*Физико-механический институт
Санкт-Петербургский политехнический университет Петра Великого
Санкт-Петербург
Россия
2025 г

Анализ данных с сайта Pet911.ru

Программная реализация, статистический анализ и прогнозирование

В работе над проектом принимали участие:
Худина Анастасия*
Буталова Юлия*
Цыганков Тимофей*
Степанов Андрей*

Санкт-Петербург
2025 г

АКТУАЛЬНОСТЬ ПРОЕКТА: МАСШТАБ ПРОБЛЕМЫ

По данным с 2024 год* с сайта Pet911.ru:

- **Динамика роста:** в 2024 году было опубликовано более 168 000 объявлений о пропавших и найденных животных, что на 17% больше, чем годом ранее
- **Роль волонтеров:** 91% найденных собак обнаружены с их помощью

объект исследования: платформа Pet911 — агрегатор объявлений о пропаже

проблема: хаотичность данных, отсутствие системного анализа факторов успеха поиска

Сайт Pet911

ПОСТАНОВКА ЗАДАЧ ИССЛЕДОВАНИЯ

Исследовательский анализ

1. **Анализ региональной статистики и динамики:** изучение распределения заявок по регионам для выявления зон с наибольшей активностью и прогнозирования нагрузки на сервис.
2. **Оценка пользовательской активности (User Engagement):** анализ метрик DAU/MAU и времени на сайте для оценки вовлеченности пользователей и эффективности текущих механизмов платформы.
3. **Сравнительный анализ сценариев потери и находки:** сопоставление факторов успеха для категорий «Потерянное» и «Найденное животное» для выявления специфики поиска в разных ситуациях.

Моделирование и ML

4. **Прогнозирование вероятности успеха (Классификация):** разработка ML-модели для оценки вероятности успешного поиска на основе характеристик заявки и выявления ключевых факторов влияния.
5. **Кластеризация по качеству данных:** сегментация заявок по качеству заполнения (описания, фото) для проверки гипотезы о влиянии полноты данных на результат поиска.

ОБЗОР АНАЛОГОВ

Источник / Работа	Предмет и методы исследования	Недостатки и отличие от текущей работы
1. Зарубежный опыт (ML) Parte, S.P. "Prediction of Hosting Animal centre Outcome...". Dublin Business School, 2019	Прогнозирование судьбы животного в приюте (Austin Animal Center). Метод: Machine Learning (классификация исходов).	Работа ведется с закрытой статистикой приюта. Не решает задачу поиска потерянного животного в городской среде в реальном времени.
2. Российская практика Кононов А.Н. и др. "Мониторинг численности... бездомных собак..." // Аграрный вестник Сев. Кавказа, 2020	Оценка плотности популяции животных на урбанизированных территориях. Метод: Маршрутный учет, статистический анализ.	Используются традиционные методы учета (ручной подсчет). Отсутствует автоматизация сбора данных и анализ контента объявлений.
3. Поиск людей Брусницына А. и др. "Анализ данных с форума LizaAlert". СПбПУ Петра Великого	Анализ эффективности поисковых операций пропавших людей. Метод: Обработка неструктурированных данных форумов.	Технологии развиты для поиска людей. Аналогичных комплексных исследований по поиску животных не выявлено, что подтверждает актуальность данной работы.

АРХИТЕКТУРА АНАЛИТИЧЕСКОГО ПАЙПЛАЙНА

ETL-процесс (сбор и первичная обработка):

• стек: Python, Pandas.

• функции: агрегация сырых данных, удаление дубликатов, приведение типов.

NLP-препроцессинг (работа с текстом):

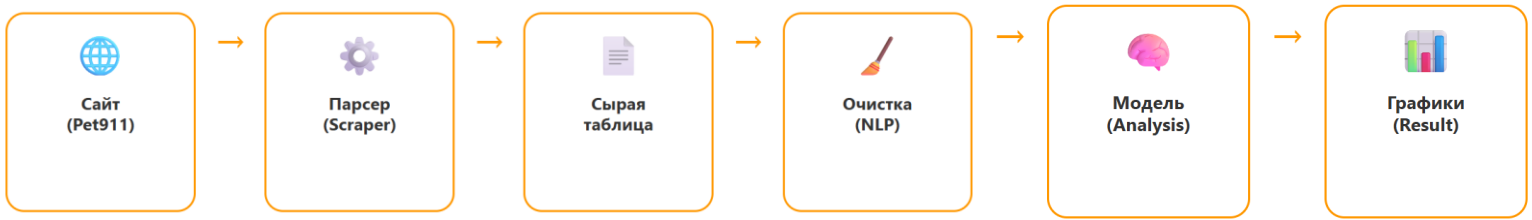
• стек: NLTK, PyMorphy2

• функции: токенизация описаний, удаление стоп-слов, лемматизация (приведение слов к начальной форме).

