Анализ и предсказание вовлеченности читателей на основе лога приложения для чтения электронных книг

Тух И. Е.

Научный руководитель: к.т.н. П. И. Браславский

Санкт-Петербургская школа физико-математических и компьютерных наук

НИУ ВШЭ – Санкт-Петербург
2020

Введение в предметную область

Задача предсказания вовлеченности читателя в процесс чтения по текстовому содержанию книги
Применения:

- Рекомендательные системы
- Рекомендации для писателей
- Интерактивное чтение

Анализ поведения читателей:

- Замеры в лабораторных условиях
- Можно использовать данные приложений

Данные Bookmate

Информация о действия пользователей за 10 месяцев:

- 130+ тысяч книг
- 800+ тысяч активных пользователей
- 33+ миллионов записей

Запись в логе (*сессия*):

Поле	Пример значения
_from	60
_to	62.8571
book_id	1639
document_id	40591
item_id	1220853
read_at	1420059070000
size	514
user_id	453753
app_user_agent	"NOKIA/RM-937_apac_hong_kong_222 WP/8.10 BOOKMATE/1.38"

Обзор работ

Анализ чтения веб-новостей:

- Предсказание общего уровня вовлеченности по текстовому содержанию[1]
- Предсказание проматывания статьи назад по текстовым признакам[2]

Нельзя обобщить на чтение художественной литературы

[1] Lagun and Lalmas, “Understanding user attention and engagement in online news reading”, 2016.

[2] Smadja et al., “Understanding Reader Backtracking Behavior in Online News Articles”, 2019.

Обзор работ

Экспериментальный анализ чтения художественной литературы:

- Скорость чтения - сигнал заинтересованности[3]
- Динамичные моменты читаются быстрее, чем спокойные[4]
- Сопоставление скорости чтения и физиологических параметров[5]
- Предсказание скорости чтения по признакам текста и контекста[6]

---

[3] Nell, “The Psychology of Reading for Pleasure: Needs and Gratifications”, 1988.

[4] Cupchik and Laszlo, “The Landscape of Time in Literary Reception: Character Experience and Narrative Action”, July 1994.

[5] Brouwer et al., “Physiological signals distinguish between reading emotional and non-emotional sections in a novel”, 2015.

[6] Tukh, Braslavski, and Buraya, “Log-Based Reading Speed Prediction: A Case Study on War and Peace”, 2019.

Особенности предыдущих работ

Недоступны большие наборы данных:

- Мало исследуемых читателей

Можно исследовать другие *сигналы заинтересованности:*

- Пролистывания страниц
- Повторные чтения фрагментов
- Прерывания в процессе чтения
- Высокая скорость чтения
- Чтение в необычное время
- Остановка чтения книги

Цель и задачи

Цель: научиться предсказывать интерес читателя по текстовому содержанию книги.
Задачи:

- Извлечь сигналы заинтересованности из логов Bookmate
- Проанализировать полученные сигналы заинтересованности
- Построить интегрированный сигнал заинтересованности
- Обучить модели для предсказания интегрированного сигнала

Извлечение сигналов заинтересованности

- «Пятьдесят оттенков серого» и «На пятьдесят оттенков темнее» за авторством Э.Л.Джеймс
- 447 и 501 пользователей (прочитали $\geqslant$ 50% книги)
- Произведения разбиты на 380 и 390 фрагментов
- Для каждой пары фрагмент и пользователь определяется, проявился ли сигнал заинтересованности

