Стилометрическая классификация реплик персонажей Stardew Valley

Можно ли различить голоса персонажей,
если все реплики написаны одним автором?

А. А. Аладьина · В. Н. Копнин · М. В. Савостиков
ПИЭ-13МО



Часть 1

Решаемая задача и корпус

Постановка задачи · сбор и очистка данных · характеристики корпуса

Докладчик: А. А. Аладьина



Задача и постановка

Решаемая задача

- Дано: 2005 коротких реплик, 34 персонажа, один автор.
- Требуется: по одной реплике определить персонажа.
- Каждый объект относится ровно к одному из 34 классов.
- Это не авторская атрибуция: мы различаем *речевые маски персонажей*, а не авторов.
- Проверяется, статистически различимы ли эти маски.

Обозначения

$D = \{(x_i, y_i)\}_{i=1}^{n}, \quad n = 2005;$
x_i — реплика;
$y_i \in Y, \ |Y| = 34$ — персонаж;
f — классификатор.

Критерий качества — макро-F1: средняя F-мера по 34 классам.

Корпус: сбор, разметка, очистка

1. Сбор

- Реплики из диалогов игры.
- Единица анализа — одна реплика.
- 34 персонажа, тексты на английском.

2. Разметка

- Метка — имя говорящего персонажа.
- Ручная проверка реплики и метки.
- Метаданные (пол, возраст, вид) хранятся отдельно и не подаются модели.

3. Очистка

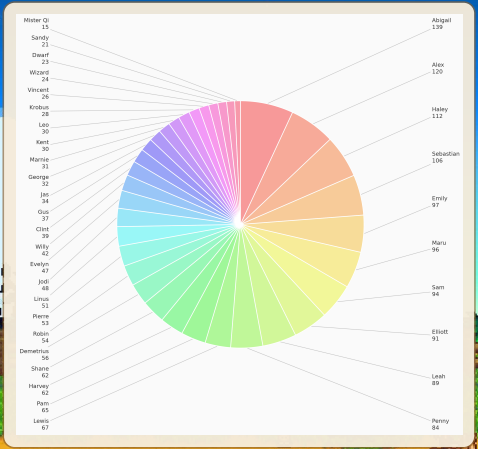
- **Unicode-очистка**: разные варианты кавычек, тире, многоточий и пробелов приводятся к единому виду.
- Плейсхолдеры вида `<PLAYER>` заменены нейтральными словами.
- Пунктуация и регистр сохранены: это стилометрический сигнал.

Ограничения данных: ручная фильтрация; нет идентификаторов сцен.

Игровые персонажи



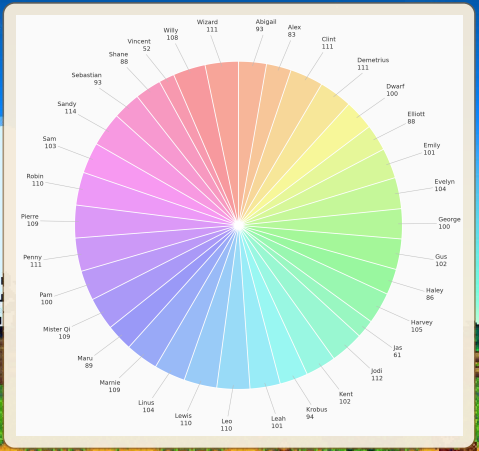
Характеристики корпуса: классы



Реплик	2005
Персонажей	34
Максимум класса	Abigail, 139
Минимум класса	Mister Qi, 15
Точные дубли	0
Почти-дубли	13 пар, но только внутри класса

Дисбаланс почти десятикратный: от 15 до 139 реплик. Поэтому доля правильных ответов обманчива; основная метрика — макро усреднённая F-мера по классам.

Характеристики корпуса: длина реплик



Средняя длина	19,1 слова
Медиана	17 слов
Диапазон	3–78 слов

- Короткие реплики дают мало информации для стилометрических характеристик.
- Длина сама по себе — сильный, но опасный признак: она может подменять «стиль» просто узнаванием персонажа.
- Поэтому её вклад проверялся отдельно: с исключением признаков длины и по группам длины.