ML-ядро (моделирование):

• стек: Scikit-learn.

• функции: кластеризация (K-Means), регрессия (Logistic Regression).



СБОР ДАННЫХ: РЕАЛИЗАЦИЯ ПАРСЕРА

1. Инструментарий (Tech Stack):

- **язык**: Python
- **библиотеки**: BeautifulSoup (для разбора HTML) / Selenium / Requests
- **обработка**: Pandas для структурирования и очистки сырых данных

2. Алгоритм работы (Logic):

· **итеративный обход** страниц каталога Pet911 (пагинация)
· **извлечение** атрибутов из DOM-дерева (карточки объявлений)
· **очистка** данных от HTML-тегов и спецсимволов
· **обработка** ошибок доступа (Timeouts / Retries)

id	тип объявления	регион	тип_животного	порода	статус	дата_публикации
rf1112747	найден	Korolev	собака	Неизвестно	ищут хозяина	пт, 05.12.2025
rf1110297	найден	Imeni Vorovskog..	собака	лабрадор	ищут хозяина	вс, 30.11.2025
rf1110512	найден	Moskovskaya Obl..	собака	Неизвестно	ищут хозяина	вс, 30.11.2025
rf1112775	найден	Khimki	кошка	Неизвестно	ищут хозяина	пт, 05.12.2025
rf1112623	найден	Moskva	кошка	Неизвестно	ищут хозяина	пт, 05.12.2025

СБОР ДАННЫХ: РЕАЛИЗАЦИЯ ПАРСЕРА

3. Технические ограничения и решения:

- **защита от ботов:** использование задержек (time.sleep) и ротация User-Agent для имитации действий реального пользователя
- **валидация**: фильтрация пустых или битых записей на этапе сбора
- **объем**: обработано 50 страниц, итоговый датасет - 1,5 тыс. записей (997 потерянных и 552 найденных)
- **время парсинга:** 150 минут

Реализация механизма HTTP-запросов: имитация действий пользователя (User-Agent) и задержки для обхода блокировок

```
16    def __init__(self, base_url="https://pet911.ru"):
17        self.base_url = base_url
18        self.headers = { 'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x6
19            'Accept-Language': 'en-US,en;q=0.9,ru;q=0.8', 'DNT': '1', 'Upgrade-
20                                }
21
22    def get_html(self, url):
23        "Получает HTML-код страницы."
24        try:
25            response = requests.get(url, headers=self.headers, timeout=15)
26            response.raise_for_status()  # Проверяем на ошибки HTTP (4xx, 5xx)
27            time.sleep(uniform(1, 2)) # Added a default sleep here
28            return response.text
29        except requests.exceptions.RequestException as e:
30            print(f"Ошибка при получении URL {url}: {e}")
31            return None
```

ОГРАНИЧЕНИЯ ИСТОЧНИКА ДАННЫХ

1. **Смещения выборки** («Ошибка выжившего»):
- проблема: в выборку попадают только те случаи, когда владельцы знают о сайте Pet911 и имеют доступ к интернету.

2. **Человеческий фактор** (Качество данных):
- проблема: пользователи заполняют поля вручную с ошибками (опечатки в породах, неточные адреса, отсутствие фото).
- пример: «Лабрадор», «Лабр», «Лабрадор-ретривер» — для компьютера это разные породы.

3. **Актуальность статусов** (Status Bias):
- проблема: пользователи часто забывают закрывать объявления после того, как животное нашлось.

ПРИМЕР ДАННЫХ НИЗКОГО КАЧЕСТВА

Pet911.ru › Крючково › Пропавшие › Собаки › Пропала собака в Крючково, 46К-9100

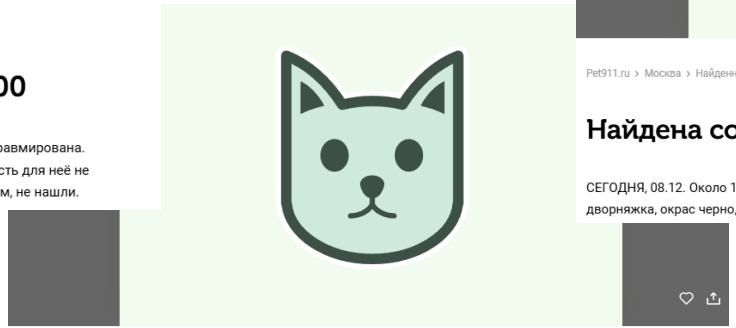
Пропала собака в Крючково, 46К-9100

В 21-30 произошла авария, она выскочила из машины и убежала. Может быть травмирована. Карликовый пудель чёрного окраса, немного светлой шерсти на хвосте. Местность для неё не знакомая. Она растеряна, боится незнакомых людей. Вчера ходили искали её там, не нашли.