Извлечение сигналов заинтересованности

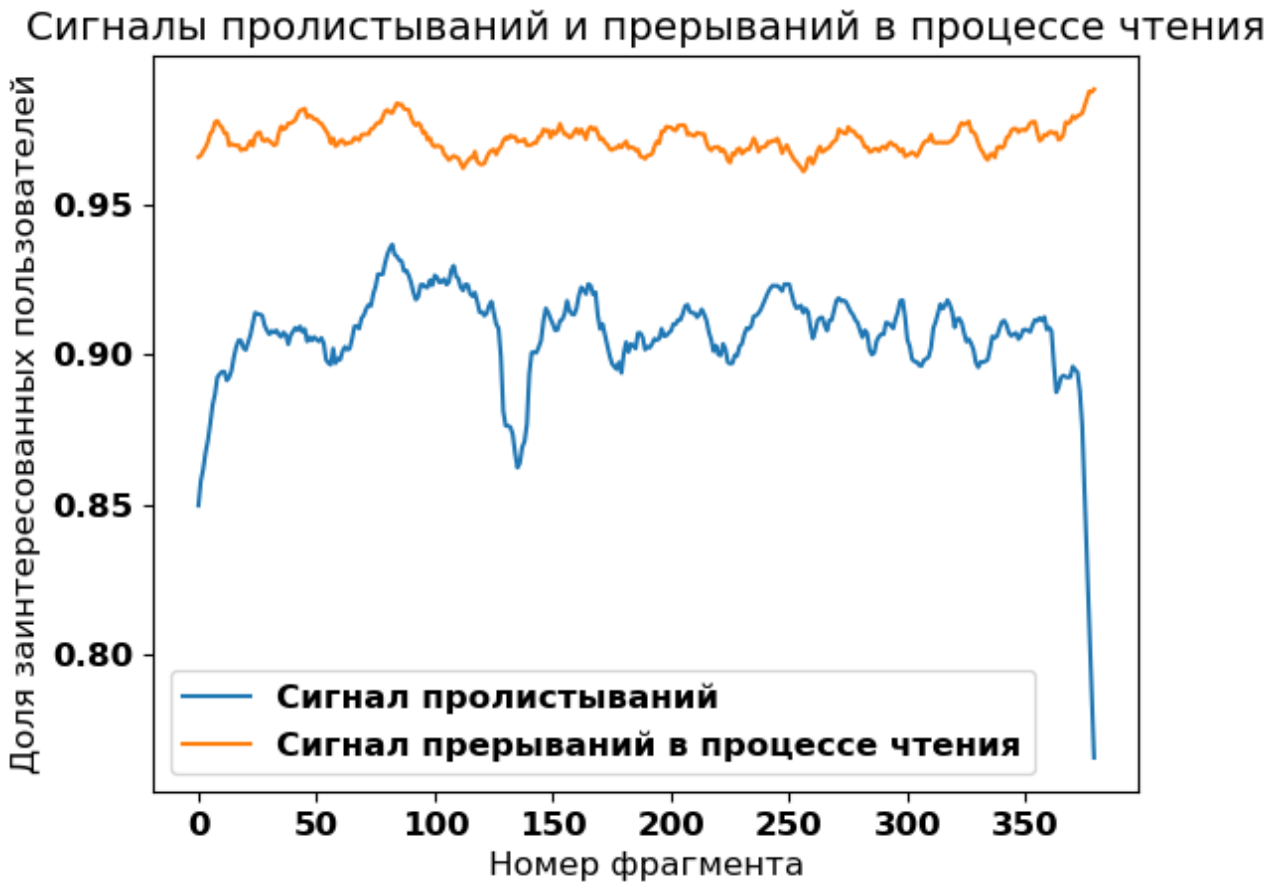
Пример: сигнал высокой скорости чтения

- Для каждого пользователя рассматриваются сессии в хронологическом порядке
- Рассматриваются соседние сессии, вычисляется скорость
- Отфильтровываются слишком близкие во времени сессии (сигнал пролистывания)
- Скорость чтения сравнивается с пороговым значением
- Пороговое значение подбирается с точки зрения вариативности сигнала
- Для каждого фрагмента считается доля пользователей, которым он интересен

Анализ сигналов заинтересованности

Вывод: сигналы интерпретируются, при сопоставлении их экстремумов с сюжетом.

