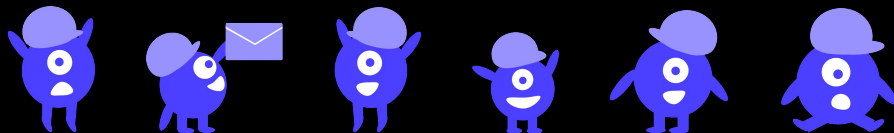
Примеры предсказаний

Успешные случаи:

- "Нимесил гранулы 100 мг" → "НИМЕСУЛИД ГРАН 100 МГ"
- "пошел в магаизн за огурцами, а купил капусту" → "unknown"

Ошибки:

- Редкие названия (например, "Эмоцивит капс 440 мг" → "unknown").
- Опечатки (например, "Парацеатол" → не распознан).



Заключение

Итоги:

- CatBoost показал наилучший баланс точности (87.3%) и времени обучения.
- Нейронные сети — лидеры по точности (88.1%), но требуют слишком больших ресурсов.

Рекомендации:

- Для production: Logistic Regression или CatBoost.
- Для исследований: нейронные сети.