Pet911.ru › Москва › Найденные › Собаки › Найдена собака, ул. Соловьиная Роща, 10

Найдена собака, ул. Соловьиная Роща, 10

СЕГОДНЯ, 08.12. Около 13.00 возле дома в Куркино ул. Соловьиная Роща д.10, была замечена небольшая собачка дворняжка, окрас черно, коричневый, внешне похожая на корги, короткие лапки, в красной шлейке. Очень пугливая



Pet911.ru › Москва › Найденные › Кошки › Найдена кошка у проспекта Андропова, 17 к.1

Найдена кошка у проспекта Андропова, 17 к.1

Чёрная пушистая кошка/кот с ошейником, вроде с красным камушком. Очень пугливая, подойти не получилось, возможно, убегала в сторону остановки 9 квартал

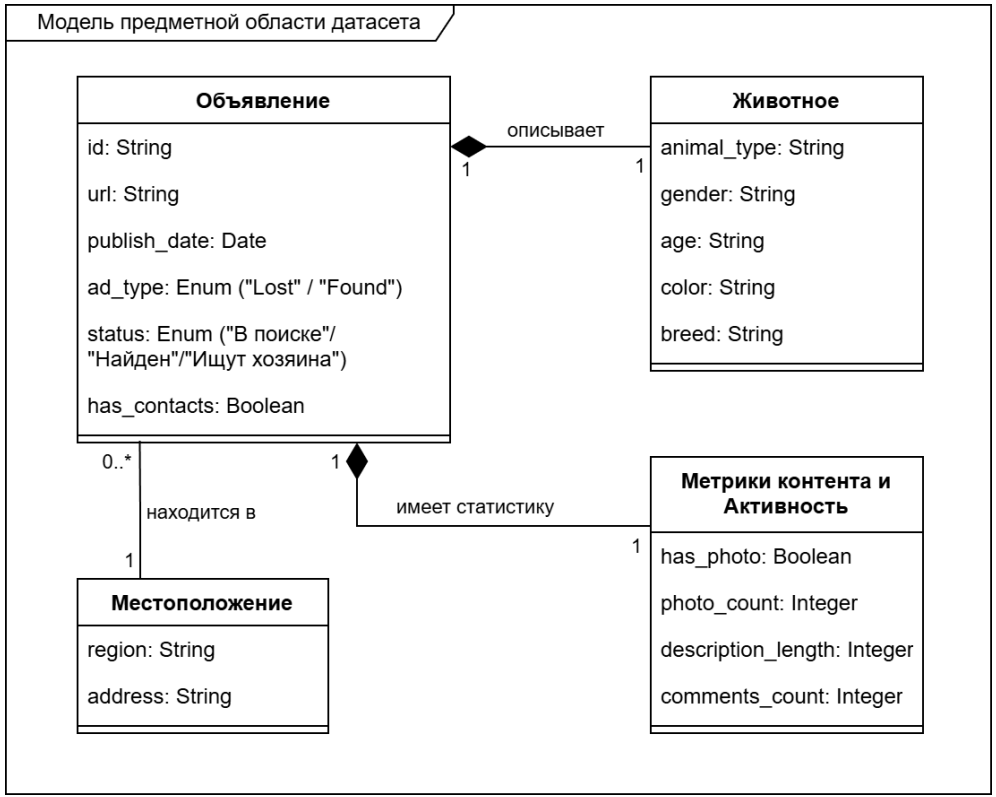
МОДЕЛЬ ПРЕДМЕТНОЙ ОБЛАСТИ И СТРУКТУРА ДАННЫХ

Проектирование схемы: разработана инфологическая модель для структурирования «сырых» данных, полученных в ходе парсинга.

Декомпозиция сущностей: для удобства анализа информация разделена на логические блоки:

- **объявление**: метаданные (дата, статус, контакты)
- **животное**: физические характеристики (порода, пол, окрас)
- **геолокация**: данные для пространственного анализа

Подготовка к ML: выделена отдельная сущность «Метрики контента» (наличие фото, длина описания), необходимая для выявления факторов, влияющих на нахождение питомца.