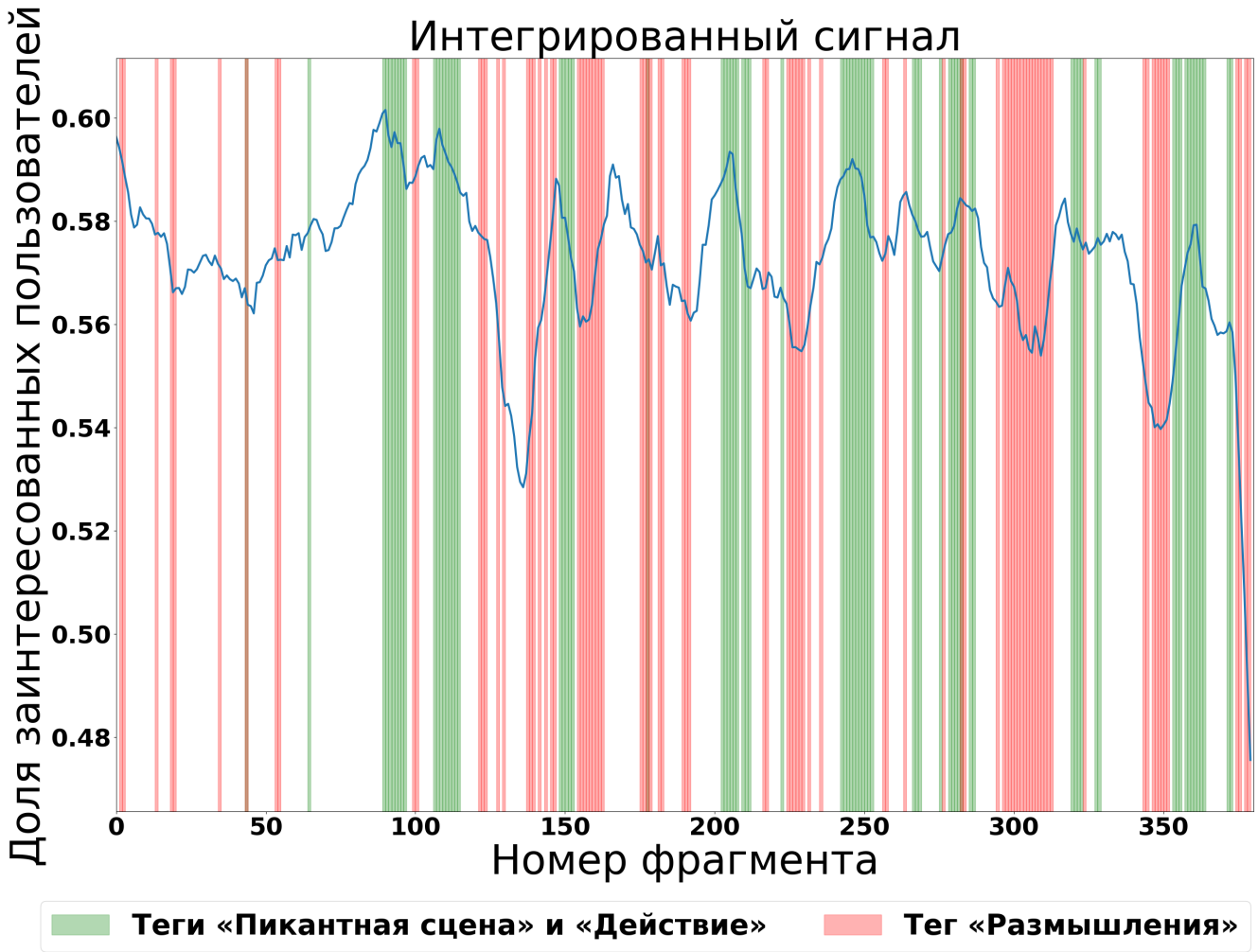
Анализ сигналов заинтересованности



Вывод: сигналы заинтересованности дополняют друг-друга. Высокие значения у нескольких сигналов усиливают уверенность в «интересности».

Интегрированный сигнал заинтересованности

Интегрированный сигнал – средняя по всем сигналам доля заинтересованных



Предсказание сигналов заинтересованности

Основные признаки:

- Удобочитаемость:
