

Применение алгоритмов составления фильмотеки в системах индивидуальных рекомендаций

Студент: Козлов Сергей Андреевич

Научный руководитель: Мусатов Даниил Владимирович

Теория общественного выбора

Коллективное решение на основе индивидуальных предпочтений.

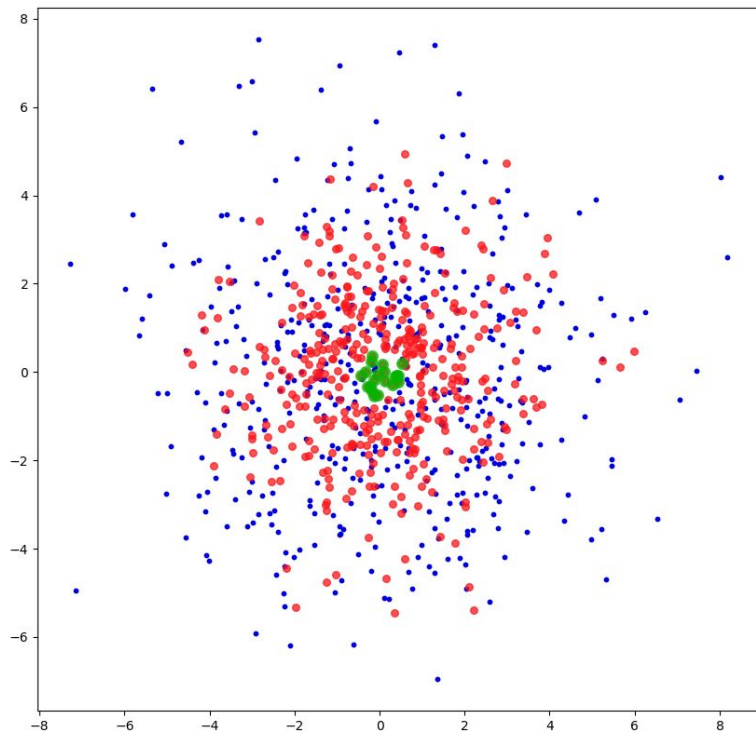
- Задача выбора комитета;
- Задача составления фильмотеки.

Рекомендательные системы

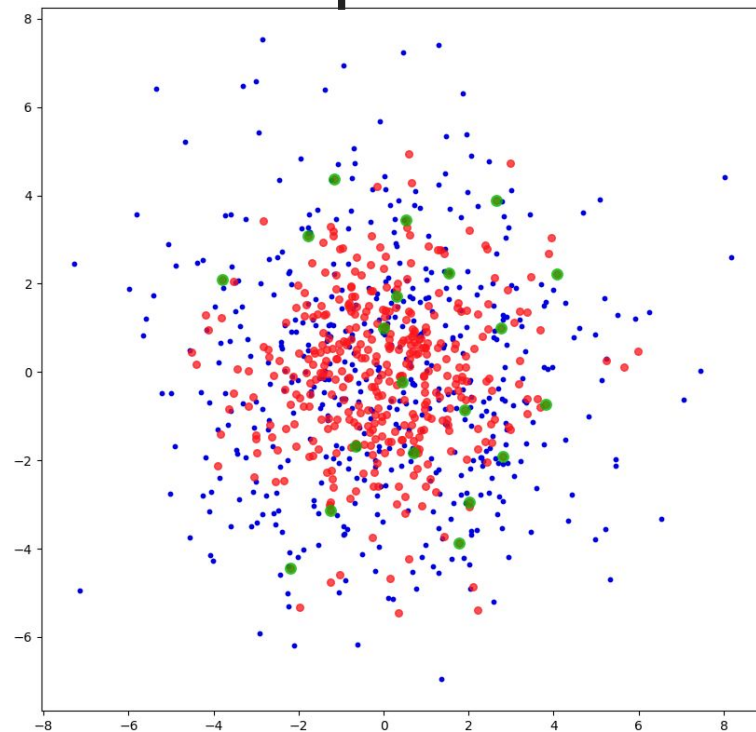
Персонализация контента с помощью индивидуальных рекомендаций объектов пользователю на основе его предыдущих действий.

Составление фильмотеки

плохо



хорошо



Задача составления фильмотеки

Выборы: $E = (V, C)$

Множество избирателей: $V = \{v_1, \dots, v_n\}$

Множество кандидатов: $C = \{c_1, \dots, c_m\}$

Комитет: $W = \{w_1, \dots, w_k\} \subset C$

Правило поиска комитета: $R : (E, k) \rightarrow W \subset C, |W| = k$

Функция полезности: $u_i : C \rightarrow \mathbb{R}_+$

$$U_i(W) = \max_{w \in W} u_i(w)$$

Проблемы рекомендательных систем

- **Разреженность данных** — матрица взаимодействия часто разреженная;
- **Переобучение на популярность** — более популярные объекты чаще попадают в рекомендации;
- **Вычислительная сложность** — некоторые современные алгоритмы требовательны к вычислительным ресурсам;
- **Использование пользовательских данных** — улучшение качества рекомендаций достигается за счёт сбора информации о пользователе;
- **Неинтерпретируемость** — алгоритмы машинного обучения обычно не дают обоснование выбранным рекомендациям;
- **Информационные пузыри** — небольшое разнообразие рекомендаций приводит к взаимодействию с объектами внутри одной категории.

Задача построения индивидуальных рекомендаций

Индивидуальные рекомендации

Множество
пользователей: $\mathcal{U} = \{u_1, \dots, u_N\}$

Множество объектов: $\mathcal{I} = \{i_1, \dots, i_M\}$

Матрица
взаимодействия: $\mathcal{R} \in \mathbb{R}^{N \times M}$

Рекомендации: $\mathcal{J}^u \subset \mathcal{I} : |\mathcal{J}^u| = k, \forall i \in \mathcal{J}^u \mathcal{R}_{ui} = 0$

Индивидуальные рекомендации

Множество
пользователей: $\mathcal{U} = \{u_1, \dots, u_N\}$

Множество объектов: $\mathcal{I} = \{i_1, \dots, i_M\}$

Матрица
взаимодействия: $\mathcal{R} \in \mathbb{R}^{N \times M}$

Рекомендации: $\mathcal{J}^u \subset \mathcal{I} : |\mathcal{J}^u| = k, \forall i \in \mathcal{J}^u \mathcal{R}_{ui} = 0$

Расширенная матрица
взаимодействия: $\mathcal{R}^* \in \mathbb{R}^{N \times M}, \forall u, i : \mathcal{R}_{ui}^* > 0$

Оптимальные
рекомендации:

$$\mathcal{J}^* = \arg \max_{\mathcal{J}^u} \|\{\mathcal{R}_{ui}^*\}_{i \in \mathcal{J}^u}\|_p$$

Метрики индивидуальных рекомендаций

- Релевантность:

- $$\bullet \text{ Precision@k} = \frac{\text{Число релевантных объектов в рекомендациях}}{k}$$

- $$\bullet \text{ Recall@k} = \frac{\text{Число релевантных объектов в рекомендациях}}{\text{Число всех релевантных объектов}}$$

- Ранжирование:

- $$\bullet \text{ DCG@k}$$

- $$\bullet \text{ nDCG@k} = \frac{\text{DCG@k}}{\text{Идеальное DCG@k}}$$

- Разнообразие:

- $$\bullet \text{ MIUF@k}$$

- $$\bullet \text{ Serendipity@k}$$

Сведение индивидуальных рекомендаций к задаче составления фильмотеки

Идея алгоритма

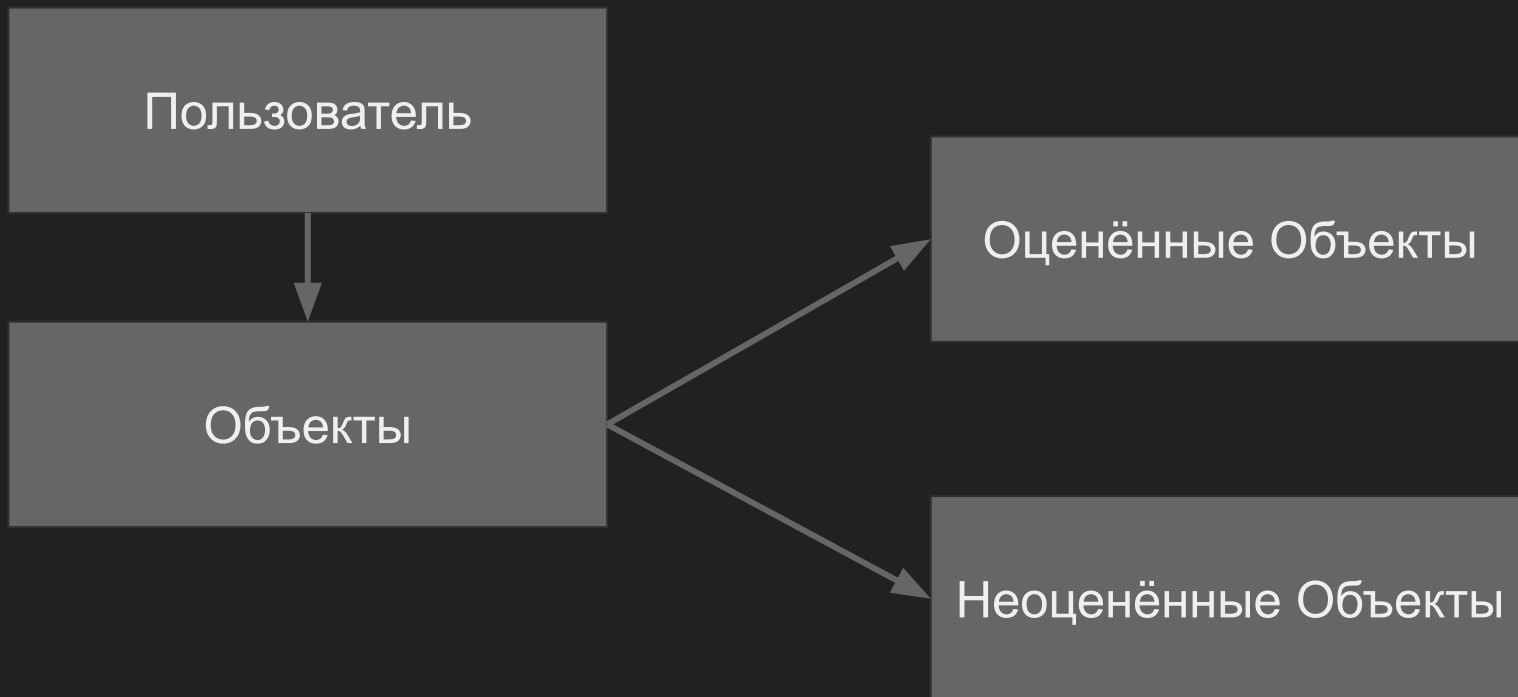
Пользователь

Объекты

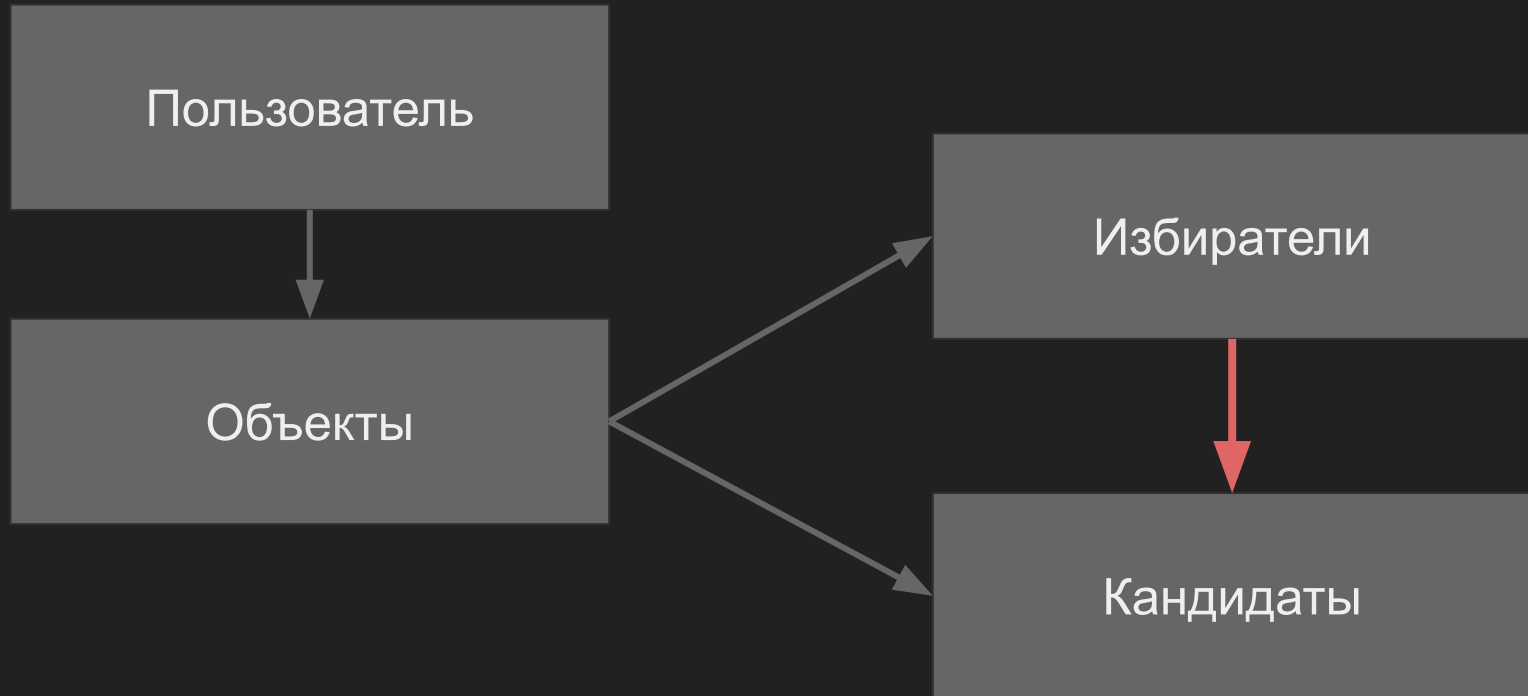
Идея алгоритма



Идея алгоритма



Идея алгоритма



Проблемы применения алгоритмов составления фильмотеки

- Для определения функций полезности необходимо расстояние между объектами;
- Любой выборный алгоритм предполагает равенство избирателей.

Алгоритмы составления фильмотеки

Алгоритмы составления фильмотеки

SNTV

$$PS(c_i) = \sum_{j=1}^n [pos_{v_j}(c_i) = 1]$$

STV

$$q = \lceil \frac{|V|}{k+1} \rceil$$

Кандидат для избрания должен набрать PS хотя бы q. Если ни один не набрал, удаляем худшего и пересчитываем PS. После каждого избрания удаляем q удовлетворённых избирателей.