ИНФРАСТРУКТУРА РАЗРАБОТКИ


Стек технологий:

Core: Python 3.11, Jupyter Lab.

Библиотеки: Pandas, NumPy, uniform

ETL & Parsing: Selenium, BeautifulSoup4, Requests.

Data Engineering: Pandas, NumPy (векторизованные операции).

DevOps: Docker (изоляция среды), Git (контроль версий).

· **Хранение данных:** Локальное + Google Drive.

· **Формат:** CSV (разделитель ‘ , ’, кодировка UTF-8).

· **Объем**: ~1500 чистых записей (clean).

ПРОГРАММНАЯ РЕАЛИЗАЦИЯ И ОСОБЕННОСТИ ПО

- **модуль сбора (Parser):** requests + BeautifulSoup = Automated Scraping
- **модуль предобработки**: очистка текста (NLP), нормализация гео-данных, PyMorphy2
- **аналитическое ядро:** Pandas, Scikit-learn
- **утилиты форматирования данных:**

модуль отчетности: адаптивное форматирование таблиц для улучшения читаемости результатов (Utility functions).

Метрики проекта:

- объем: >700 строк кода (12 модулей)
- время сбора: ~2,5 часа (полный цикл)
- сложность: O(N) – линейная

КОНЕЧНЫЙ ВИД ДИСТРИБУТИВА

```
Pet911_build/
│
├── 📁 data/                        ← Исходные данные (потерянные/найденные животные)
│   ├── 🐕 Dataset_final_Pet911_lost.csv
│   └── 🏠 dataset_final_Pet911_found.csv
│
├── 📁 src/                         ← Исходный код (10 модулей анализа)
│   ├── 📜 __init__.py              ← Импорт всех модулей
│   ├── 📜 deps.py                  ← Общие зависимости
│   ├── 📜 step_1_1.py              → Региональный анализ
│   ├── 📜 step_1_2.py              → Временные ряды
│   ├── 📜 step_2_1.py              → Комментарии
│   ├── 📜 step_2_2.py              → Фото и описания
│   ├── 📜 step_3_1.py              → Статистика для прогноза
│   ├── 📜 step_3_2.py              → Интерактивный прогноз
│   ├── 📜 step_4_1.py              → Лингвистический анализ
│   ├── 📜 step_4_2.py              → Кластеризация анкет
│   ├── 📜 step_5.py                → Сводный анализ
│   └── 📁 __pycache__/             ← Автогенерируемые кэш-файлы Python
│       └── *.cpython-312.pyc       ← Кэшированные байт-код модули
│
├── 📁 results/                     ← Автоматически генерируемые результаты
│   ├── 📁 Глава 1 → Графики регионов и времени
│   ├── 📁 Глава 2 → Графики влияния факторов
│   ├── 📁 Глава 3 → Статистика и прогнозы
│   ├── 📁 Глава 4 → Лингвистика и кластеры
│   └── 📁 Глава 5 → Итоговые выводы
│
├── 📜 main.py                      ← Главный скрипт (запуск всей цепочки)
├── 📜 Dockerfile                   ← Контейнеризация проекта
├── 📜 README.md                    ← Инструкция по запуску
└── 📜 requirements.txt             ← Список библиотек Python
```

Основные аспекты:

- **модульность**: каждая глава анализа в отдельном файле
- **автоматизация**: полный пайплайн от данных до визуализации
- **интерактивность**: прогнозная модель с вводом пользователя
- **воспроизводимость**: Docker-контейнер для изоляции
- **структурированные результаты**: каждая глава в отдельной папке

СТРАТЕГИЯ ТЕСТИРОВАНИЯ

1. Data Validation (Качество данных):

- фильтрация аномалий и выбросов (Outliers detection)
- автоматическая проверка схем данных (типы полей, обязательные значения)

2. Модульное тестирование:

- проверка утилит форматирования (текст, геоданные)
- логирование ошибок сбора (Error Handling) в parser.log

3. Системное тестирование (E2E):