  - Средняя глубина в дереве зависимостей
  - Доля глаголов / существительных / прилагательных / личных местоимений
  - Средняя длина слова / предложения
  - Доля имен персонажей / главных персонажей
- Тональность / тема:
  - Тональность по словарю
  - Доля эмоционально окрашенных глаголов
  - Доля упоминаний частей тела

Векторные представления на основе BERT

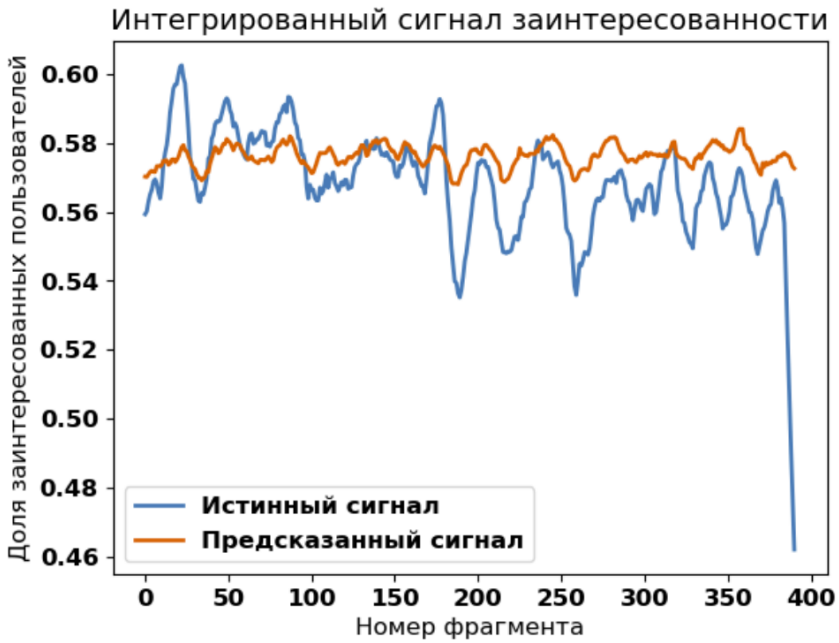
Предсказание сигналов заинтересованности