Алгоритмы составления фильмотеки

SNTV

$$PS(c_i) = \sum_{j=1}^n [pos_{v_j}(c_i) = 1]$$

STV

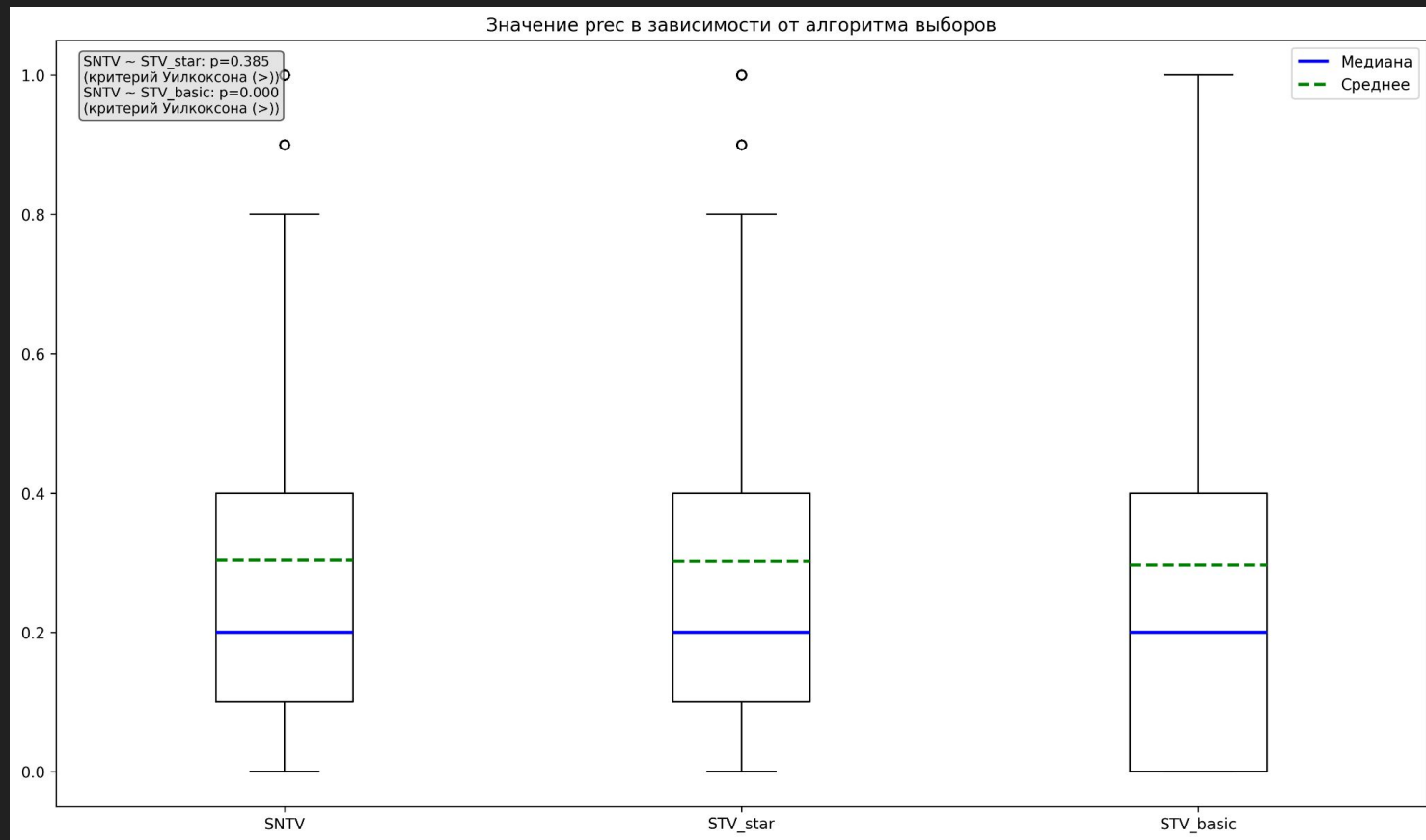
$$q = \lceil \frac{|V|}{k+1} \rceil$$

STV*

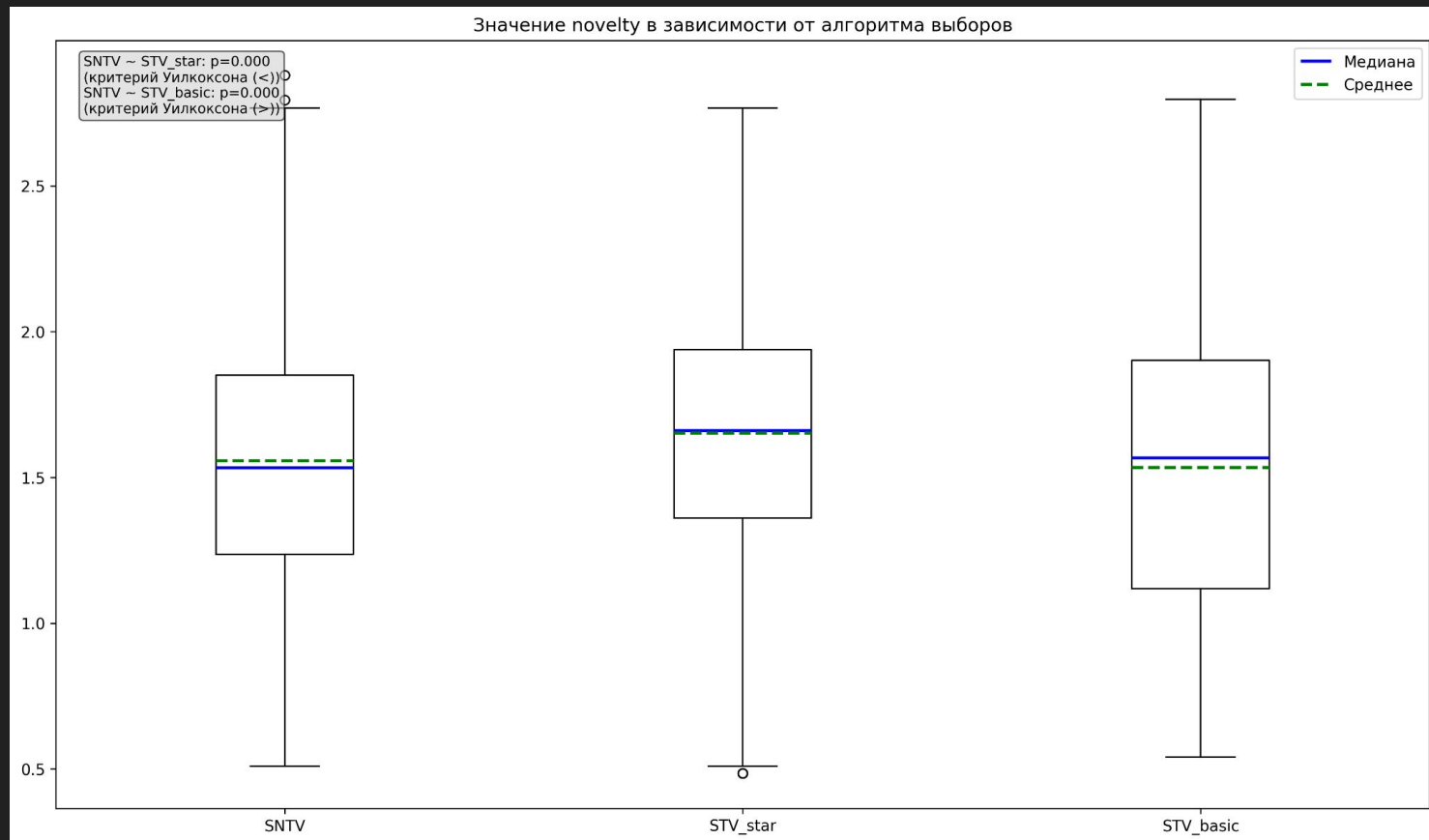
- Удаляем **всех** удовлетворённых избирателей;
- Пересчитываем q на каждом шаге;
- Также удаляем худшего кандидата, если ни один не набрал квоту.

Кандидат для избрания должен набрать PS хотя бы q . Если ни один не набрал, удаляем худшего и пересчитываем PS . После каждого избрания удаляем q удовлетворённых избирателей.

Влияние алгоритма составления фильмотеки на ТОЧНОСТЬ



Влияние алгоритма составления фильмотеки на разнообразие



Расстояния между объектами

Расстояние между объектами

- Косинусное расстояние:

- $\rho_{\cos}(i, j) = 1 - \frac{\langle r_i, r_j \rangle}{\|r_i\| \cdot \|r_j\|}$

- Корреляционное расстояние:

- $\rho_p(i, j) = 1 - \frac{\langle r_i - \bar{r}_i e, r_j - \bar{r}_j e \rangle}{\|r_i - \bar{r}_i e\| \cdot \|r_j - \bar{r}_j e\|}$

- $\rho_s(i, j) = 1 - \frac{6}{N^3 - N} \sum_{u=1}^N (R_u - S_u)^2$

- $\rho_k(i, j) = 1 - \frac{4}{N(N-1)} R$

Расстояние между объектами

- Расстояние Хэмминга:

- $ham(c, d) = |A(c) \setminus A(d)| + |A(d) \setminus A(c)|$

- Расстояние Жаккара:

- $jac(c, d) = 1 - \frac{|A(c) \cap A(d)|}{|A(c) \cup A(d)|} = \frac{ham(c, d)}{|A(c) \cup A(d)|}$

- Модификация расстояния Жаккара:

- $$\rho_n(c, d) = n - \sum_{i,j=1}^n S_{ij} \frac{|A_i(c) \cap A_j(d)|}{|A_i(c) \cup A_j(d)|}$$

где $S_{ij} = 1 - 2 \frac{|i-j|}{n-1}$

Расстояние между объектами

- Модификация расстояния Жаккара:

- $$\rho_n(c, d) = n - \sum_{i,j=1}^n S_{ij} \frac{|A_i(c) \cap A_j(d)|}{|A_i(c) \cup A_j(d)|} \quad \text{где } S_{ij} = 1 - 2 \frac{|i-j|}{n-1}$$

- Частный случай при $n = 2$:

- $$\rho_2(c, d) = 2 - \frac{|A(c) \cap A(d)|}{|A(c) \cup A(d)|} - \frac{|D(c) \cap D(d)|}{|D(c) \cup D(d)|} + \frac{|A(c) \cap D(d)|}{|A(c) \cup D(d)|} + \frac{|D(c) \cap A(d)|}{|D(c) \cup A(d)|}$$

Свойства методов вычисления расстояний

- *Косинусное расстояние и модифицированное расстояние Жаккара не чувствительны к добавлению пользователя;*
- *Все корреляционные расстояния чувствительны к добавлению пользователя;*

Свойства методов вычисления расстояний

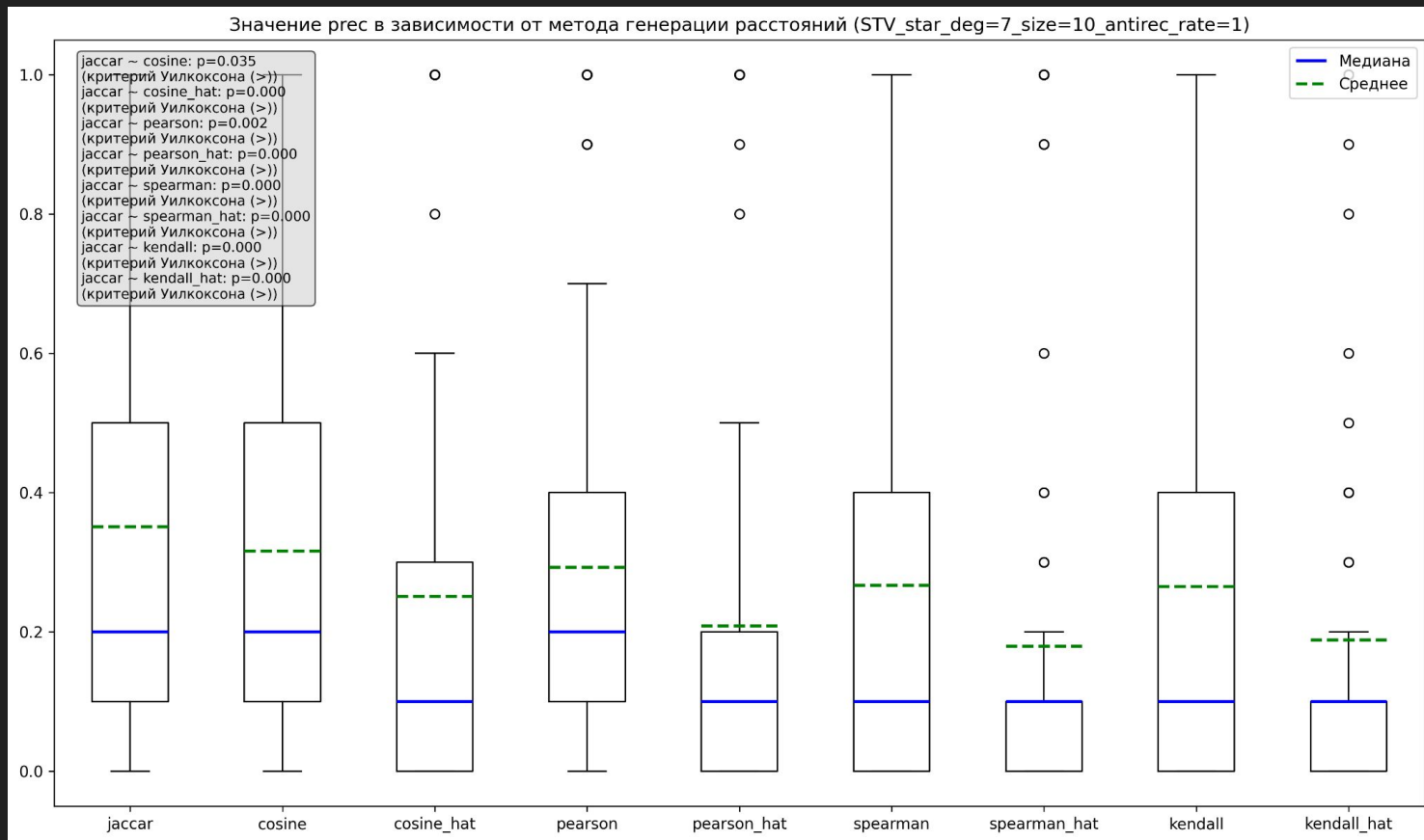
- Косинусное расстояние и модифицированное расстояние Жаккара не чувствительны к добавлению пользователя;
- Все корреляционные расстояния чувствительны к добавлению пользователя;
- Косинусное расстояние всегда положительно чувствительно к асимметричному вкладу;
- Расстояние Пирсона положительно чувствительно к асимметричному вкладу, если

$$\frac{\bar{r}_j}{2} < \langle r_i - \bar{r}_i e, r_j - \bar{r}_j e \rangle \leq \frac{\|r_i - \bar{r}_i e\|^2}{2}$$

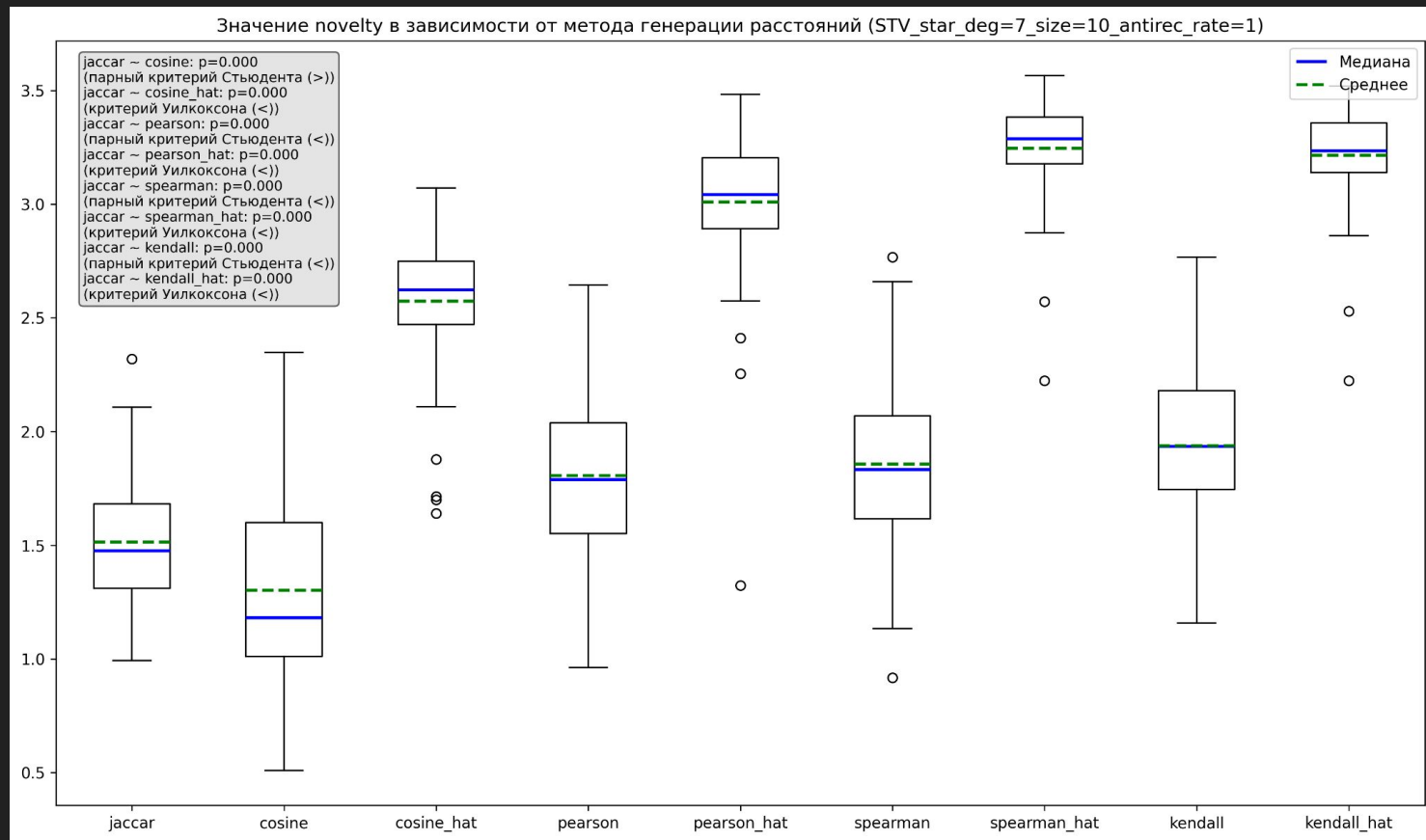
Чувствительность к асимметричному вкладу

Любой метод вычисления расстояний между объектами либо чувствителен к асимметричному вкладу, либо для разных пар объектов размерность пространства, в котором определяется расстояние, отличается.