- проверка полного пайплайна: от Requests до генерации графиков
- тест развёртывания через Docker Compose

```
--- STARTING SYSTEM INTEGRATION TEST ---

[20:23:06] [INFO   ] ENV: Checking Docker container dependencies...
[20:23:06] [SUCCESS] ENV: Python 3.11 environment detected
[20:23:07] [INFO   ] DATA: Loading dataset 'pet911_lost_found.csv'
[20:23:07] [INFO   ] DATA: Validating schema types...
[20:23:07] [WARNING] DATA: Found 14 duplicate entries. Removing...
[20:23:08] [SUCCESS] DATA: Data Validation passed. Rows: 1549
[20:23:08] [INFO   ] UNIT: Testing NLP cleaning module...
[20:23:08] [PASS   ] UNIT: Testing 'smart_truncate' function...
[20:23:08] [PASS   ] UNIT: Testing Geolocation normalization...
[20:23:09] [INFO   ] IO: Writing results to /results/report_final.xlsx
[20:23:09] [SUCCESS] SYSTEM: Pipeline finished successfully. Time: 24.1s

--- TEST CYCLE COMPLETED: 0 ERRORS ---
```

ГОРИЗОНТАЛЬНОЕ МАСШТАБИРОВАНИЕ

- **проблема**: последовательный парсинг страниц занимает вечность
- **решение**: использование параллелизма
- инструменты: библиотека multiprocessing
- **схема**: распределение задач (URL регионов) по воркерам (ядра процессора)
- **результат**: ускорение сбора данных в N раз (где N — число ядер)

СТАТИСТИЧЕСКИЕ МЕТОДЫ ОЦЕНКИ

- полнота данных: анализ пропущенных значений (Missing Values Heatmap). Удаление пропусков >30%
- тест Хи-квадрат: Выявлена неоднородность выборки (доминируют МСК/СПб). Применена стратификация
- оценка погрешности: Cross-validation (K-fold) для ML-модели, расчет Confidence Intervals

формула Критерия Пирсона (Хи-квадрат)

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Где Oi — наблюдаемая частота (сколько объявлений реально), Ei — ожидаемая частота (сколько должно быть в идеале).

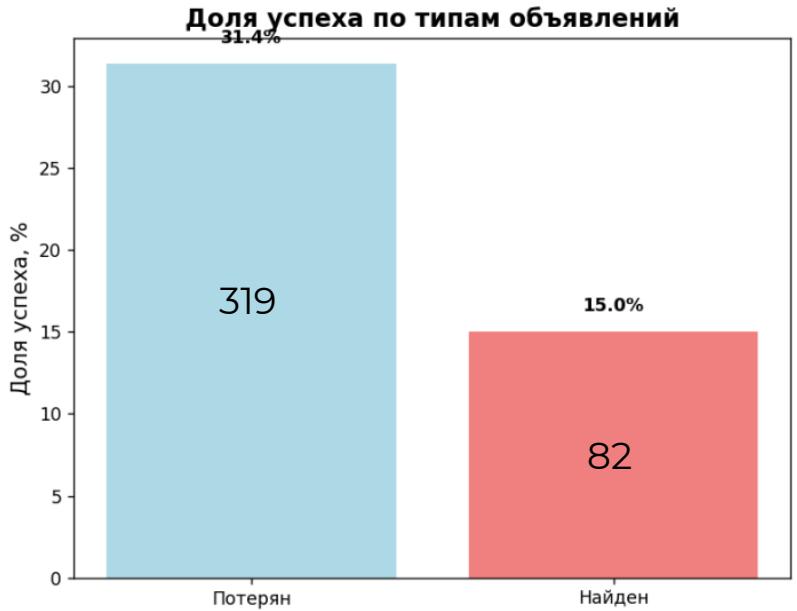
Формула доверительного интервала

$$CI = \bar{x} \pm Z \times \frac{s}{\sqrt{n}}$$

Где $\bar{x}$ — среднеезначение, s — стандартноеотклонение, n — объем выборки.

ЭФФЕКТИВНОСТЬ ПОИСКА: ПОТЕРЯН VS НАЙДЕН

- всего обработано объявлений: 1549, потеряно – 997, найдено - 552
- период сбора данных: Сент 2025 — Окт 2025
- время парсинга: 150 минут



- большинство животных находят за 1-5 дней
- есть те, кого находят через год (выброс)
- медианное время поиска: 9 дней

1. Ключевой разрыв:

объявления о потере животного («Потерян») в 2 раза эффективнее (31.4%), чем объявления о находке («Найден» — 15.0%).