Признаки	Разбиение	Лучшая модель	Средняя абсолютная ошибка	% ошибки от длины диапазона значений сигнала	% ошибки от значения сигнала	Средняя квадратичная ошибка
«Пятьдесят оттенков серого»						
BERT	Случайное	Random Forest Regressor	**0,0145**	6%	4%	**0,0003**
BERT	Упорядоченное	Random Forest Regressor	0,0185	8%	5%	0,0007
Основные	Случайное	Linear Regression	0,0156	7%	4%	0,0004
Основные	Упорядоченное	ElasticNetCV	0,0187	8%	5%	0,0007
Объединенные	Случайное	Random Forest Regressor	**0,0145**	6%	4%	0,0004
Объединенные	Упорядоченное	Random Forest Regressor	0,0186	8%	5%	0,0007
«Пятьдесят оттенков серого» → «На пятьдесят оттенков темнее»						
BERT	-	Ada Boost Regressor	0,0169	7%	4%	**0,0005**
Основные	-	Random Forest Regressor	0,0171	7%	4%	**0,0005**
Объединенные	-	ElasticNet	**0,0168**	7%	4%	**0,0005**

Вывод: предсказание лучше получается в рамках одной книги

Предсказание сигналов заинтересованности

Предсказание в рамках одной книги



Предсказание в рамках двух книг

Вывод: предсказание на второй книге улавливает «основные тенденции»

Результаты

- Построены шесть сигналов заинтересованности, которые дополняют друг друга
- Построен хорошо интерпретируемый интегрированный сигнал заинтересованности
- Построены регрессионные модели для предсказания интереса
  - Предсказание получается лучше в рамках одной книги
  - Лучшие результаты дает предсказание по представлениям BERT
  - Предсказание на второй книге показывает основные тенденции
- По результатам работы планируется статья

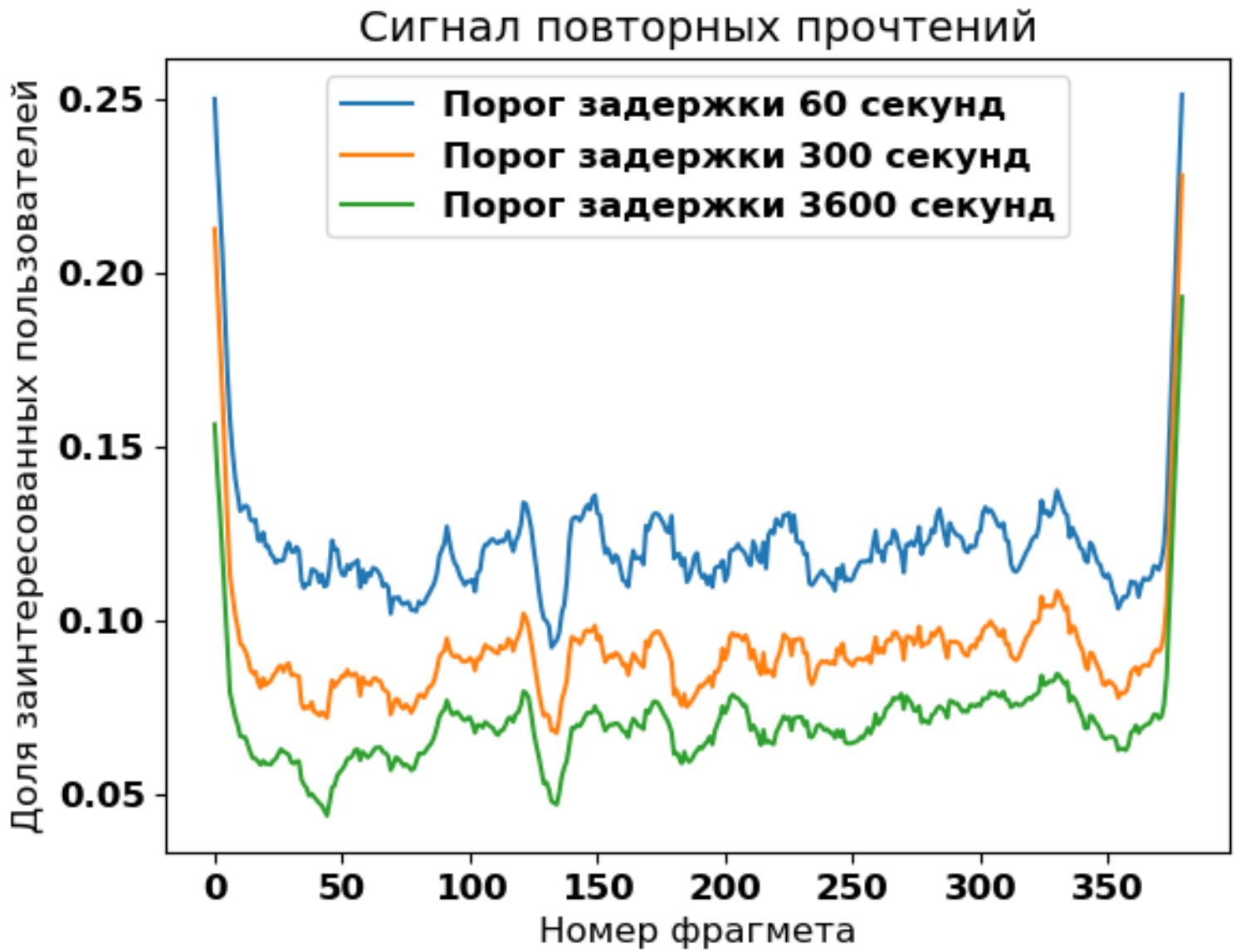
https://github.com/Igor-Tukh/bookmate/tree/metasessions_processing/src/metasessions_module