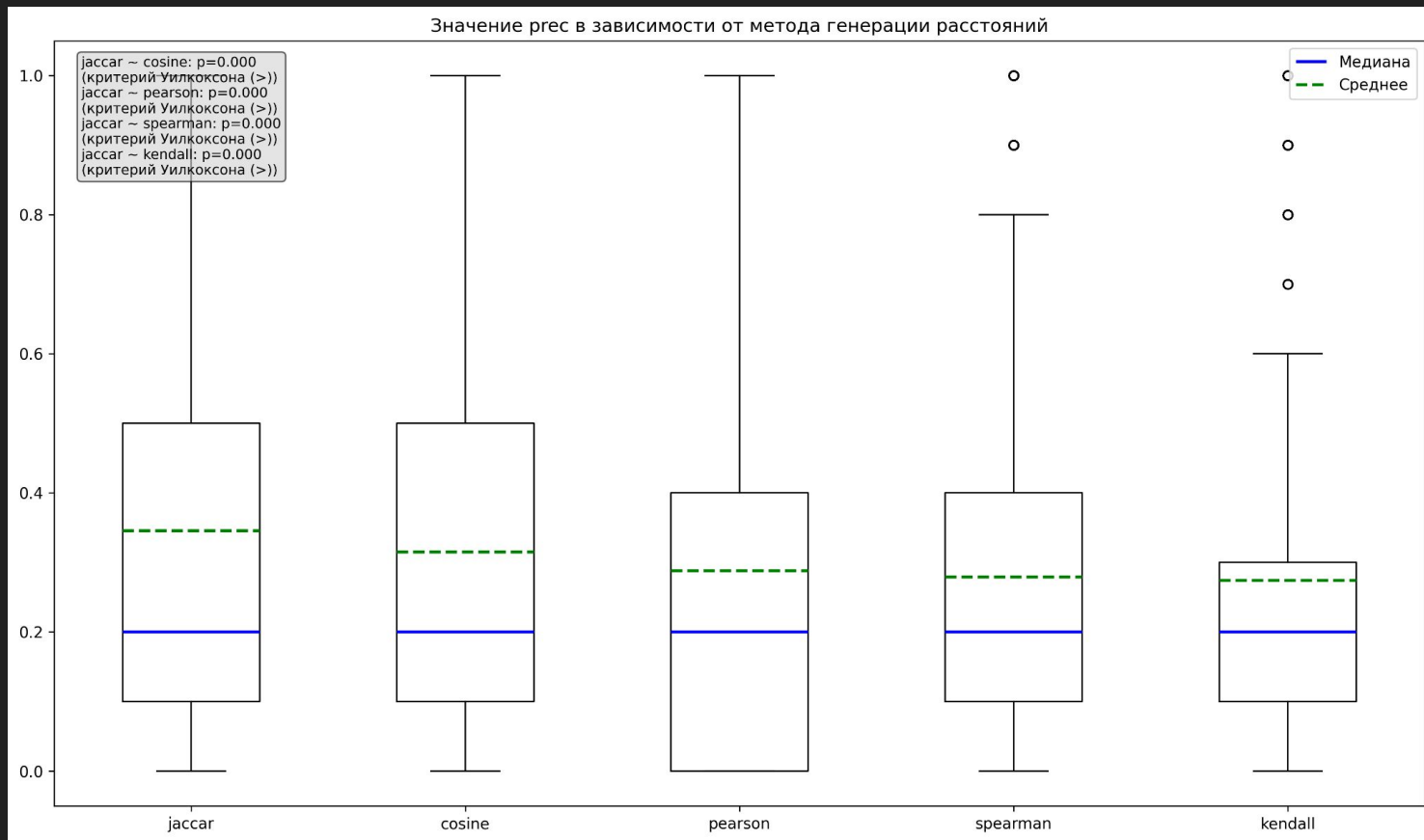
Влияние метода вычисления расстояний на точность (с нечувствительными к асим. вкладу расстояниями)



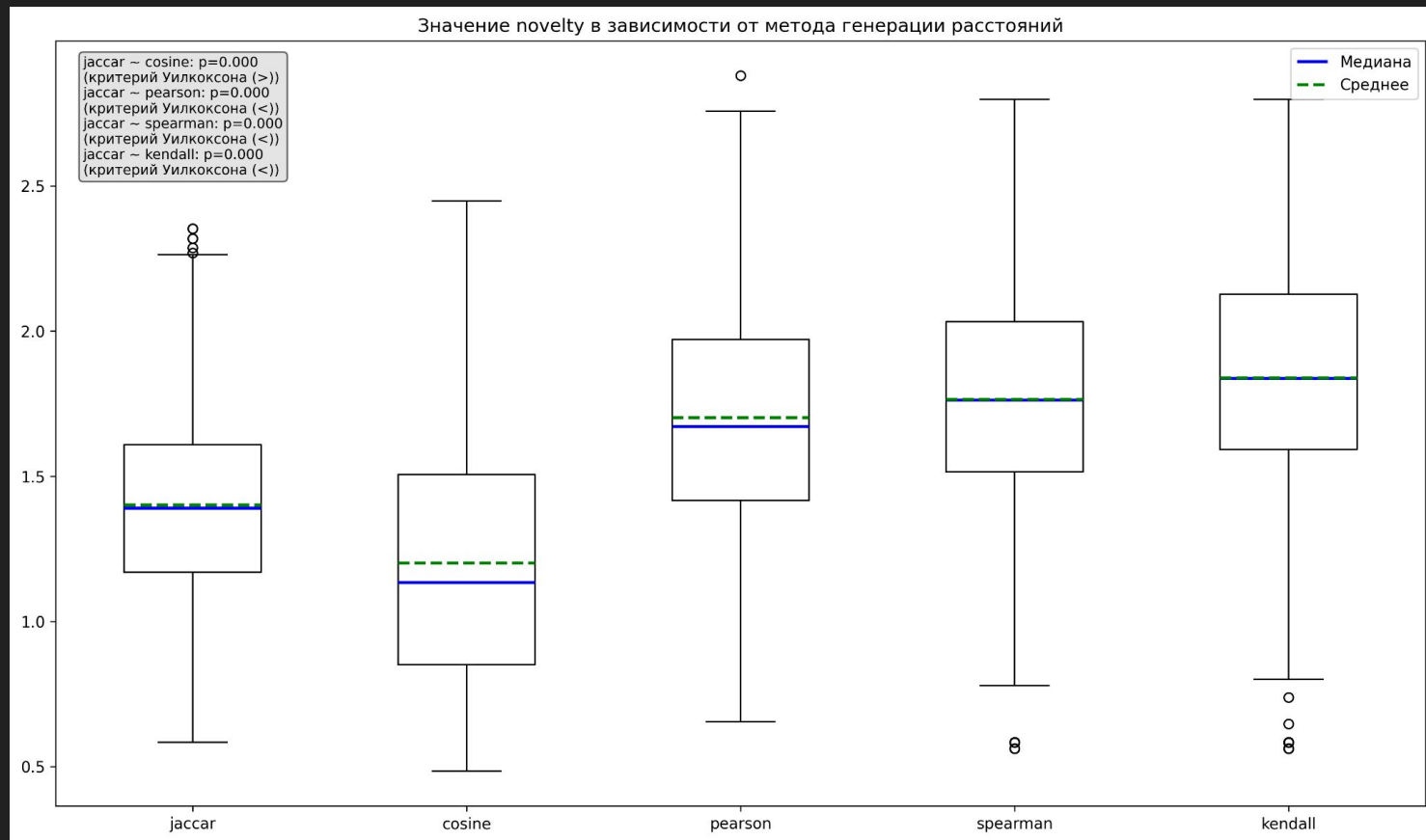
Влияние метода вычисления расстояний на разнообразие (с нечувствительными к асим. вкладу расстояниями)



Влияние метода вычисления расстояний на точность



Влияние метода вычисления расстояний на разнообразие



Выборы с неравными избирателями

Методы учёта оценок избирателей

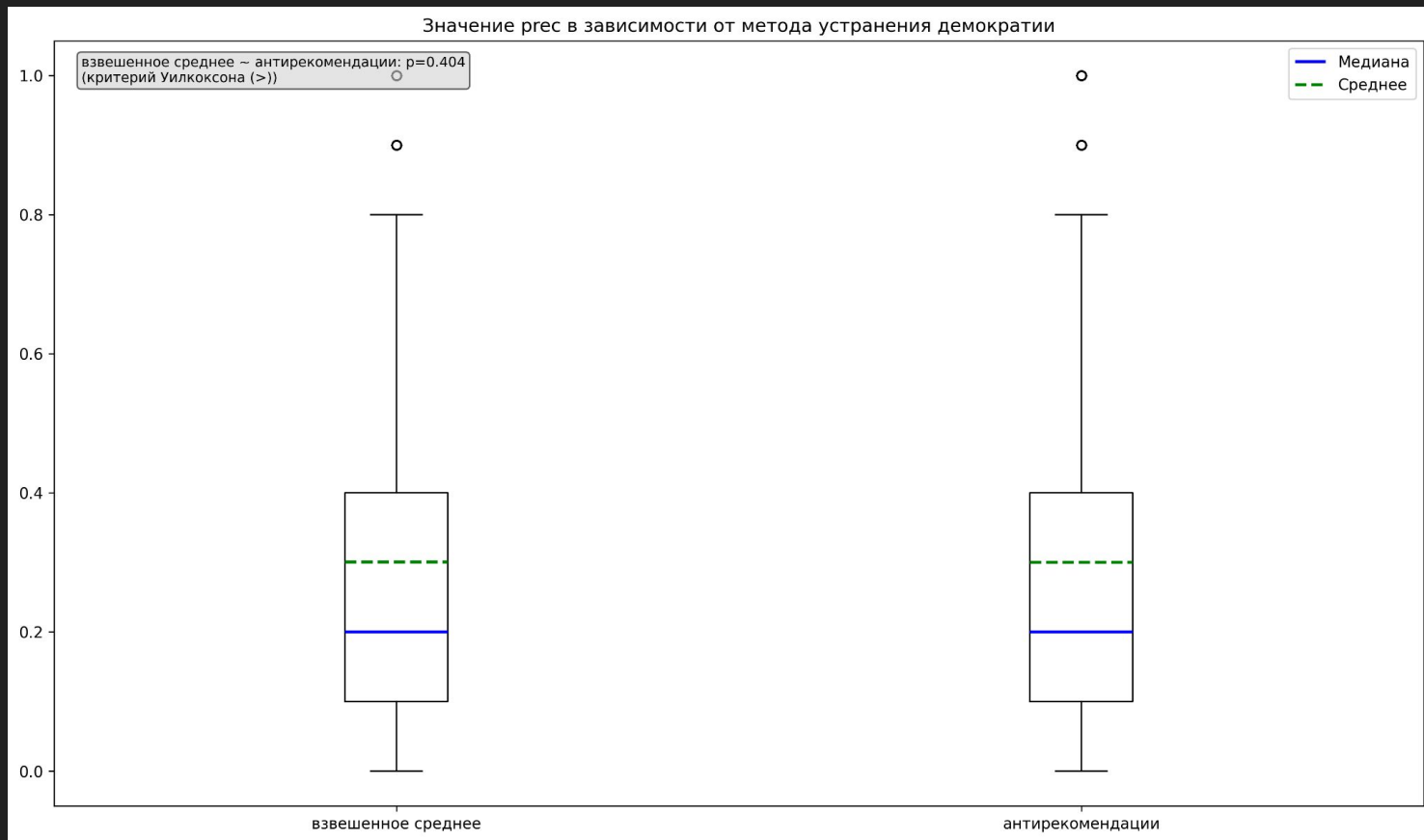
- Взвешенное среднее:

$$U(W) = \frac{1}{|V^u|} \sum_{i \in V^u} \mathcal{R}_{ui} \max_{w \in W} (\eta_i(w)).$$

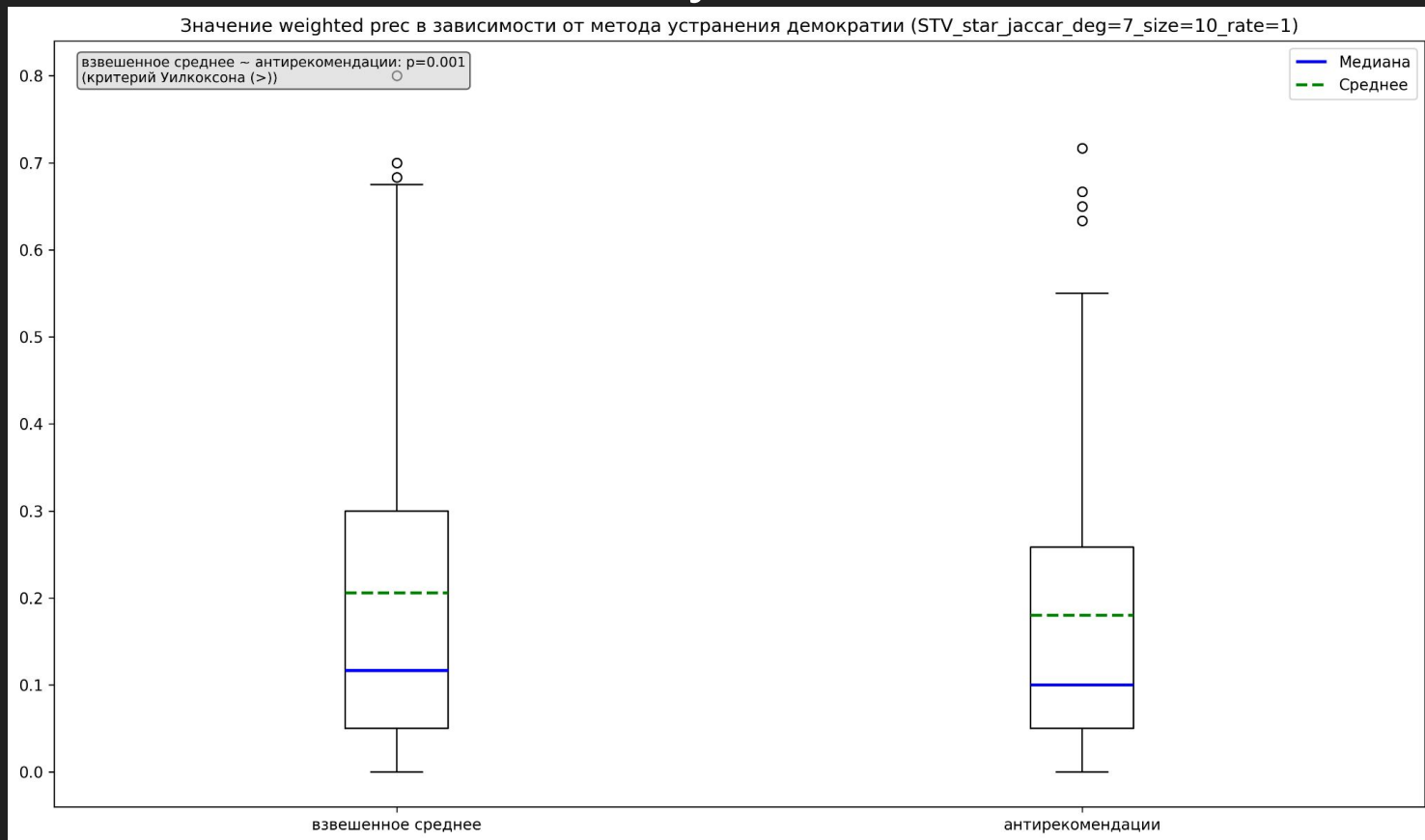
Методы учёта оценок избирателей

- Взвешенное среднее:
$$U(W) = \frac{1}{|V^u|} \sum_{i \in V^u} \mathcal{R}_{ui} \max_{w \in W} (\eta_i(w)).$$
- Антирекомендации:
 - Проводим выборы с худшими по оценкам избирателями;
 - Удаляем избранный комитет из списка кандидатов;
 - Проводим выборы с остальными избирателями и оставшимися кандидатами.