Часть 2

Методология: как измеряем стиль

Признаки · LSA · классификаторы · схема проверки · метрики

Докладчик: В. Н. Копнин



Термины, которые встретятся дальше

Фолд — одна из частей разбиения выборки при перекрёстной проверке. При 5 фолдах модель обучается на 4 и проверяется на 1, затем разбиение меняется.

POS (part of speech) — часть речи; POS-тег — её обозначение; POS-грамма — последовательность тегов.

Universal POS — универсальный набор тегов: NOUN, VERB, ADJ, ADV, PRON, DET, ADP, CONJ, NUM, PRT, INTJ, PUNCT, X. Единый набор для разных языков.

Клауза — синтаксическая единица с подлежащим и сказуемым; в репликах оценивается приблизительно.

Синтаксис-прокси — приблизительная оценка синтаксиса без полного разбора: число союзов, длина предложения, порядок слов.

n-грамм — последовательность n соседних единиц: символов, слов или POS-тегов.

LSA — сжатие таблицы «документ × признак» в несколько скрытых компонент.

Макро-F1 — средняя F-мера по классам; каждый персонаж весит одинаково.

Unicode-очистка — приведение разных вариантов записи символов к единому виду: кавычки, тире, многоточия, пробелы.

Класс-аттрактор — класс, к которому модель ошибочно относит чужие реплики.

Фолд и группа признаков — разные вещи: фолд — часть проверки, группа — набор признаков.

Общая схема

Реплика (текст персонажа) → очистка (Unicode, плейсхолдеры) → токены и частеречная разметка (NLTK, Universal POS) → признаки (ручные, n-граммы, LSA) → модель (вероятности по 34 классам) → проверка (5 фолдов, повторы, метрики).

Что важно про схему

- Все обучаемые преобразования настраиваются только внутри обучающих фолдах.
- Тестовые фолды не участвуют ни в настройке признаков, ни в выборе параметров модели.

Что такое «вектор признаков»

- Это не «понимание» текста, а числовое описание его статистических свойств: длины, частоты, пунктуация, темы.

454 ручных признака: 11 групп

Группа	Примеры признаков
Длина	число слов, символов, предложений; длина предложения
Лексика	длины слов, апострофы, заглавные буквы, цифры
Словарное богатство	TTR, Guiraud, Yule’s K, hapax; на коротких текстах нестабильно
Символы	частоты букв, регистр, пунктуация
Пунктуация	запятые, многоточия, вопросы, восклицания, повторы
Функциональные слова	артикли, предлоги, союзы, местоимения
Местоимения	1-е, 2-е, 3-е лицо; отношение «я/ты»
Части речи	Universal POS, Penn Treebank, время, модальность
Синтаксис	клаузы, подчинительные союзы, пассивные конструкции
Читаемость	Flesch, FK, SMOG; на коротких репликах приблизительна
Разговорные маркеры	междометия, смех, хеджи, вежливость, сокращения

Автоматические признаки: n-граммы и LSA

n-граммы

- **n-грамм** — последовательность n соседних единиц: символов, слов или POS-тегов.
- Символьные n-граммы 2–5: морфология, пунктуация, орфография.
- n-граммы POS-тегов 1–3: последовательности частей речи, синтаксические шаблоны.
- TF-IDF: вес признака с учётом частоты в реплике и редкости в корпусе.

LSA

- Сжатие таблицы «реплика $\times$ признак» в k скрытых компонент.
- Варианты: `word`, `content`, `fw`, `char`, `pos`.
- $k = 50$–400; для `content`-LSA качество росло до 400.
- LSA — сжатое представление, а не «понимание» текста.

Символьные n-граммы чувствительны к морфологии и пунктуации; n-граммы POS-тегов — к порядку частей речи; LSA сжимает разреженные признаки в несколько чисел.

LSA простыми словами

Идея

- Строим таблицу «реплика $\times$ слово», взвешиваем частоты через TF-IDF.
- Сжимаем таблицу до k скрытых осей: каждая реплика превращается в короткий вектор.
- Оси не обязаны быть «темами»: их порядок и знак произвольны.