- Brouwer, Anne-Marie et al. “Physiological signals distinguish between reading emotional and non-emotional sections in a novel”. In: *Brain-Computer Interfaces* 2.2-3 (2015), pp. 76–89.
- Cupchik, Gerald and Janos Laszlo. “The Landscape of Time in Literary Reception: Character Experience and Narrative Action”. In: *Cognition & Emotion - COGNITION EMOTION* 8 (July 1994), pp. 297–312. DOI: 10.1080/02699939408408943.
- Lagun, Dmitry and Mounia Lalmas. “Understanding user attention and engagement in online news reading”. In: *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. 2016, pp. 113–122.
- Nell, Victor. “The Psychology of Reading for Pleasure: Needs and Gratifications”. In: 1988.


Smadja, Uzi et al. “Understanding Reader Backtracking Behavior in Online News Articles”. In: *The World Wide Web Conference*. WWW ’19. San Francisco, CA, USA: Association for Computing Machinery, 2019, pp. 3237–3243. ISBN: 9781450366748. DOI: 10.1145/3308558.3313571. URL: https://doi.org/10.1145/3308558.3313571.

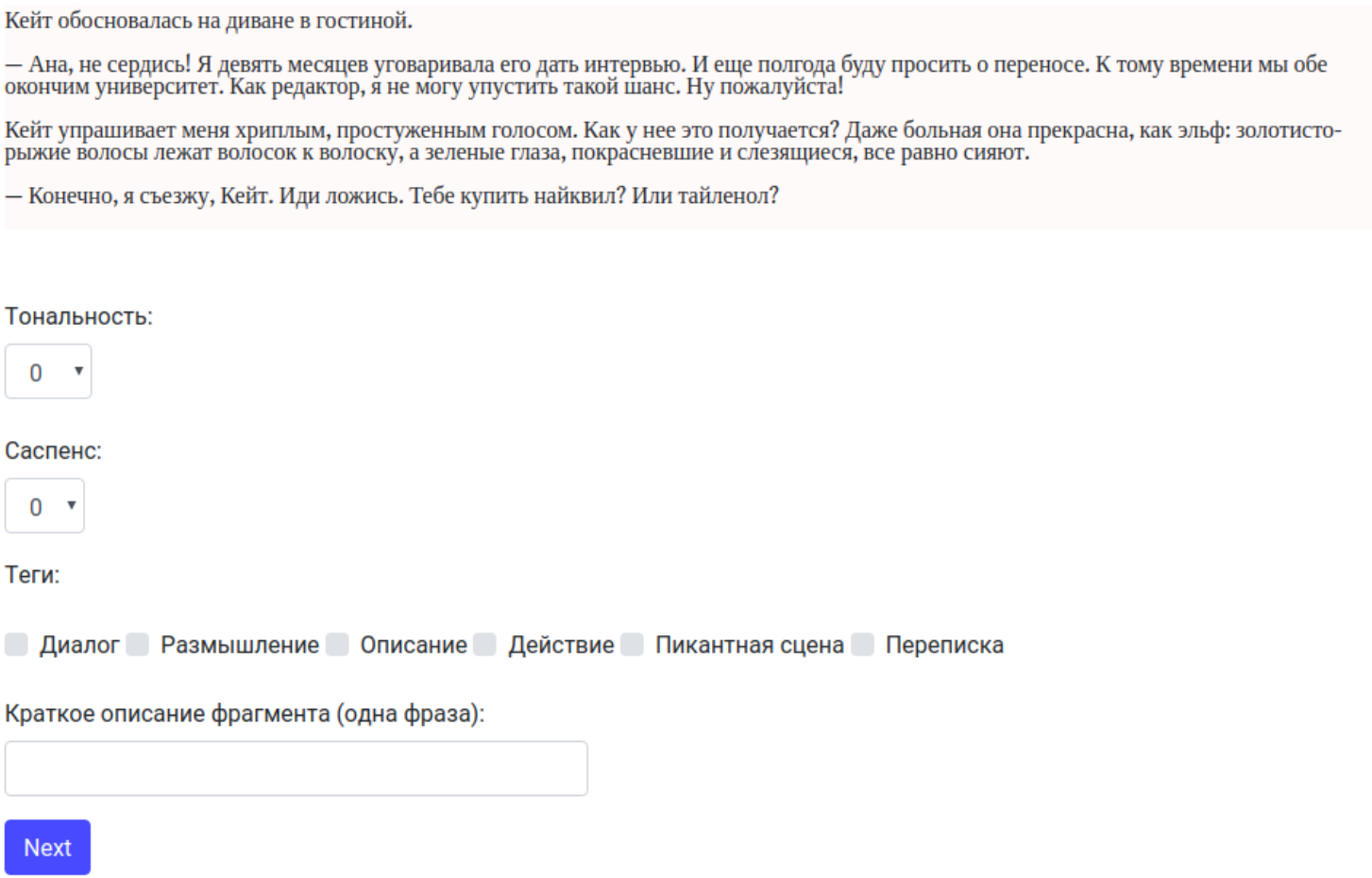

Tukh, Igor, Pavel Braslavski, and Kseniya Buraya. “Log-Based Reading Speed Prediction: A Case Study on War and Peace”. In: *Analysis of Images, Social Networks and Texts*. Ed. by Wil M. P. van der Aalst et al. Cham: Springer International Publishing, 2019, pp. 122–133. ISBN: 978-3-030-37334-4.

Извлечение сигналов заинтересованности



Сбор дополнительной разметки

Интерфейс инструмента для сборки разметки



https://fifty-shades-of-grey-marking.herokuapp.com/

Анализ сигналов заинтересованности

Таблица «похожести» сигналов

	1	2	3	4	5	6
1	1,00	0,97	0,02	0,45	0,91	0,09
2	0,97	1,00	0,02	0,45	0,89	0,09
3	0,02	0,02	1,00	0,02	0,02	0,03
4	0,45	0,45	0,02	1,00	0,48	0,02
5	0,91	0,89	0,02	0,48	1,00	0,07
6	0,09	0,09	0,03	0,02	0,07	1,00

1. Бросания
2. Прерывания
3. Необычные времена чтения
4. Высокая скорость
5. Пролистывания
6. Повторные чтения