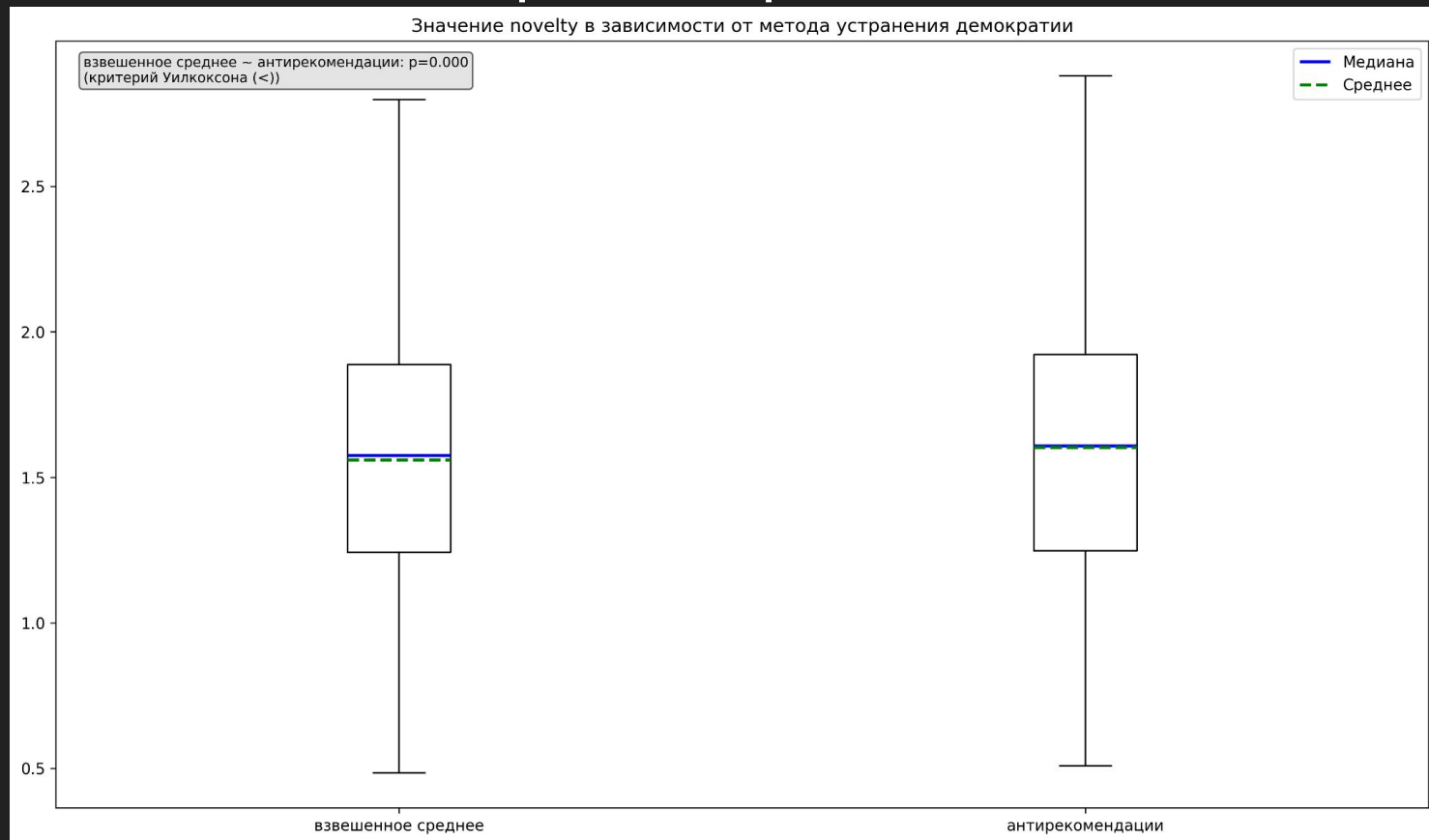
Влияние метода учёта оценок избирателей на точность



Влияние метода учёта оценок избирателей на взвешенную точность

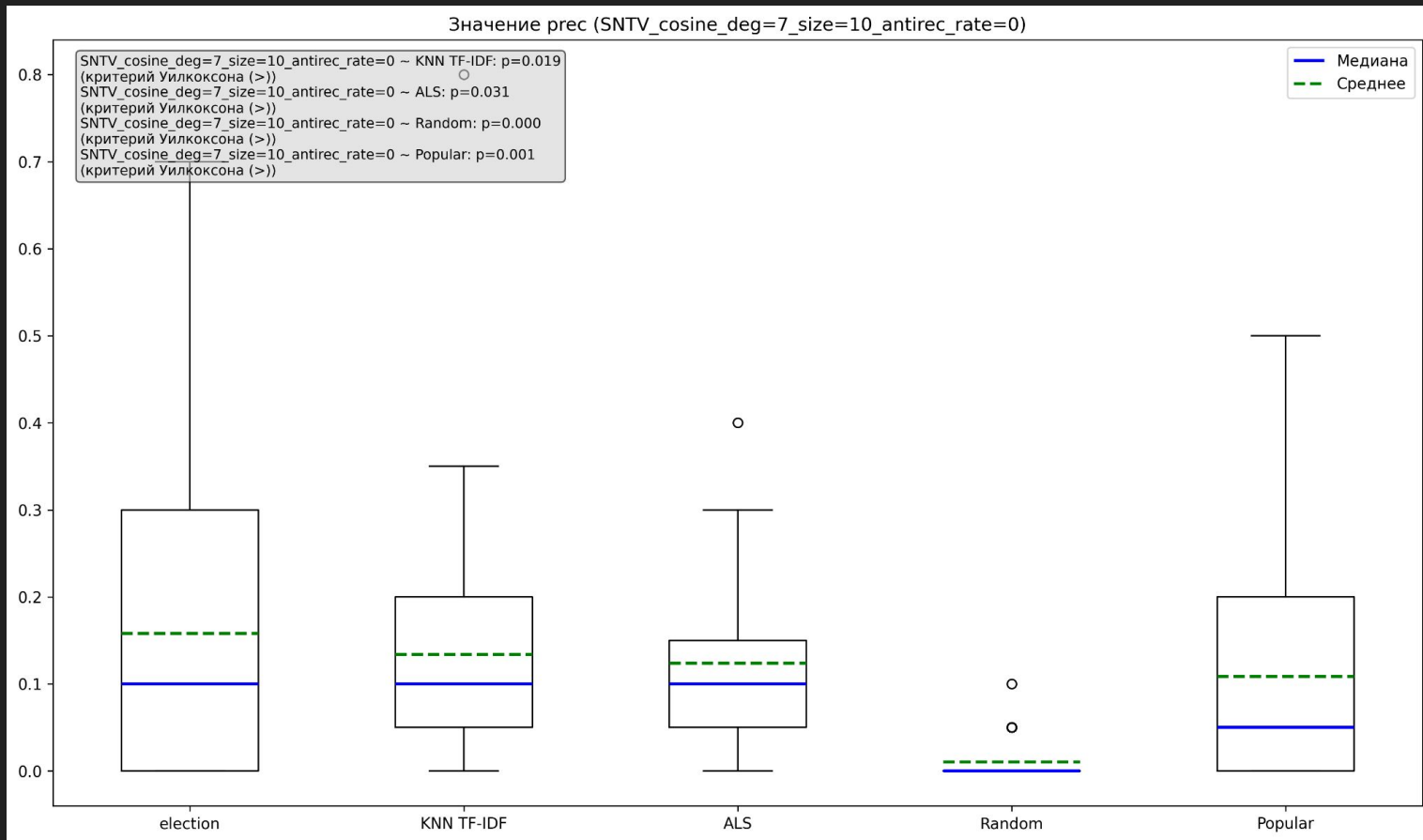


Влияние метода учёта оценок избирателей на разнообразие

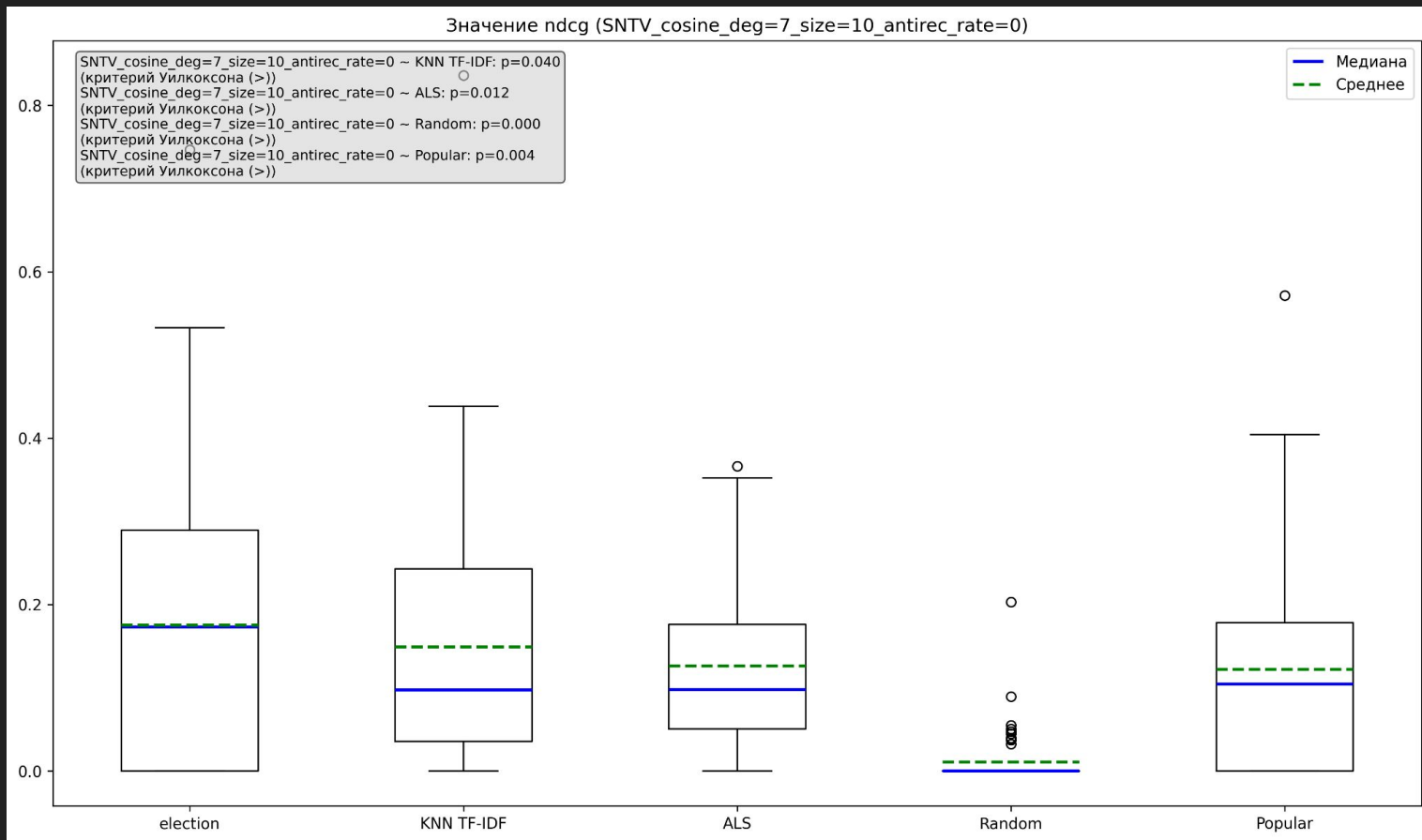


Сравнение с традиционными рекомендательными алгоритмами

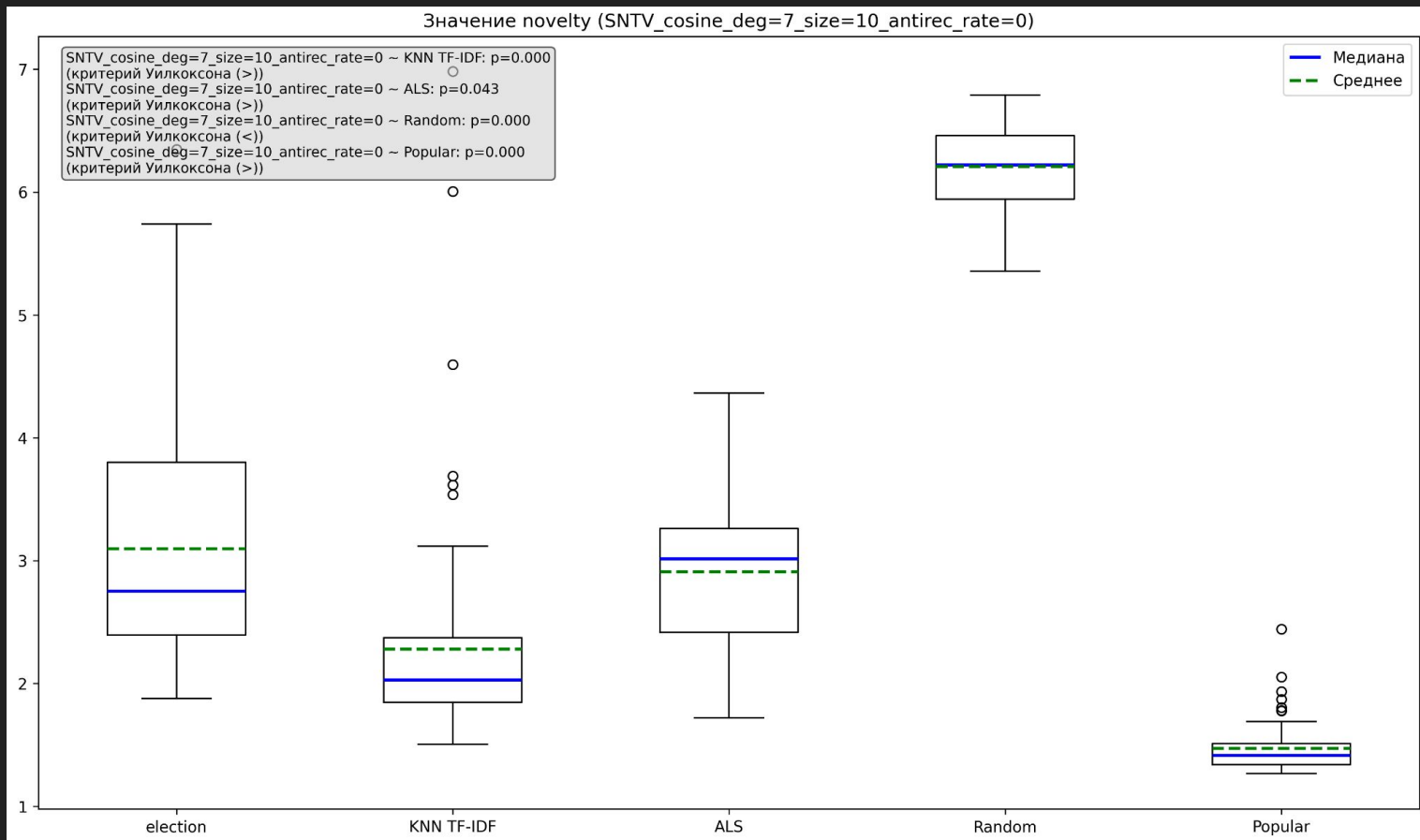
Сравнение рекомендательных алгоритмов — релевантность



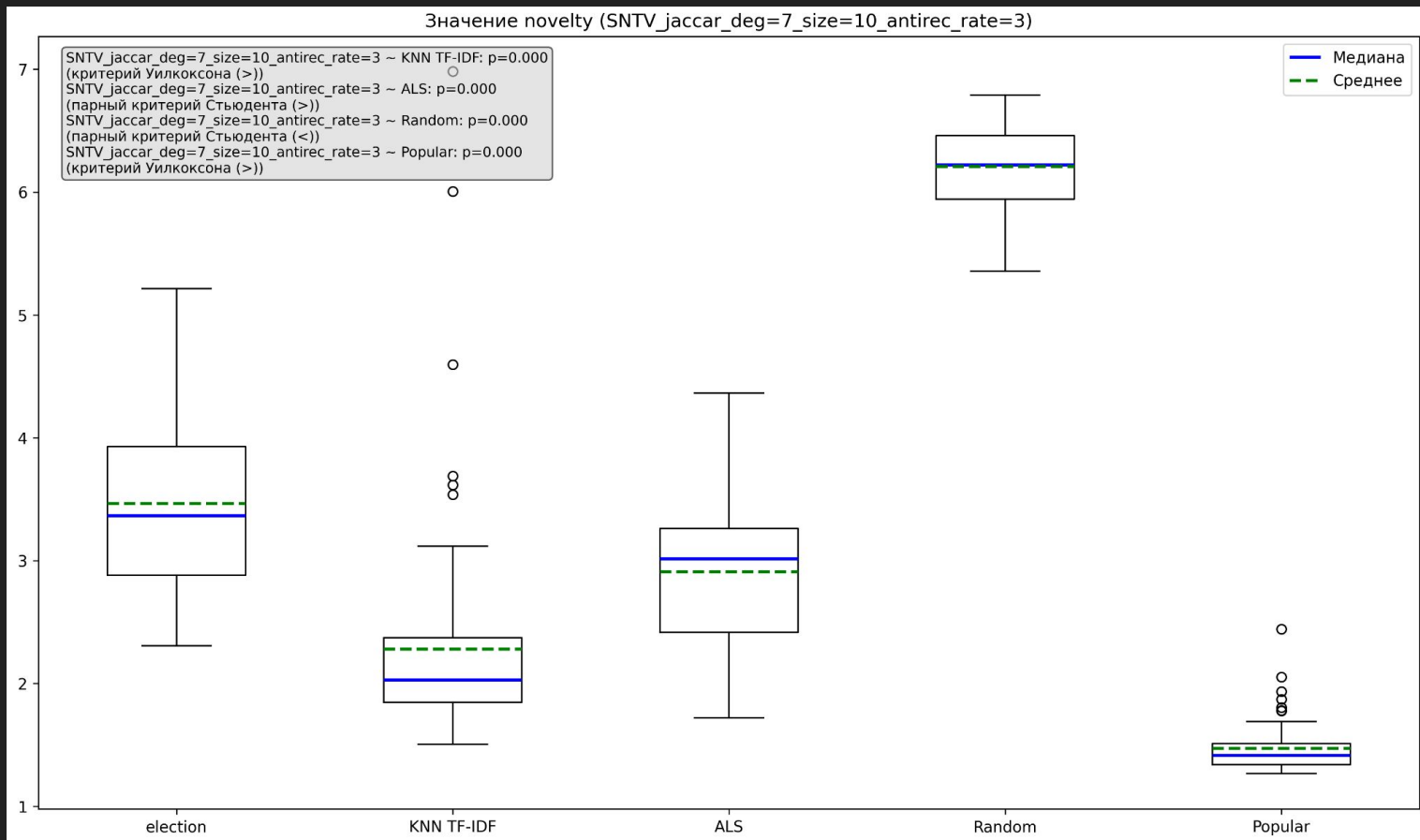
Сравнение рекомендательных алгоритмов — ранжирование