Варианты LSA

- `word` — лексика и тема.
- `content` — содержание без служебных слов.
- `fw` — функциональные слова и синтаксис.
- `char` — морфология и пунктуация.
- `pos` — частеречные последовательности.

k подбирался в диапазоне 50–400; для `content`-LSA качество устойчиво росло до 400. Ключевой вопрос: ловит ли LSA по содержательным словам *тему*, а не только стиль?

Классификаторы

Линейные модели

- Логистическая регрессия: «признак × вес», затем вероятности; редкие классы получают повышенные веса.
- Линейный SVM: граница с максимальным зазором; часто силён на тексте, но вероятностей не выдаёт.

Деревья и ансамбли

- Случайный лес и ExtraTrees: сотни деревьев; у ExtraTrees пороги выбираются случайно — меньше разброс.
- Мягкое голосование: усредняем вероятности моделей.
- Объединение через мета-классификатор (стекинг) у нас переобучилось и оказалось хуже.

Сравнивали несколько семейств, чтобы вывод не зависел от одного алгоритма: близкие результаты у разных моделей делают вывод надёжнее.

Как проверяем

Схема оценки

- Стратифицированная перекрёстная проверка: 5 фолдов (частей разбиения); на 4 фолдах обучаемся, на 1 проверяем. Доли классов внутри фолдов сохраняются.
- Повторяем 3–5 раз с разными разбиениями — оцениваем разброс.
- Вложенная схема: внешние фолды — оценка, внутренние — настройка параметров модели.
- Отдельная отложенная выборка 80/20 для финальных кандидатов.

Контроль ошибок исследователя

- TF-IDF, сжатие и масштабирование обучаются только на обучающих фолдах — борьба с утечкой данных.
- Контроль перестановкой меток дал 0,027–0,028 — воспроизведён случайный уровень.
- Почти-дубликаты находятся только внутри одного класса и не пересекают обучение и тест.



Метрики: что и зачем

Метрика	Что показывает
Macro-F1	Средняя F-мера по 34 классам: каждый персонаж важен одинаково, включая редких. Основная метрика.
Доля правильных ответов	Доля совпавших предсказаний; при дисбалансе обманчива: модель, всегда выбирающая частый класс, выглядит неплохо.
Сбалансированная доля	Средняя полнота по классам: находит ли модель редких персонажей.
Тройка лучших (top-3)	Истинный персонаж оказался в тройке наиболее вероятных ответов.
MCC / каппа	Согласие с поправкой на случайное угадывание.

Ориентиры при 34 классах: случайный уровень macro-F1 $\approx$ 0,029, top-3 $\approx$ 0,088.

Гипотезы исследования

H1. Функциональные слова, части речи и символьные n-граммы (последовательности соседних символов) различают персонажей выше случайного уровня.

H2. На коротких репликах части речи и символьные n-граммы устойчивее словарного богатства и читаемости.

H3. LSA по всем словам отражает тему, а LSA по частям речи и функциональным словам — стиль.

H4. Качество зависит от длины реплики.

H5. Ошибки группируются вокруг персонажей со сходным возрастом, полом, ролью или видом.

Как проверяем

- H1 — сравнение моделей и поочерёдное исключение групп.
- H2 — исключение групп и анализ по длине.
- H3 — варианты LSA и контроль содержания.
- H4 — метрики по группам длины.
- H5 — анализ ошибок по метаданным.