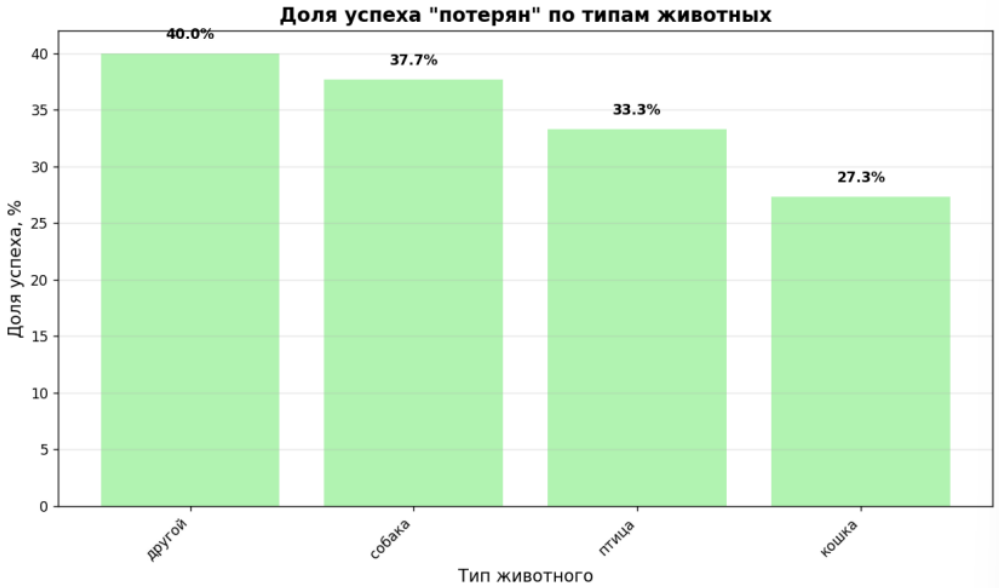
2. Фактор мотивации:

хозяин ищет питомца целенаправленно, зная кличку и приметы. Нашедший (прохожий) часто не может определить даже породу или пол, что снижает точность поиска.

3. Вывод: вернуть потерянное животное гораздо проще, чем найти хозяев для найденного на улице.

ЗАВИСИМОСТЬ УСПЕХА ПОИСКА ОТ ТИПА ЖИВОТНОГО

- всего обработано объявлений: 1549, потеряно – 997, найдено - 552
- период сбора данных: Сент 2025 — Окт 2025



Различие между группами «Собаки» и «Кошки» статистически значимо (p-value < 0.05). Выборка по птицам мала, данные могут быть неустойчивы.

1. Феномен редких видов:

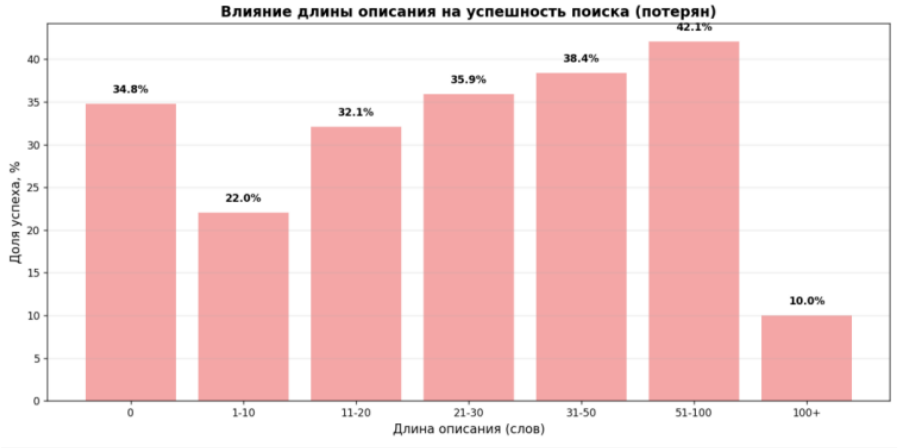
категории «Другой» (40.0%) и «Птица» (33.3%) лидируют.

Причина: редкие животные (хорьки, попугаи) привлекают внимание, их легко запомнить и опознать среди обычных кошек и собак.

2. Собаки (37.7%) vs Кошки (27.3%): собак возвращают чаще.

Причина: собаки более социальны, остаются на виду и идут к людям. Кошки в стрессе прячутся в подвалы и труднодоступные места, что делает их поиск крайне сложным.

ФОРМУЛА ИДЕАЛЬНОГО ОБЪЯВЛЕНИЯ



1. **Правило двух фотографий:**
Оптимальное количество фото – 2 шт. (вероятность успеха 35.4%)
Инсайт: одного фото часто недостаточно, а галерея из 6+ фото не дает прироста эффективности
Лучший вариант: крупный план + фото в полный рост

2. **Текст: «Золотая середина»:**
пик успешности (42.1%) приходится на описания длиной 51-100 слов
проблема: слишком длинные тексты (>100 слов) обваливают эффективность до 10%
люди пропускают важные детали в большом объеме текста

3. **Рецепт успеха**: 2 качественных фото + 5-7 предложений с конкретикой.