Сравнение рекомендательных алгоритмов — разнообразие



Сравнение рекомендательных алгоритмов — разнообразие



Интерпретации рекомендаций

рекомендации	ближайший сосед	оценка соседа
Вечное сияние чистого разума (2004) Eternal Sunshine of the Spotless Mind	Загадочная история Бенджамина Баттона The Curious Case of Benjamin Button, 2008	8.0
Игры разума A Beautiful Mind, 2001	Форрест Гамп Forrest Gump, 1994	10.0
Король говорит! The King's Speech, 2010	Дюна: Часть вторая (2024) Dune: Part Two	8.0
Залечь на дно в Брюгге (2007) In Bruges	Большой куш Snatch, 2000	9.0
Фарго (1995) Fargo	Бешеные псы (1991) Reservoir Dogs	9.0
Скотт Пилигрим против всех (2010) Scott Pilgrim vs. the World	Назад в будущее Back to the Future, 1985	8.0
Американская история X (1998) American History X	Курьер (1986)	8.0
Игра на понижение (2015) The Big Short	О чём говорят мужчины (2010)	9.0
Нечто (1982) The Thing	Бешеные псы (1991) Reservoir Dogs	9.0
Пленницы (2013) Prisoners	Семь Se7en, 1995	8.0

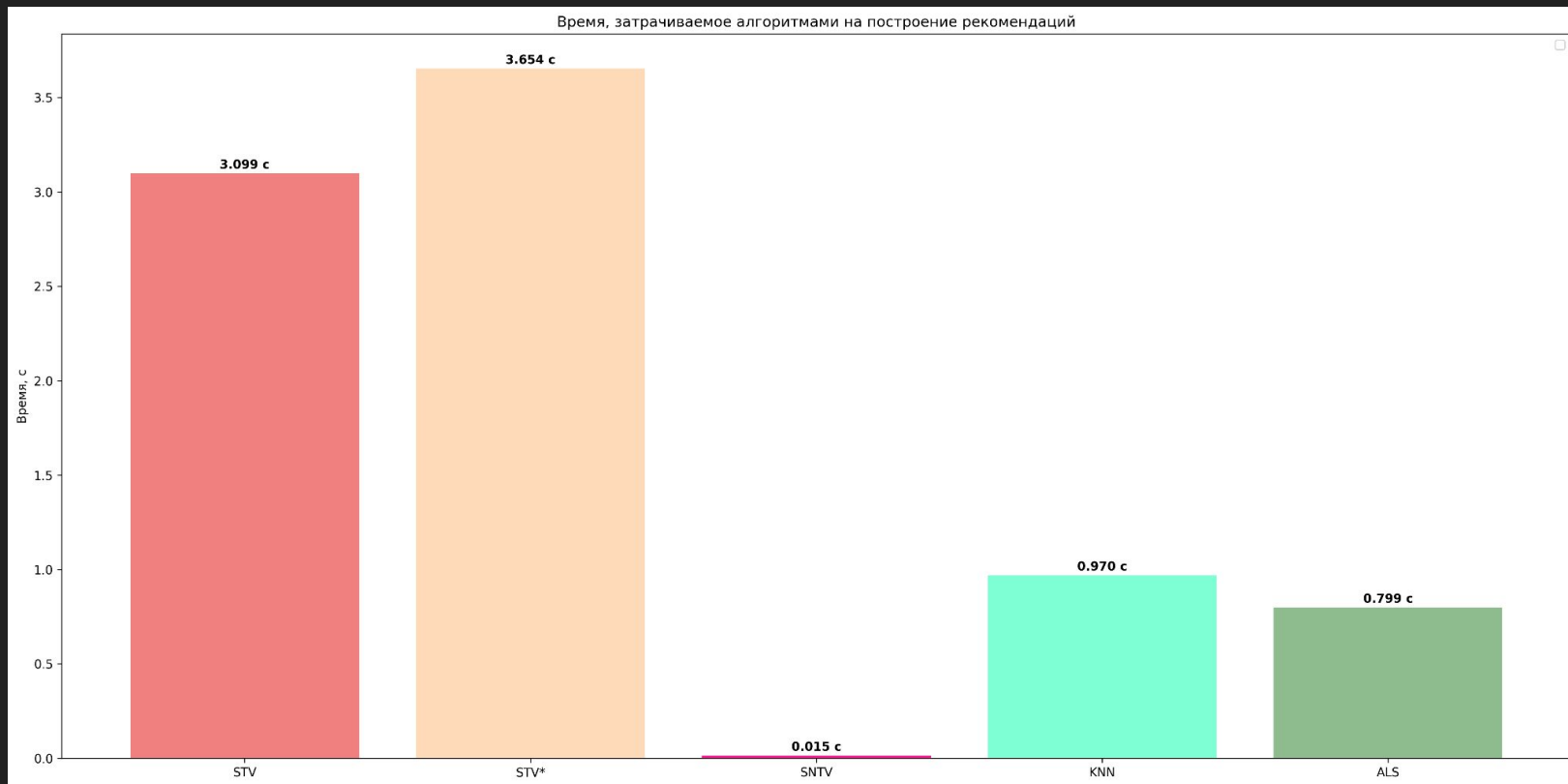
Выводы

- Предложен и реализован новый подход к индивидуальным рекомендациям;
- Исследованы различные параметры алгоритма и выявлены лучшие конфигурации;
- Разработаны новые методы вычисления расстояния, составления фильмотеки и учёта пользовательских оценок;
- Предложенный алгоритм обеспечивает высокие показатели всех метрик на разреженных данных;
- Частичное решение проблемы переобучения на популярность;
- Интерпретируемость результата;
- Высокая вычислительная эффективность в случае SNTV;
- Использование предложенного метода в рекомендательных системах целесообразно, и имеет смысл развивать этот метод в сторону гибридных рекомендательных систем.

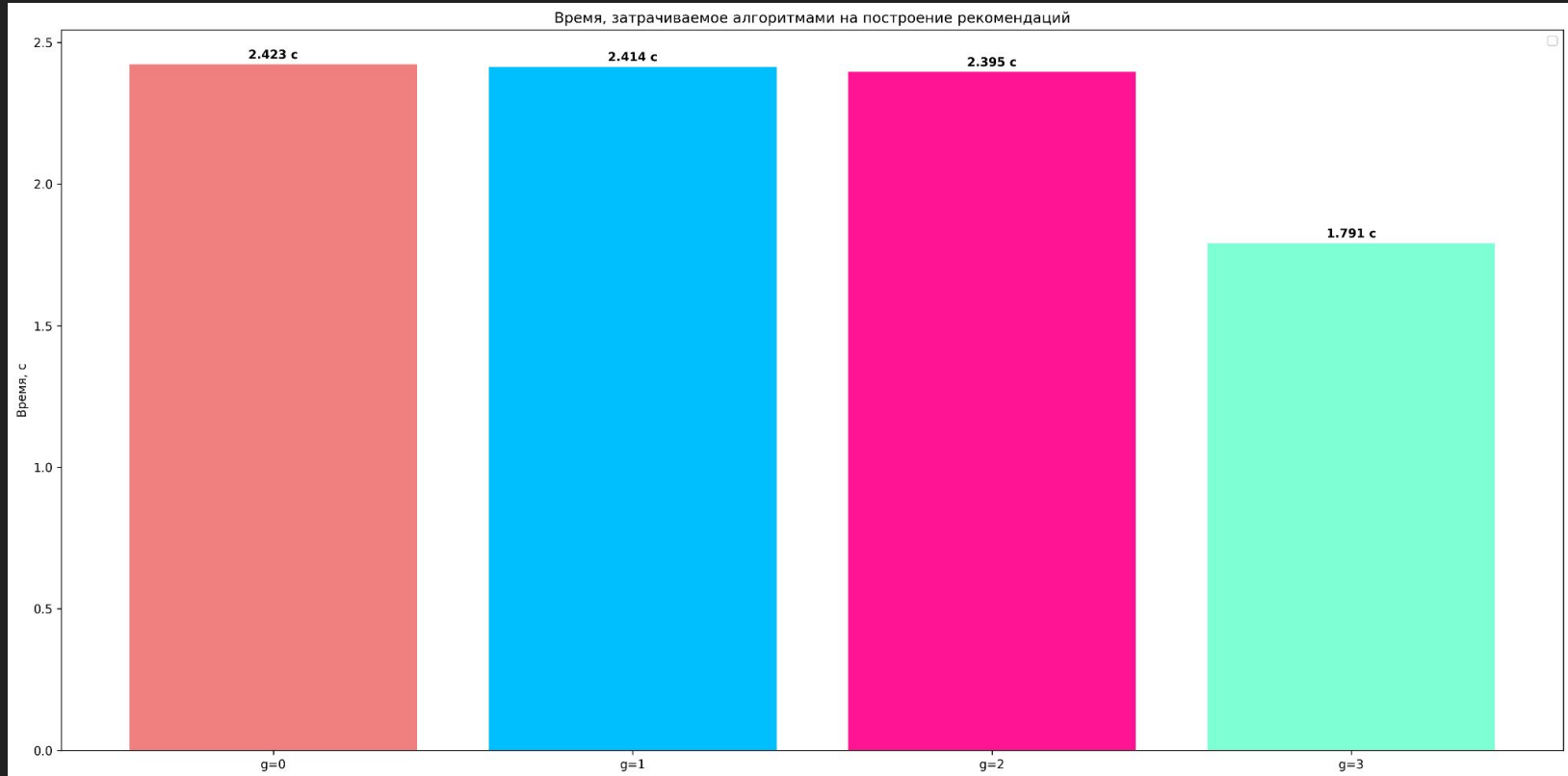
Дальше скрытые слайды

Время работы

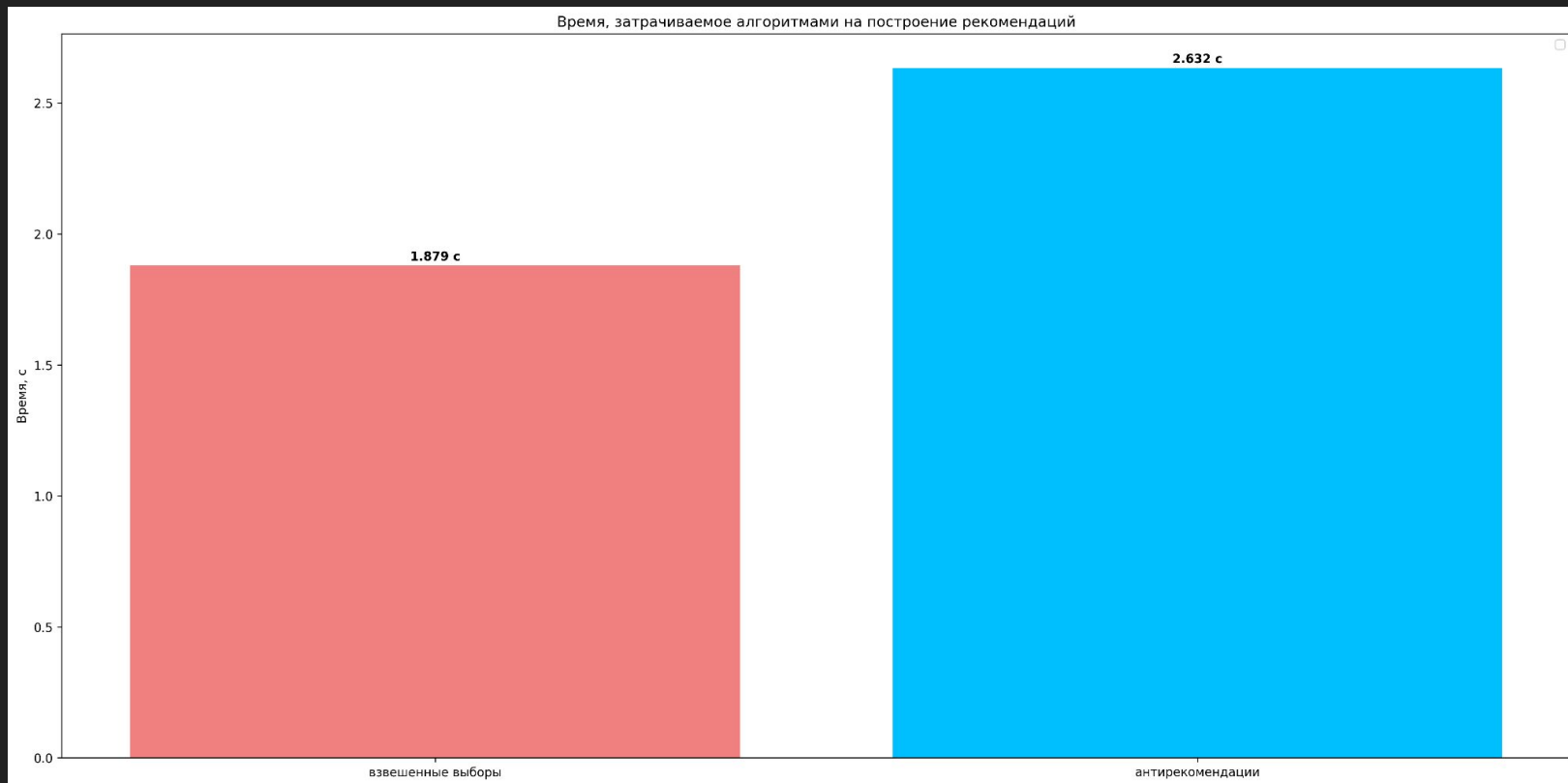
Сравнение рекомендательных алгоритмов — время