Часть 3

Результаты и интерпретация

Оценки качества · вклад признаков · матрицы ошибок · ограничения · выводы

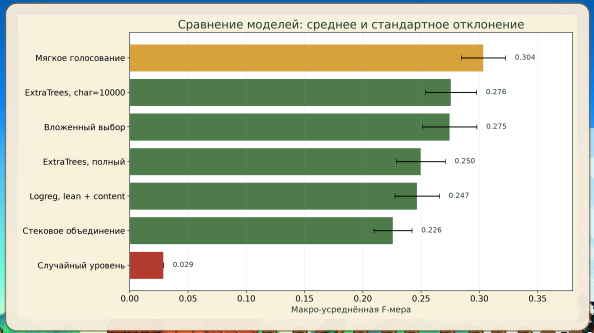
Докладчик: М. В. Савостиков



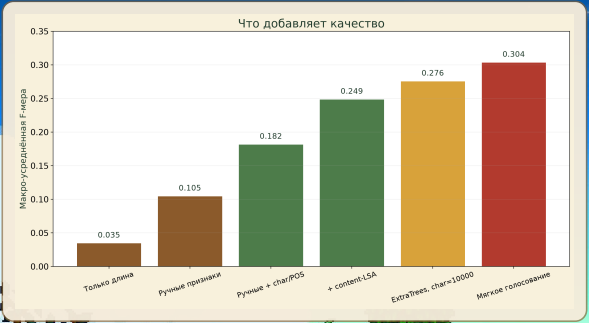
Главный результат

Мягкое голосование (5 фолдов)	0,304
ExtraTrees, char = 10 000	0,276
Вложенный выбор	0,275
Отложенная выборка	0,333
Перестановка меток	0,029

Числа — macro-F1. Случайный уровень для 34 классов — 0,029; контроль перестановкой меток вышел на этот уровень, значит в извлеченных характеристиках есть информация.



Что добавляет качество



- Только длина: 0,035.
- Ручные стилометрические признаки: 0,105.
- + символьные и частеречные n-граммы: 0,182.
- + content-LSA: 0,249.
- ExtraTrees, char = 10 000: 0,276.
- Мягкое голосование: 0,304.

Итого: основной прирост — лексика, n-граммы и content-LSA; «чистый синтаксис» (функциональные слова и POS-LSA) сам по себе слаб — важный отрицательный результат.

Стиль или содержание?



- Проверка: содержательные слова (существительные, глаголы, прилагательные, наречия) заменены нейтральным маркером.
- Исходный текст: 0,248; после маскирования: 0,113.
- Падение более чем вдвое, но остаток 0,113 — в 3,9 раза выше случайного уровня.

Вывод контроля: более половины качества даёт лексическое *содержание*; стилевой сигнал сохраняется, но слабее тематического.

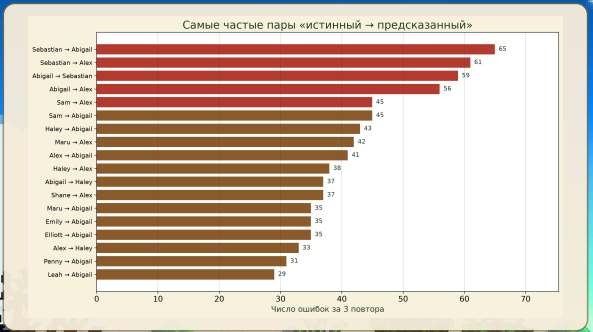
Матрица ошибок: как читать



- **Строки** — истинный персонаж, **столбцы** — предсказание; **диагональ** — верные ответы, вне диагонали — ошибки.
- Полная матрица 34×34 нечитаема: показываем top-12.

Что видно: крупная доля вне диагонали у многих строк — редкие классы «утекают» в популярные.

Кого модель путает чаще всего



Частые пары ошибок (число перепутанных реплик)

- Sebastian → Abigail: 65
- Sebastian → Alex: 61
- Abigail → Sebastian: 59
- Abigail → Alex: 56

Основной клубок — молодые персонажи: Abigail, Alex, Sebastian, Haley, Sam. Один автор и общие сюжетные контексты делают сходство голосов ожидаемым. Это коллапс редких классов в классы-аттракторы.

Классы-аттракторы

Классы-аттракторы

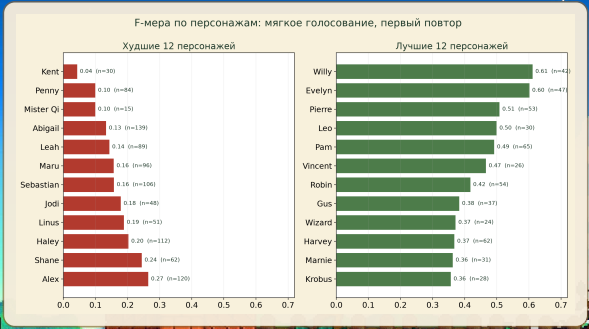
- Alex и Abigail собирают заметную долю *чужих* предсказаний: на разных повторах Abigail получает порядка 10–18 % всех ответов при истинной доле класса ≈ 6,9 %.
- Коррекция априорных вероятностей классов macro-F1 не улучшила и проблему не сняла.