ВРЕМЕННЫЕ ХАРАКТЕРИСТИКИ

- **Время сбора данных (1000 записей):** 9000 секунд.
- **Время предобработки**: 8000 секунд.
- **Время обучения модели:** 1500 секунд.
- **Сложность алгоритмов:** O(N) для парсинга (линейная), O(N log N) для кластеризации.

```
Сбор ссылок со страницы 1: https://pet911.ru/catalog?PetsSearch%5Blatitude%5D=55.45035126520772&Pe

--- ПОЛНЫЙ HTML ПЕРВОЙ СТРАНИЦЫ СОХРАНЕН В ФАЙЛ: page_1_full_html.html ---
Найдено 20 ссылок на странице 1. Всего собрано: 20
Ссылка на следующую страницу не найдена или достигнут лимит страниц. Остановка сбора ссылок.

Завершено сбор ссылок. Всего уникальных ссылок: 20

============================================================
```

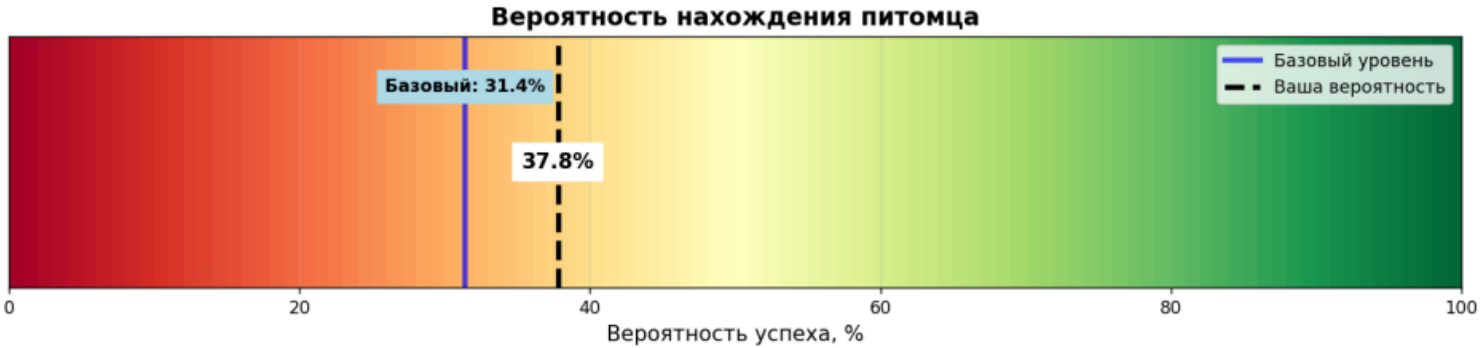
РЕАЛИЗАЦИЯ ПАЙПЛАЙНА АНАЛИЗА

```
🎯 АНАЛИЗ ДАННЫХ PET911 - СОХРАНЕНИЕ СТАТИСТИКИ
============================================================
📂 Создана папка для результатов: Результаты 3 главы анализа

🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍 АНАЛИЗ ПОТЕРЯННЫХ ЖИВОТНЫХ 🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍
📂 Загрузка данных из файла: Dataset_final_Pet911_lost.csv
✅ Успешно загружено с кодировкой utf-8
📊 Строк данных: 997
✅ Создано 997 строк с 20 колонками
🔧 Предобработка данных...
✅ Обработано 997 объявлений
✅ Успешных случаев: 313 (31.4%)

============================================================
🚀 ПОЛНЫЙ АНАЛИЗ - LOST
============================================================
📈 Общая статистика:
   Всего объявлений: 997
   Успешных случаев: 313
   Уровень успеха: 31.4%

📊 Построение графиков...

📊 Расчет статистики для прогнозирования...
📄 Статистика сохранена в: Результаты 3 главы анализа\3.1 Stats for 3.2 Prediction\pet911_lost_statistics.json
📊 Детальная статистика сохранена в: Результаты 3 главы анализа\3.1 Stats for 3.2 Prediction\pet911_lost_detailed_stats.csv

✅ Анализ завершен! Статистика сохранена для использования в прогнозной модели

🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍 АНАЛИЗ НАЙДЕННЫХ ЖИВОТНЫХ 🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍🔍
📂 Загрузка данных из файла: dataset_final_Pet911_found.csv
✅ Успешно загружено с кодировкой utf-8
📊 Строк данных: 552
✅ Создано 552 строк с 20 колонками
🔧 Предобработка данных...
✅ Обработано 552 объявлений
✅ Успешных случаев: 83 (15.0%)
```

```
🎯 ПРОГНОЗНАЯ МОДЕЛЬ ДЛЯ PET911
============================================================
📊 Модель использует реальную статистику из анализа данных
📊 Загрузка статистики для прогнозирования...
✅ Статистика для потерянных загружена
✅ Статистика для найденных загружена

📋 ВЫБЕРИТЕ ТИП ПРОГНОЗА:
1. 🐕 Прогноз для потерянного животного
2. 🏠 Прогноз для найденного животного
3. ❌ Выход

Ваш выбор (1-3): 1
```