Влияние параметра серии на время

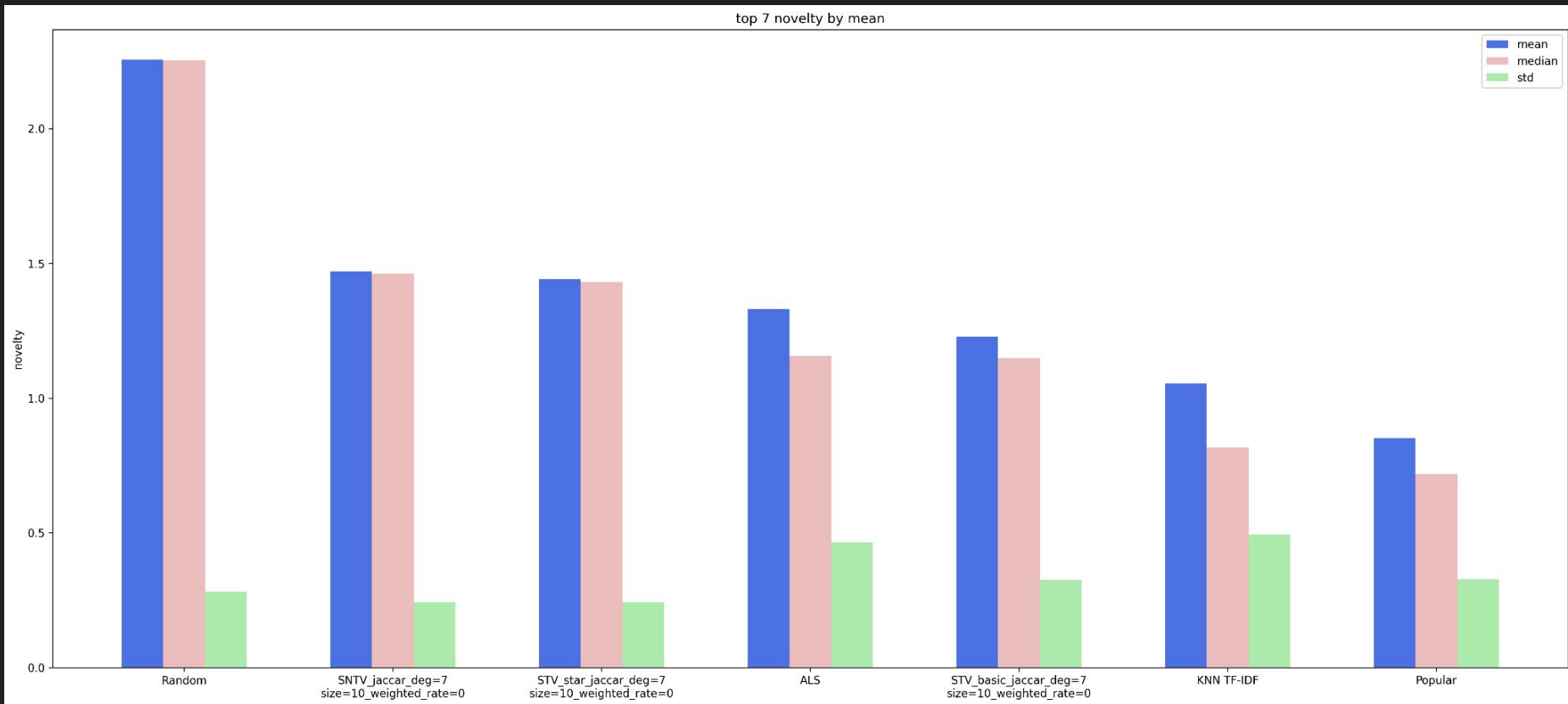


Влияние метода устранения демократии на время

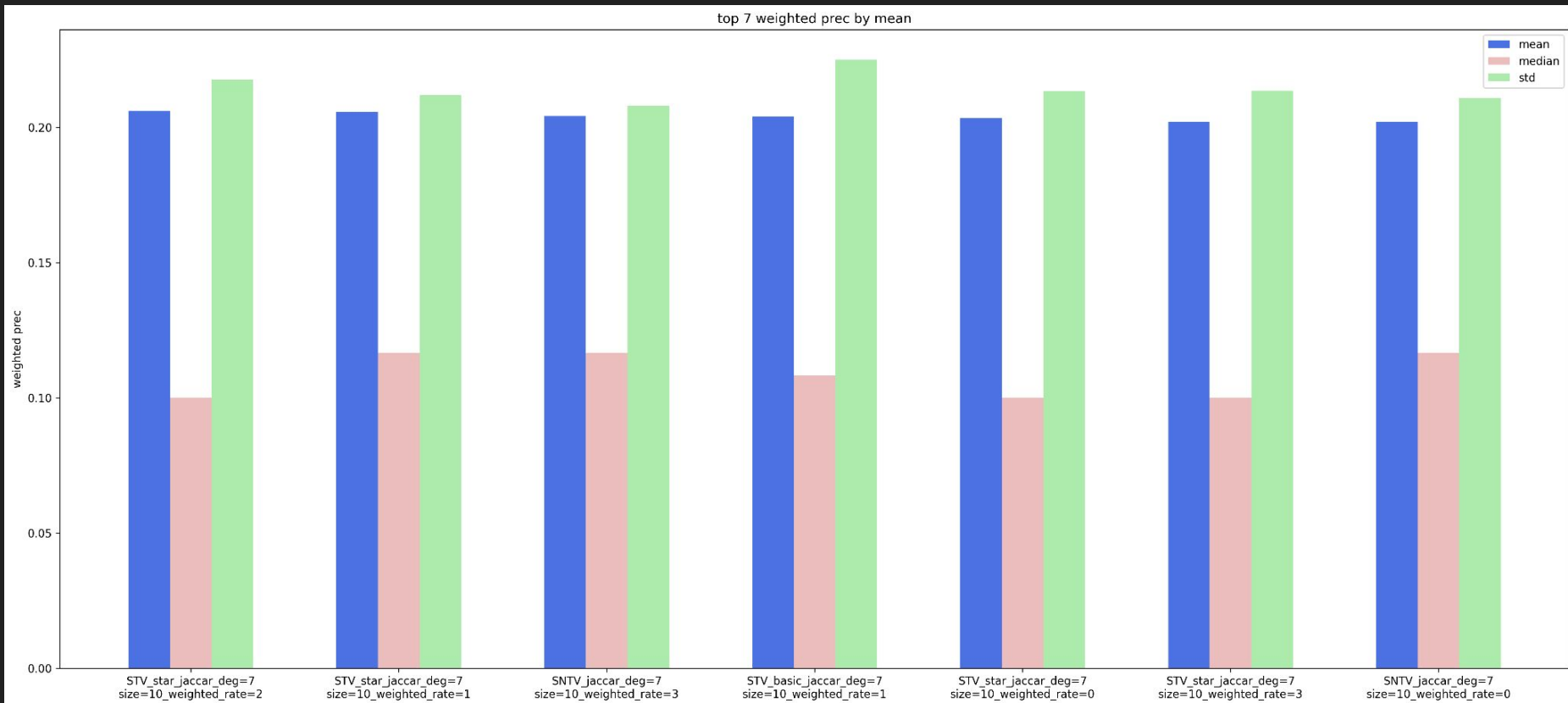


Сравнение статистик

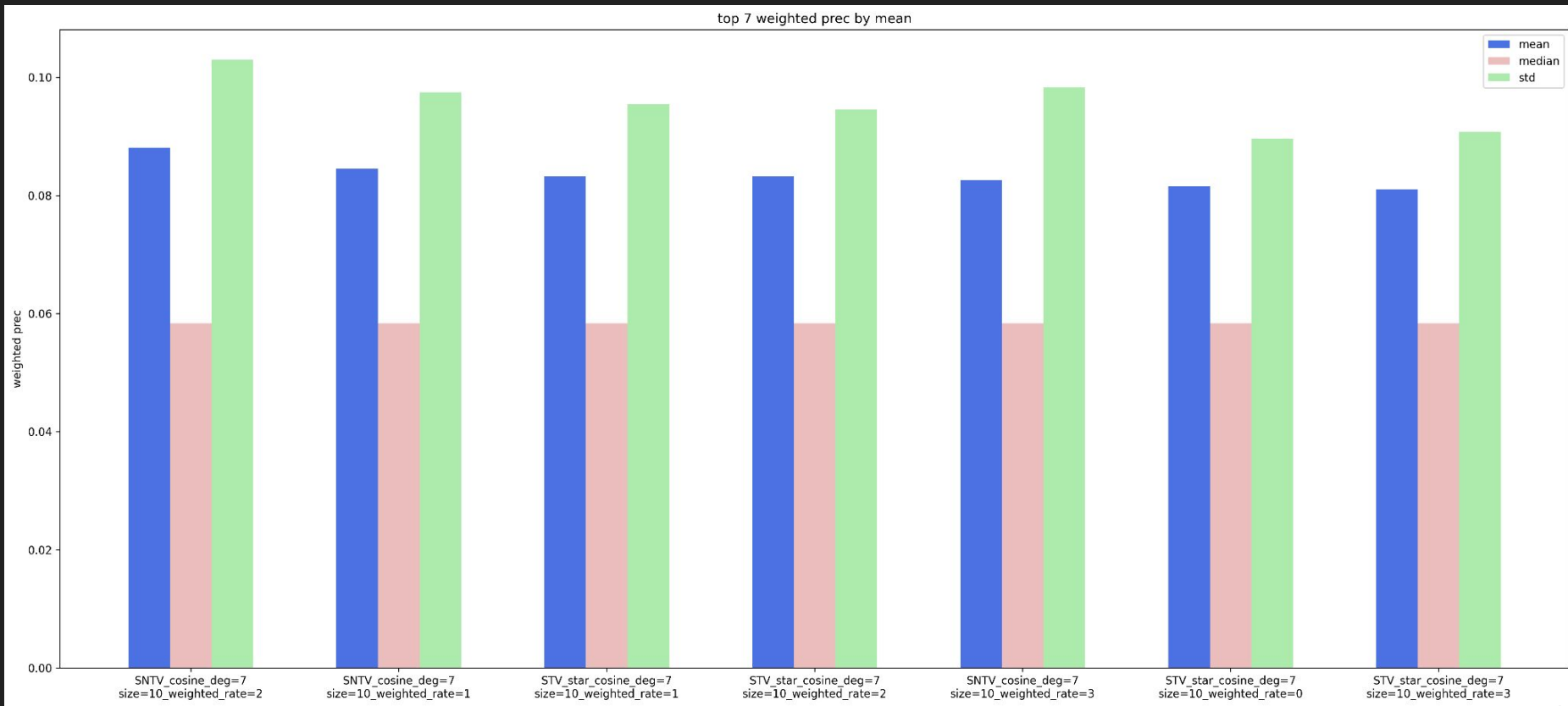
Сравнение рекомендательных алгоритмов — разнообразие



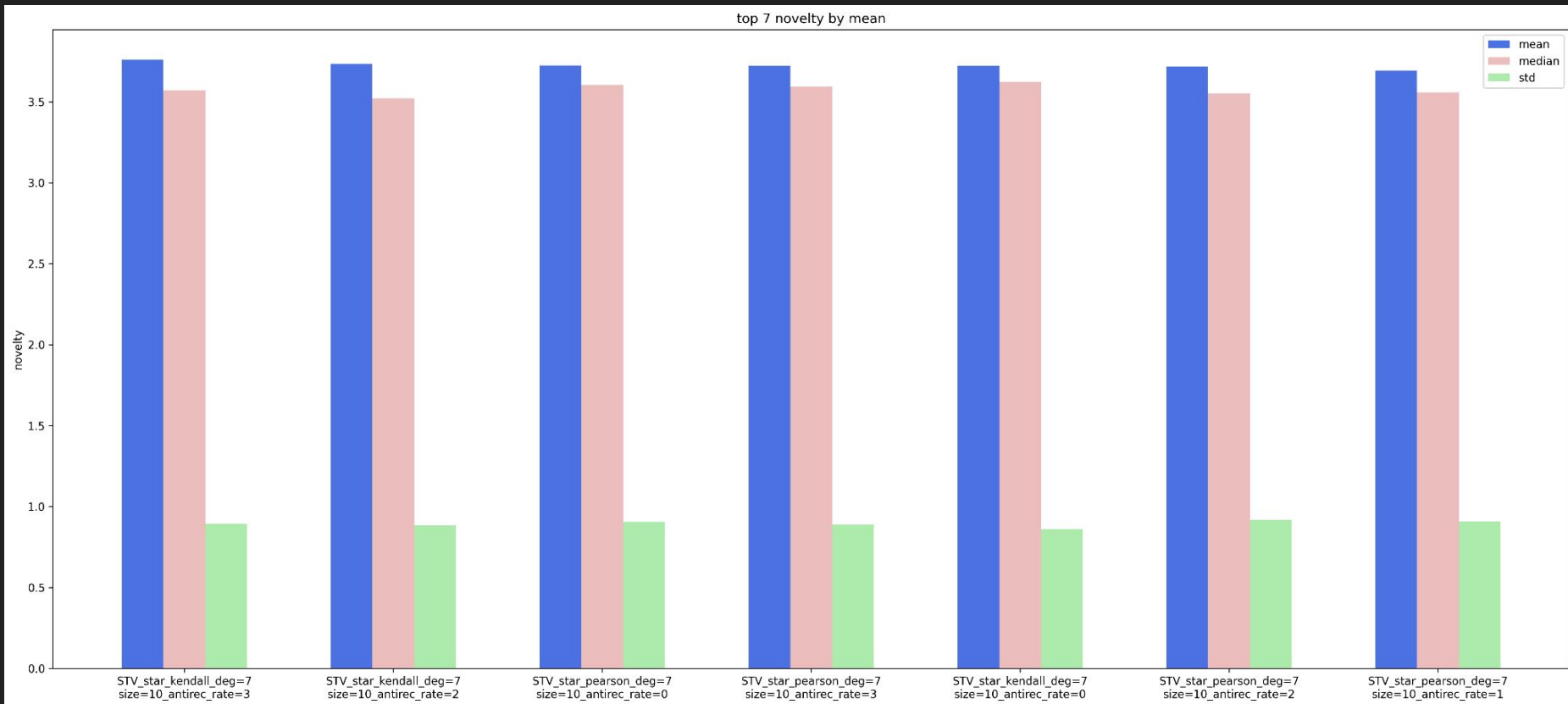
Лучшие конфигурации для плотной матрицы



Лучшие конфигурации для разреженной матрицы



Лучшие конфигурации для разреженной матрицы



Ссылка на гитхаб

Использованный для экспериментов код (github)



Формулы метрик

Метрики индивидуальных рекомендаций (ранжирование)

- $\Theta^u = \{i \in \mathcal{I} \mid \mathcal{R}_{ui} = 0, \mathcal{R}_{ui}^* > 0\}$
- $\text{Precision@k} = \frac{|T^u|}{k}, T^u = \mathcal{J}^u \cap \Theta^u$
- $\text{Recall@k} = \frac{|T^u|}{|\Theta^u|}$
- $\text{nDCG@k} = \frac{DCG@k}{IDCG@k} \qquad DCG@k = \sum_{i=1}^k \frac{2^{rel_i} - 1}{\log_2(i+1)}$

Метрики индивидуальных рекомендаций (разнообразие и новизна)

- $MIUF@k = -\frac{1}{k} \sum_{j \in \mathcal{J}^u} \log_2 \frac{\Pi(j)}{N}, \Pi(j) = \sum_{u=1}^N \mathbb{I}(\mathcal{R}_{uj} > 0)$
- $Serendipity@k = \frac{1}{k} \sum_{i=1}^k \max(p(i) - p_u(i), 0) * rel(i)$

Разнообразие и новизна

$$Serendipity@k = \frac{1}{k} \sum_{i=1}^k \max(p(i) - pu(i), 0) * rel(i)$$

$p(i) = \frac{(n_items - i)}{n_items}$ — вероятность рекомендовать объект с рангом i данному пользователю

$pu(i) = \frac{(n_items - popularity_i)}{n_items}$ — вероятность рекомендовать объект с рангом i любому пользователю

rel_i — бинарная релевантность объекта с рангом i

n_items — число объектов во всем каталоге

$popularity_i$ — ранг популярности объекта с рангом i в выдаче

Примеры добавления пользователя

Чувствительность к добавлению пользователя

- Косинусное расстояние не чувствительно к добавлению пользователя;
- Все корреляционные расстояния чувствительны к добавлению пользователя.

$$r_i = \begin{pmatrix} 0 \\ 7 \\ 5 \end{pmatrix}, r_j = \begin{pmatrix} 8 \\ 3 \\ 1 \end{pmatrix}, r'_i = \begin{pmatrix} 0 \\ 7 \\ 5 \\ 0 \end{pmatrix}, r'_j = \begin{pmatrix} 8 \\ 3 \\ 1 \\ 0 \end{pmatrix}$$

$$\rho_p(i, j) = 1 - \frac{-22}{26} \approx \boxed{1.85}, \quad \rho'_p(i, j) = 1 - \frac{-10}{38} \approx \boxed{1.26}$$

Чувствительность к добавлению пользователя

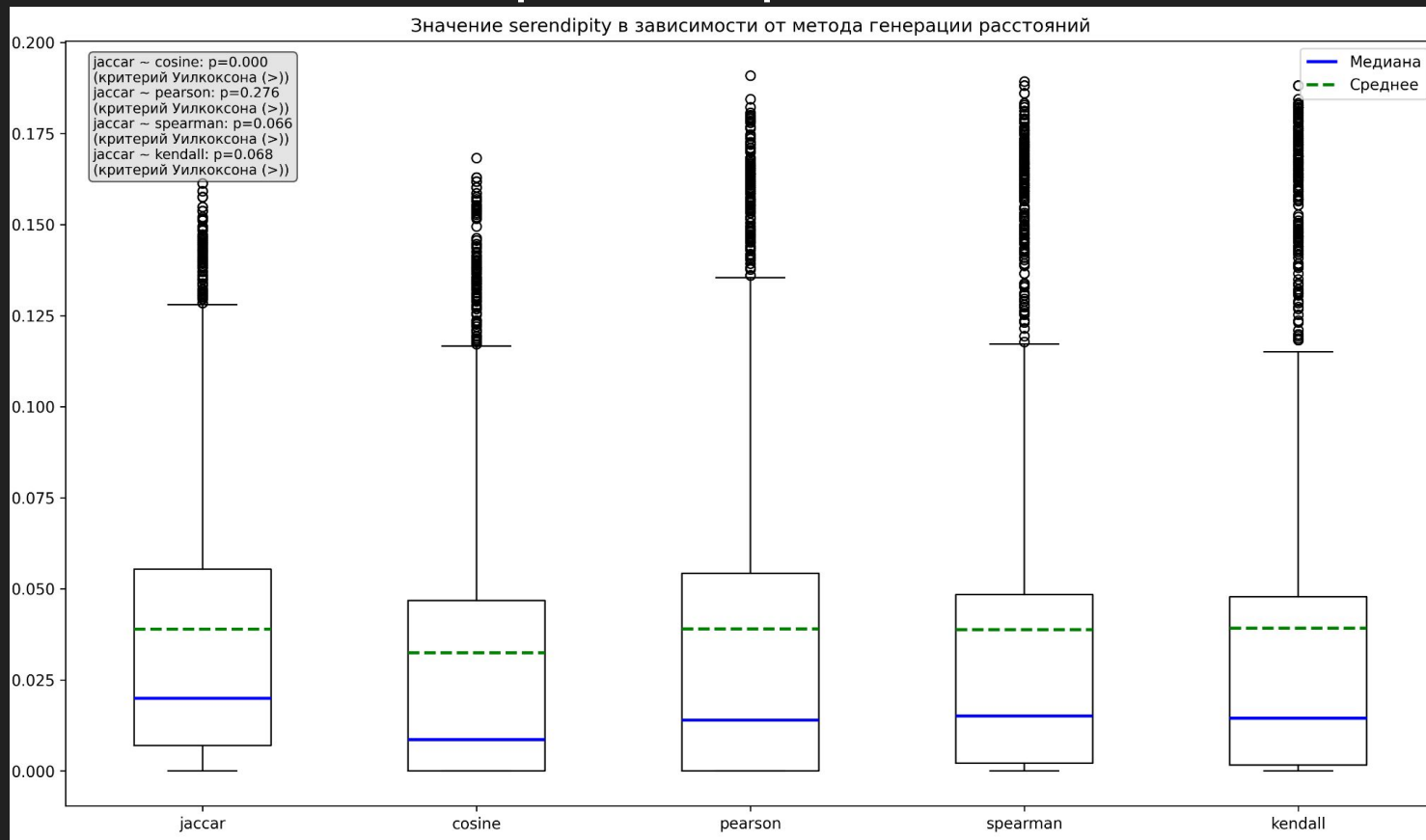
- Косинусное расстояние не чувствительно к добавлению пользователя;
- Все корреляционные расстояния чувствительны к добавлению пользователя.