Персонажи: кто различается

Лучшие и худшие по F-мере

- Лучшие: Evelyn, Willy, Pam, Pierre, Leo.
- Худшие: Kent, Wizard, Penny, Mister Qi, Leah, Abigail.
- У редких классов — широкие интервалы оценок.
- Мягкое голосование улучшает часть редких классов (Wizard, Dwarf, Vincent), но Kent остаётся F1 = 0.



Что означают полученные оценки

- 0,30 macro-F1 при 34 классах — примерно **в 10 раз выше случайного уровня** 0,029: устойчивый сигнал различимости.
- Значение умеренное и это ожидаемо: реплики по ∼19 слов, дисбаланс классов и общие темы не позволяют «узнать» персонажа по одной фразе.
- Умеренность объясняется *условиями задачи*, а не тем, что автор «не старался»: превышение случайного уровня как раз говорит о намеренно созданных речевых масках.

- Мы решаем задачу «*какой персонаж* говорит», а не «*он или не он* писал»: автор один.
- И о «чистых» идиолектах говорить нельзя: основной вклад дают лексико-тематический профиль и символьные n-граммы, а после маскирования содержательных слов качество падает.

Ограничения

Данные

- Один корпус, одна игра, один автор; внешняя валидация не проводилась.
- Короткие реплики и сильный дисбаланс классов.
- Ручная фильтрация корпуса невоспроизводима.
- Нет идентификаторов сцен и диалогов.

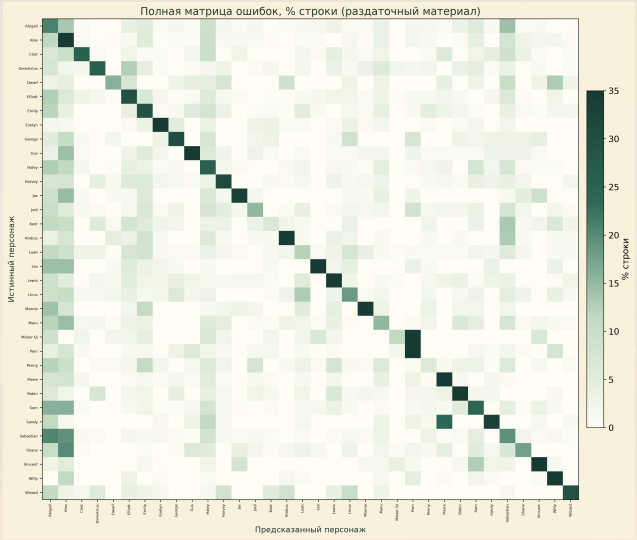
Методология

- Выбор признаков и ансамбля опирался на прошлые серии; вложенная проверка снижает этот риск, но не устраняет.
- Основной вклад — содержательные слова и тема.
- Простая коррекция приоров не помогла редким классам.
- Объединение через мета-классификатор переобучается на малой выборке.

Выводы

- **Персонажи статистически различимы:** лучшая модель — 0,304 macro-F1, консервативная оценка — 0,275; случайный уровень — 0,029.
- **Основной сигнал — лексико-тематический:** символьные n-граммы и content-LSA дают основной прирост; функциональные слова и частеречная LSA сами по себе слабы.
- **После замены содержательных слов** качество падает с 0,248 до 0,113: содержание доминирует, но стилевой сигнал сохраняется.
- **Главная проблема — редкие классы и классы-аттракторы;** мягкое голосование помогает, но не решает её полностью.
- **Корректная формулировка результата:** различимы речевые маски по лексико-тематическому профилю; утверждение о «чистом стиле» требует внешней валидации и тематического контроля.

Приложение: полная матрица ошибок 34×34



Спасибо за внимание