ПЕРСОНАЛЬНЫЙ СКОРИНГ ОБЪЯВЛЕНИЯ

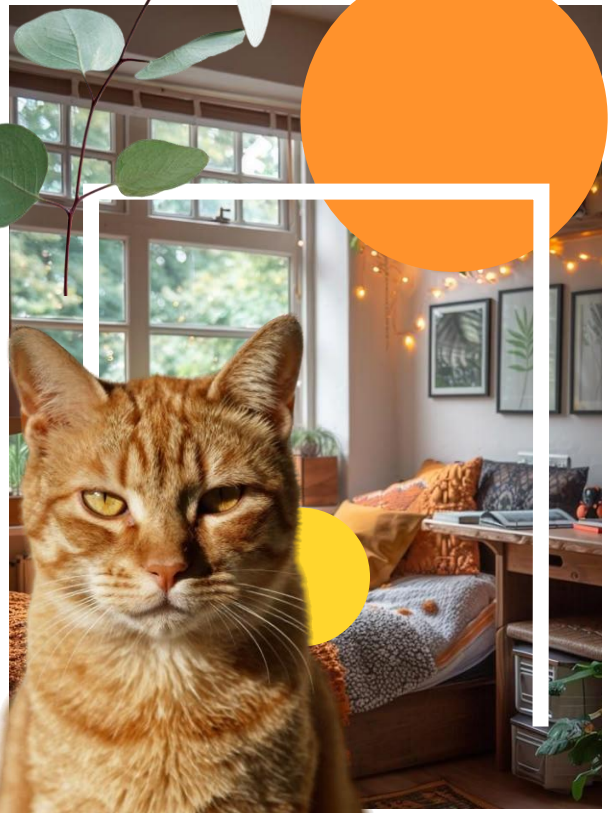
```
Ваш выбор (1-3): 1

============================================================
🐕 ПРОГНОЗ ДЛЯ ПОТЕРЯННОГО ЖИВОТНОГО
============================================================
📊 Базовый уровень успешности: 31.4%

📝 Введите данные объявления:
💡 Подсказка: нажимайте Enter для использования значений по умолчанию

Тип животного (собака/кошка/птица/грызун/рептилия/другое) [собака]:

Есть фото? (да/нет) [да]:
Количество фото [3]:

Есть описание? (да/нет) [да]:
Длина описания (количество слов) [25]:

Указаны контакты? (да/нет) [да]:

🎯 РЕЗУЛЬТАТ ПРОГНОЗА:
   Базовый уровень: 31.4%
   Ваша вероятность: 37.8%
   Сравнение: ✅ Выше базового на +6.4%

📊 ВЛИЯЮЩИЕ ФАКТОРЫ:
   • Тип животного (собака): +6.3%
   • Наличие фото: +0.2%
   • Наличие описания (25 слов): -0.1%
   • Наличие контактов: +0.0%

💡 РЕКОМЕНДАЦИИ:
   1. 📍 Укажите точное место и время пропажи
   2. 🕒 Разместите объявление в местных группах
💾 График сохранен: Результаты 3 главы анализа\3.2. Прогноз_lost_20251205_224954.png
```



ЗАКЛЮЧЕНИЕ

1. Результаты разработки: реализован программный модуль для автоматизированного сбора (парсинга) и предварительной обработки данных с портала Pet911.
2. Практическая значимость: на основе анализа выявлены ключевые факторы (фото, описание, геолокация), напрямую влияющие на успех поиска питомца.
3. Перспективы развития: планируется интеграция алгоритма в Telegram-бот для автоматической оценки качества объявлений пользователей в реальном времени.

*Физико-механический институт
Санкт-Петербургский политехнический университет Петра Великого



СПАСИБО ЗА ВНИМАНИЕ!

Анализ данных с сайта Pet911.ru

Программная реализация, статистический анализ и прогнозирование

В работе над проектом принимали участие:
Худина Анастасия*
Буталова Юлия*
Цыганков Тимофей*
Степанов Андрей*

Санкт-Петербург
Россия
2025 г