$$r_i = \begin{pmatrix} 2 \\ 4 \\ 6 \end{pmatrix}, r_j = \begin{pmatrix} 0 \\ 8 \\ 4 \end{pmatrix}, r'_i = \begin{pmatrix} 2 \\ 4 \\ 6 \\ 0 \end{pmatrix}, r'_j = \begin{pmatrix} 0 \\ 8 \\ 4 \\ 0 \end{pmatrix}$$

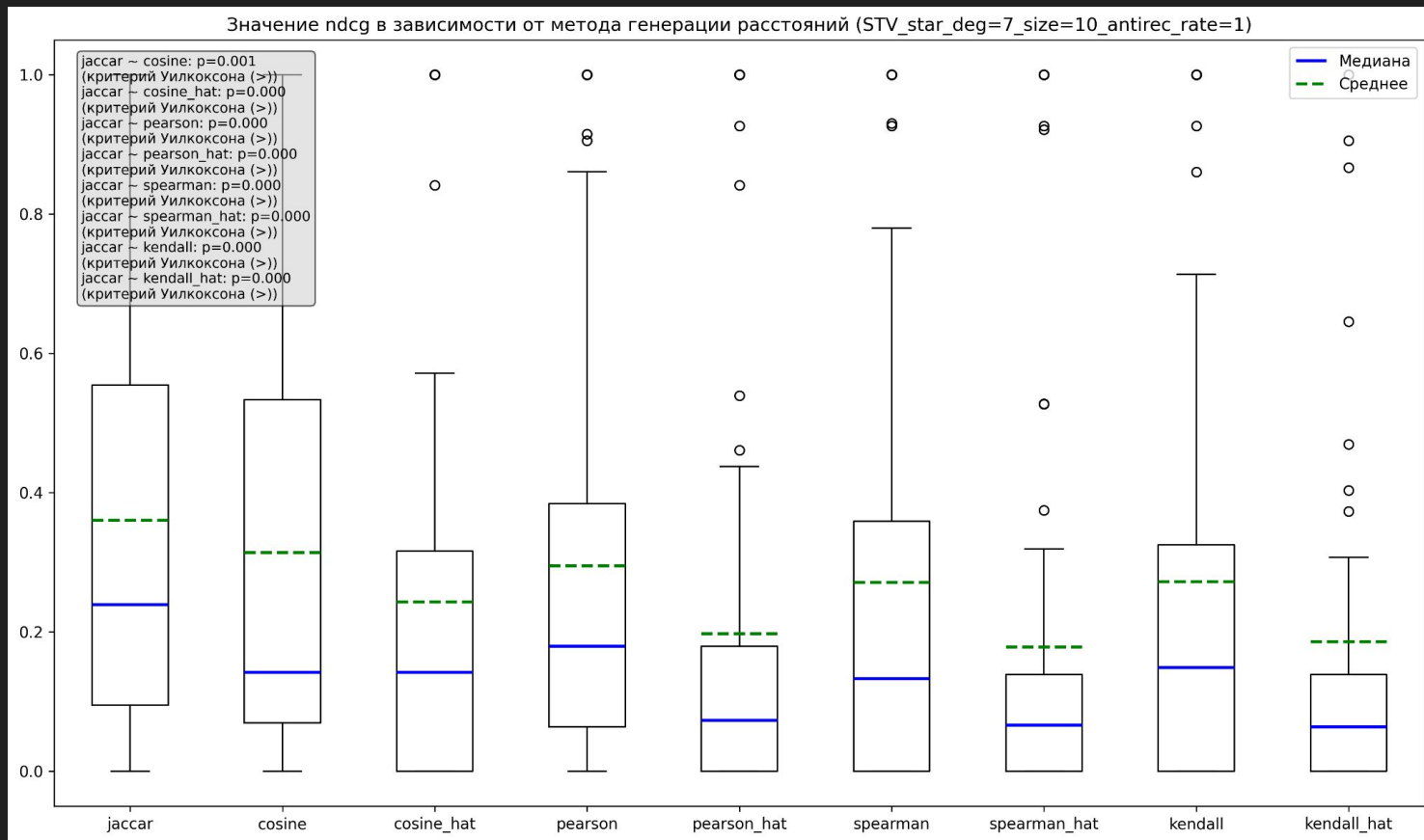
$$\rho_p(i, j) = 1 - \frac{8}{16} = \boxed{0.5}, \quad \rho'_p(i, j) = 1 - \frac{11}{4\sqrt{55}} \approx \boxed{0.63}$$

Невошедшие графики

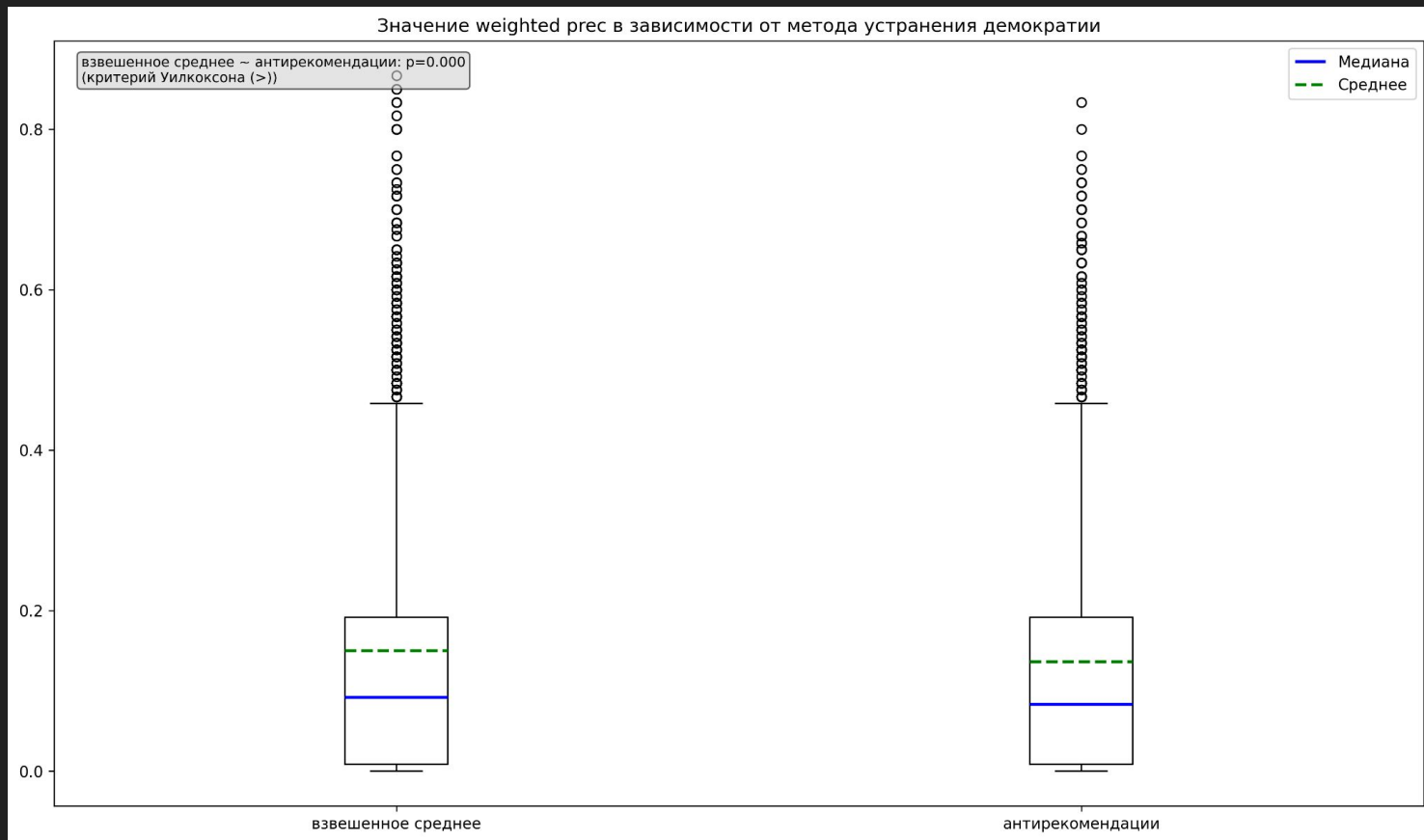
Влияние метода вычисления расстояний на разнообразие



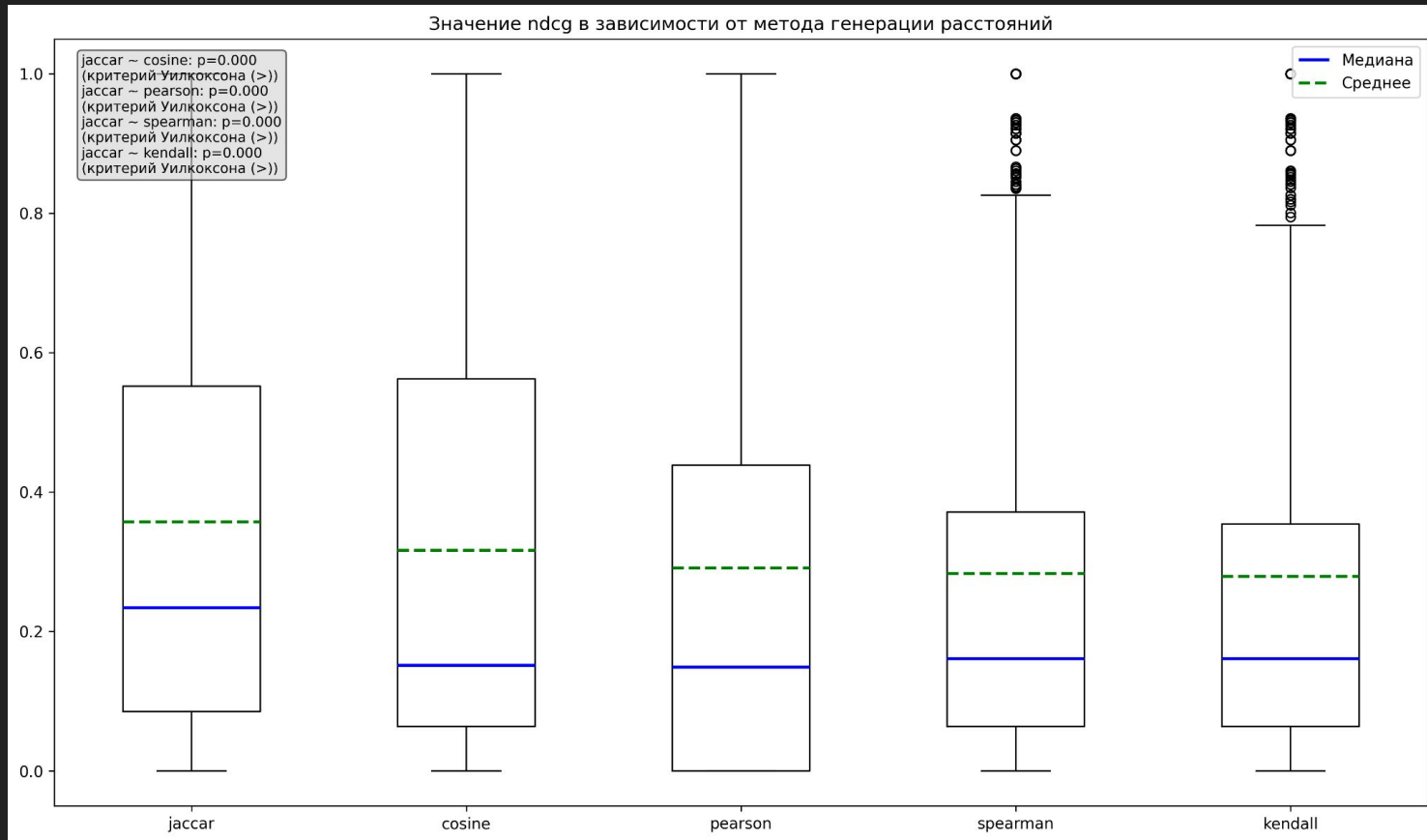
Влияние метода вычисления расстояний на качество ранжирования



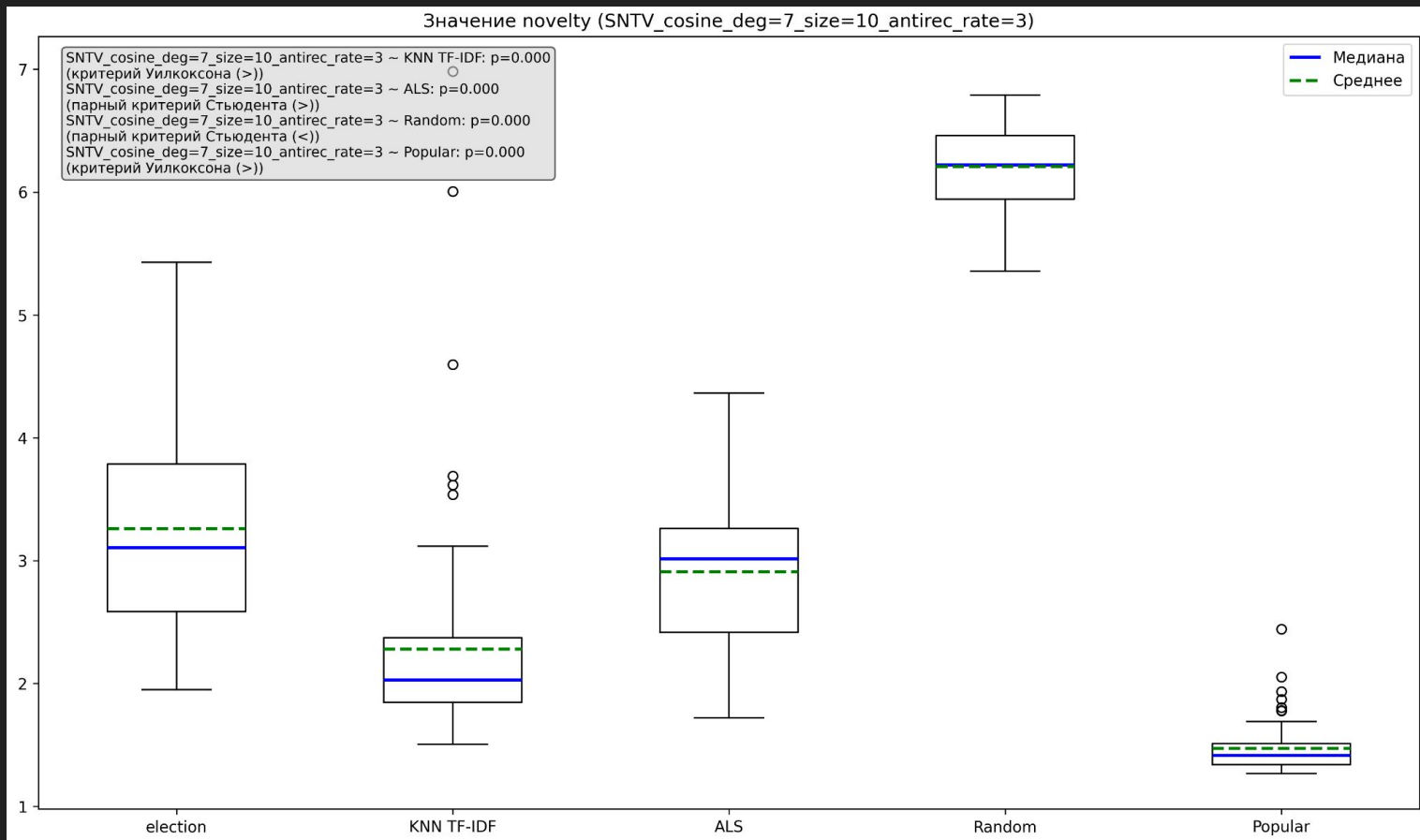
Влияние метода учёта оценок избирателей на точность