Анализ сигналов заинтересованности

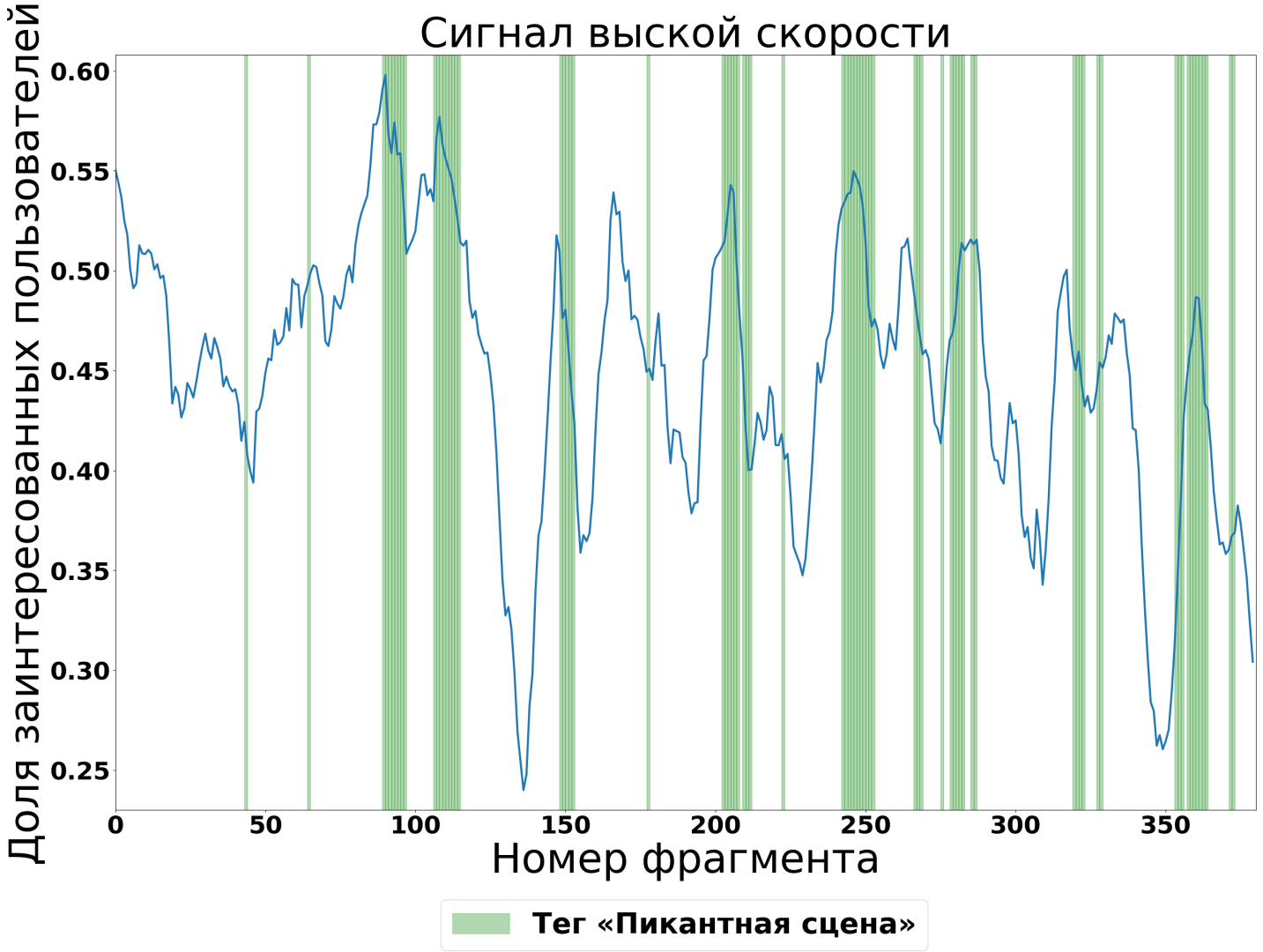
Наиболее частые «проявления» сигналов:

1	2	3	4	5	6	Процент проявлений
1	1	0	1	1	0	40,08
1	1	0	0	1	0	39,16
1	1	0	0	0	0	4,52
1	1	0	0	1	1	5,48
1	1	0	0	0	1	1,92
0	0	0	0	0	0	1,76

1. Бросания
2. Прерывания
3. Необычные времена чтения
4. Высокая скорость
5. Пролистывания
6. Повторные чтения

Анализ сигналов заинтересованности

Сопоставление сигнала с тегом «Пикантная сцена»



Кластеризация пользователей

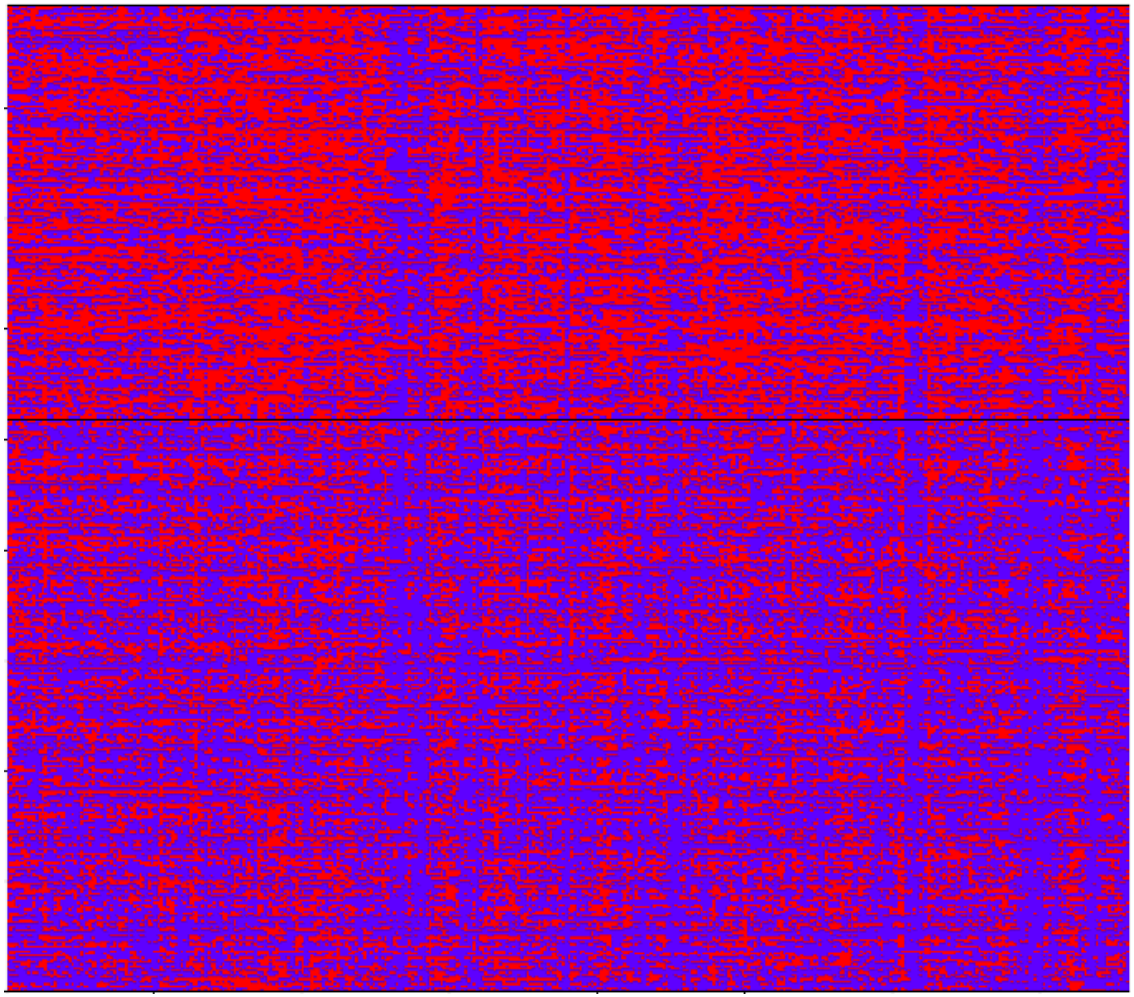
Можно ли выделить «характерные» группы читателей для каждого из сигналов заинтересованности?

- Кластеризированы все «хорошие» читатели
- Различные подходы (KMeans / Agglomirative / Spectral)
- Различные метрики для оценки качества кластеризации (Silhouette Coefficient / Calinski-Harabasz Index / Davies-Bouldin Index)
- Хорошо кластеризуется только сигнал высокой скорости
- Небольшие кластеры определяются для сигналов пролиыстываний и повторного чтения

Вывод: Сигналы дополняются друг друга. При этом проявляются примерно одинаково для всех пользователей.

Кластеризация пользователей

Проявления сигнала высокой скорости



Интегрированные сигналы заинтересованности

Предсказание сигналов заинтересованности

- Linear Regression
- Ridge
- Lasso
- ElasticNet
- ElasticNetCV
- LARS
- SGD Regressor
- SVM Regressor
- K Neighbors Regressor
- Decision Tree Regressor
- Ada Boost Regressor
- Random Forest Regressor

Предсказание сигналов заинтересованности

Сигнал	Разбиение	Средняя абсолютная ошибка	Минимальное значение сиганала	Максимальное значение сиганала	% от значения сигнала	% от длины диапазона
Интгерированный сигнал	Одна книга	0,0145	0,4044	0,6341	4%	6%
	Две книги	0,0168	0,3948	0,6341	4%	7%
Взвешенный интгерированный сигнал	Одна книга	0,0435	0,3940	0,8018	11%	11%
	Две книги	0,0399	0,3940	0,8018	10%	10%
Сигнал остановки чтения	Одна книга	0,0004	0,1842	1,0000	0%	0%
	Две книги	0,0027	0,1192	1,0000	2%	0%
Сигнал высокой скорости чтения	Одна книга	0,0809	0,0748	0,7596	108%	12%
	Две книги	0,0879	0,0748	0,7806	118%	12%
Сигнал повторных прочтений	Одна книга	0,0126	0,0276	0,3184	46%	4%
	Две книги	0,0174	0,0276	0,3333	63%	6%
Сигнал прерываний в процессе чтения	Одна книга	0,0089	0,9233	0,9954	1%	12%
	Две книги	0,0103	0,9230	0,9954	1%	14%
Сигнал необычных времен чтения	Одна книга	0,00510	0,0046	0,0362	111%	16%
	Две книги	0,00421	0,0019	0,0350	222%	13%
Сигнал пролситываний	Одна книга	0,0166	0,6763	0,9750	2%	6%
	Две книги	0,0165	0,6163	0,9750	3%	5%

Анализ признаков текста произведений