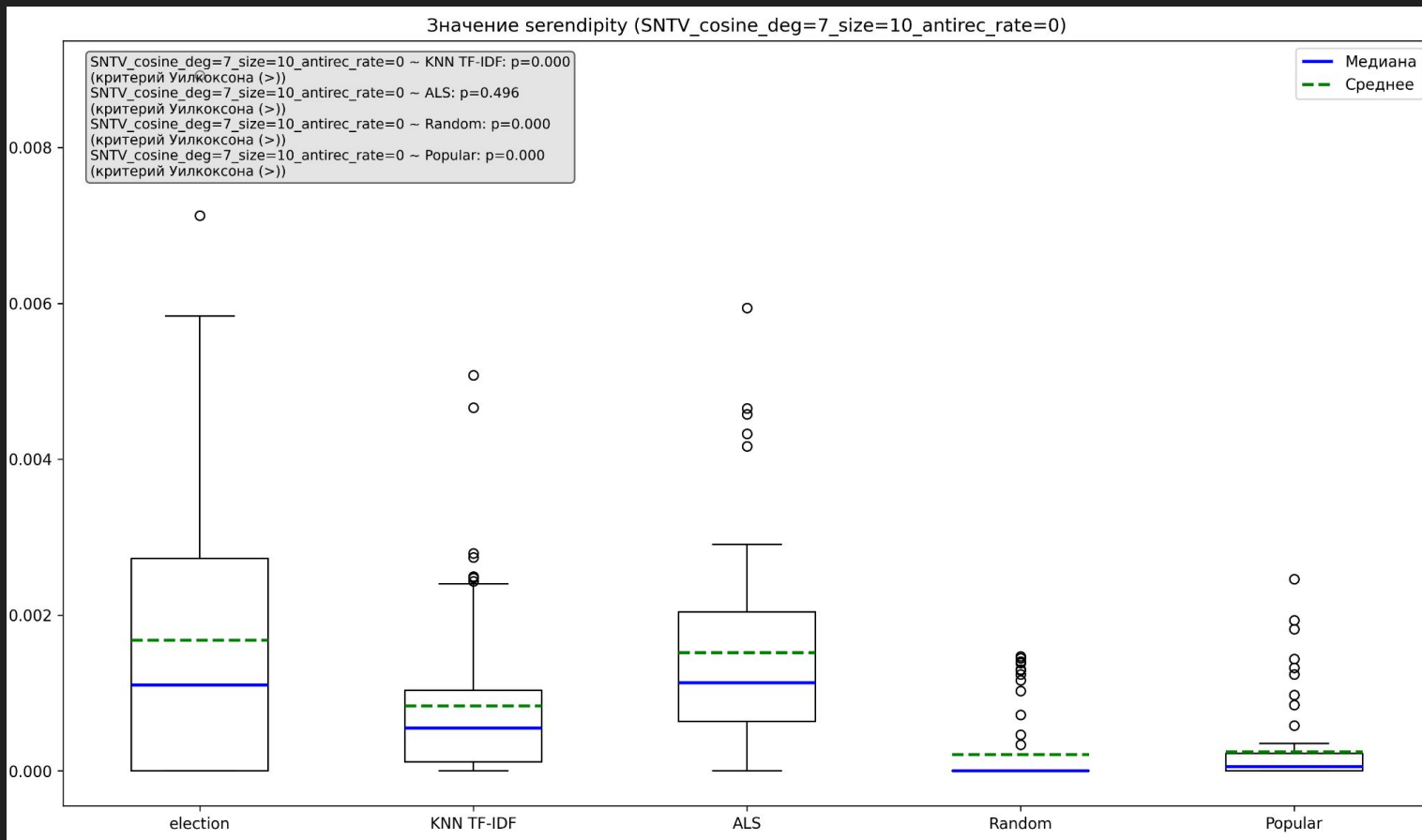
Влияние метода вычисления расстояний на качество ранжирования



Сравнение рекомендательных алгоритмов — разнообразие



Сравнение рекомендательных алгоритмов — разнообразие



Повышение разнообразия

Гипотеза разнообразия

$$c_0 = k \int_0^{+\infty} \int_{-\phi}^{\phi} r \cos(\theta) d\mathbb{P}_{\rho}(r) \frac{1}{2\pi} d\theta$$

Гипотеза разнообразия

$$R = \frac{k}{\sqrt{2\pi}} \sin\left(\frac{\pi}{k}\right)$$

(для нормального распределения)

R — радиус оптимальных кандидатов

k — количество избираемых кандидатов



Гипотеза разнообразия

Избираем больше
кандидатов, чем нужно



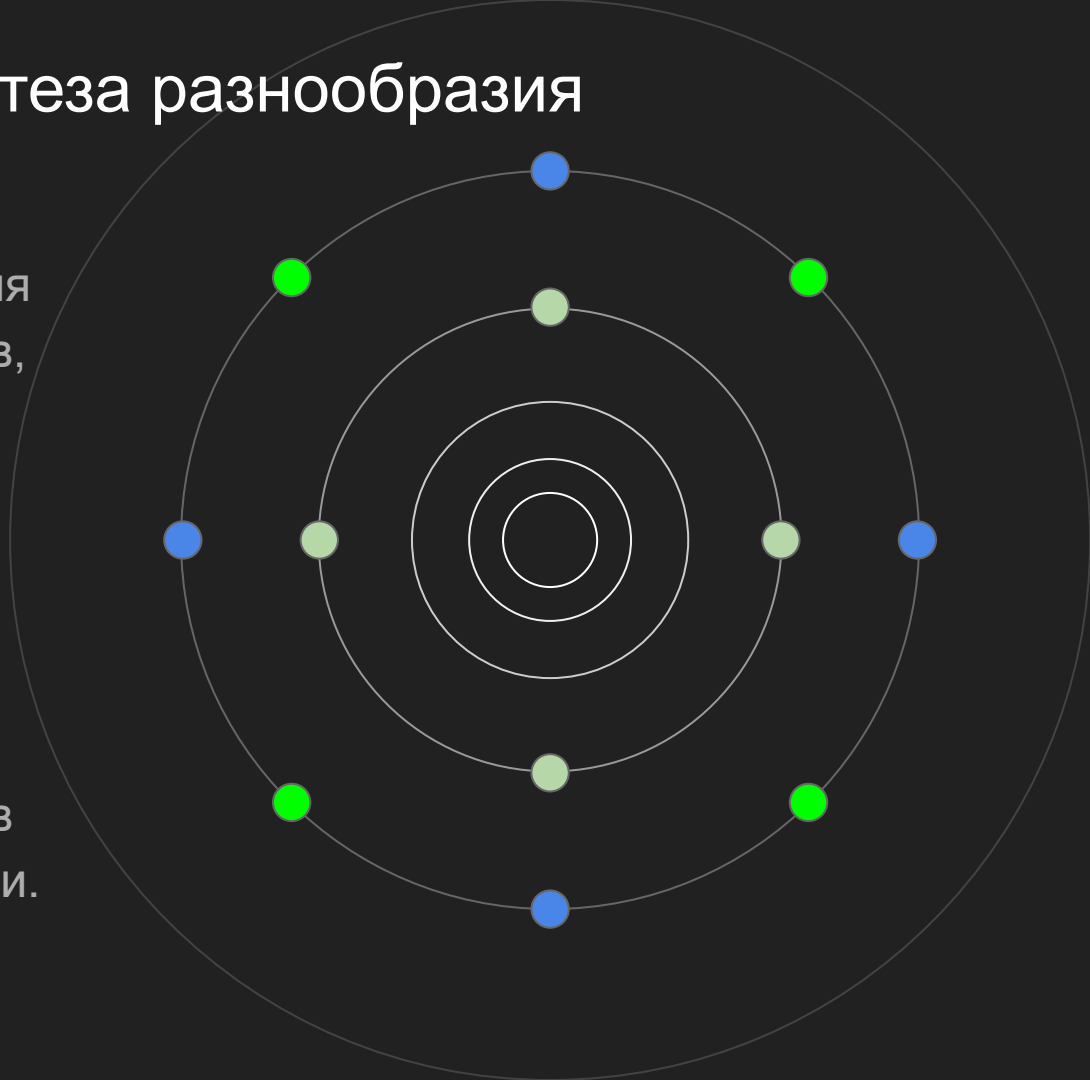
Гипотеза разнообразия

Из них уже выбираем
нужное количество

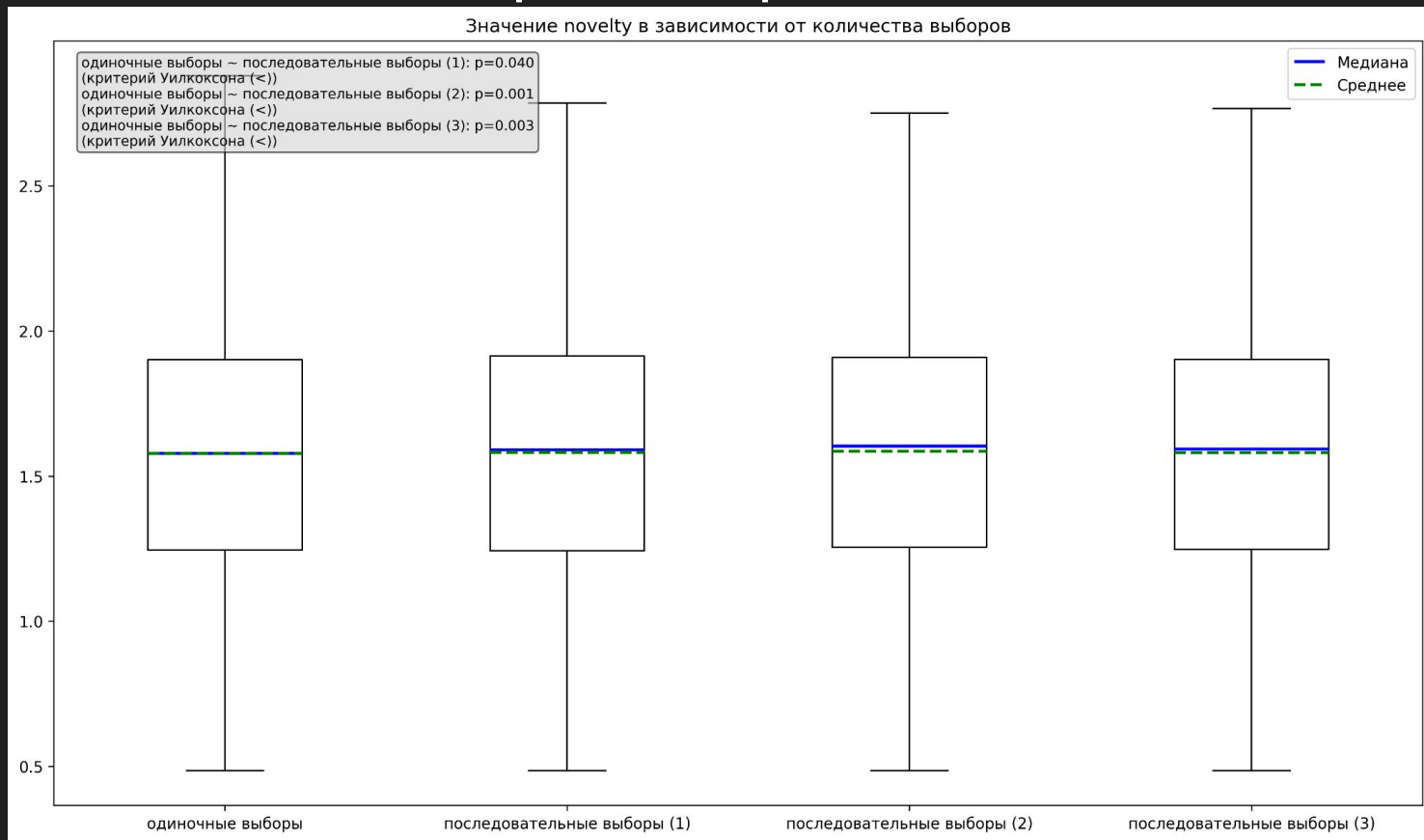


Гипотеза разнообразия

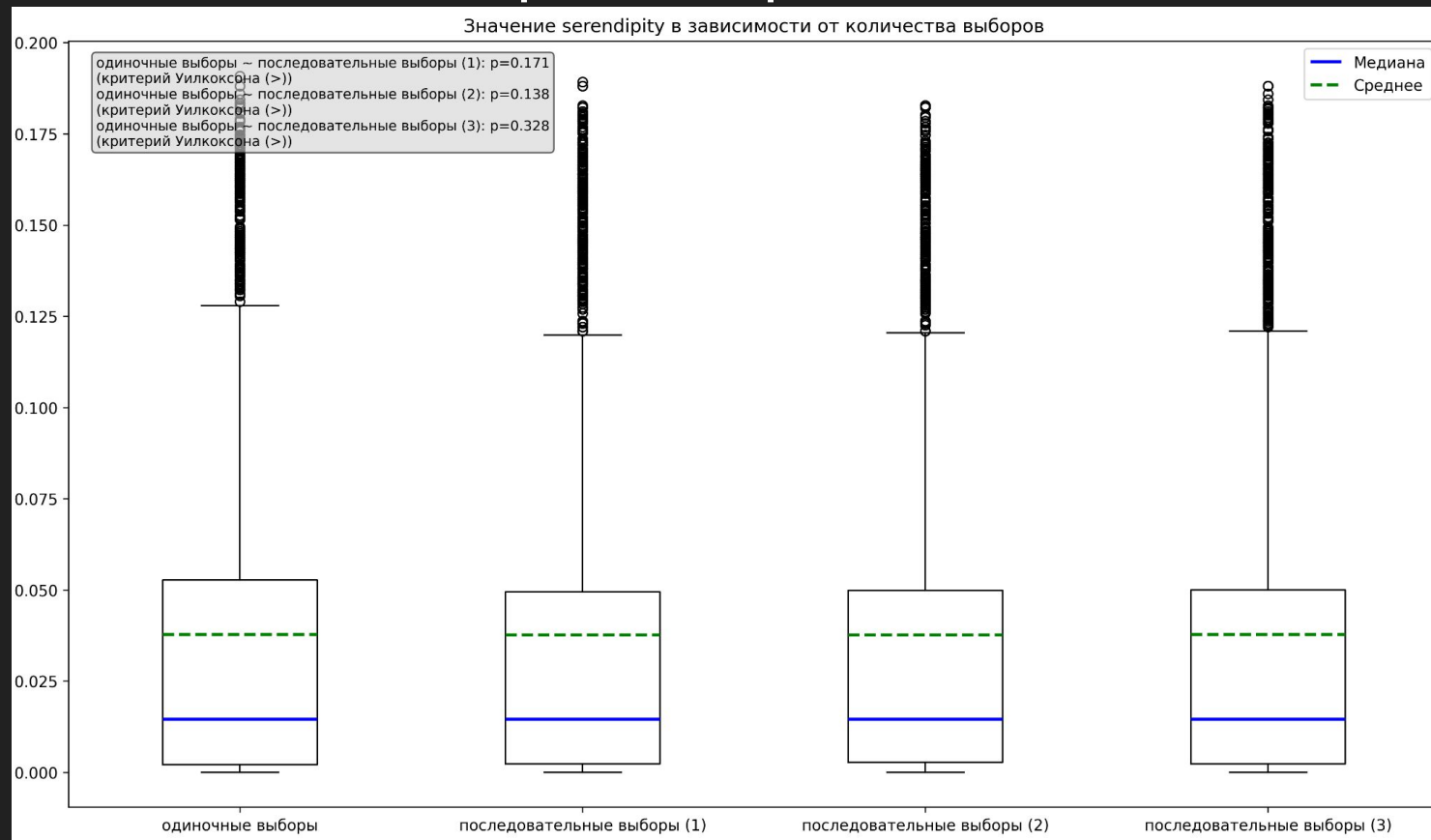
- Ожидаем, что метрики Serendipity и MIUF выше для последовательных выборов, чем для одиночных;
- Для пользователей с маленькой дисперсией оценённых фильмов ожидаем ухудшения релевантности последовательных выборов по сравнению с одиночными.



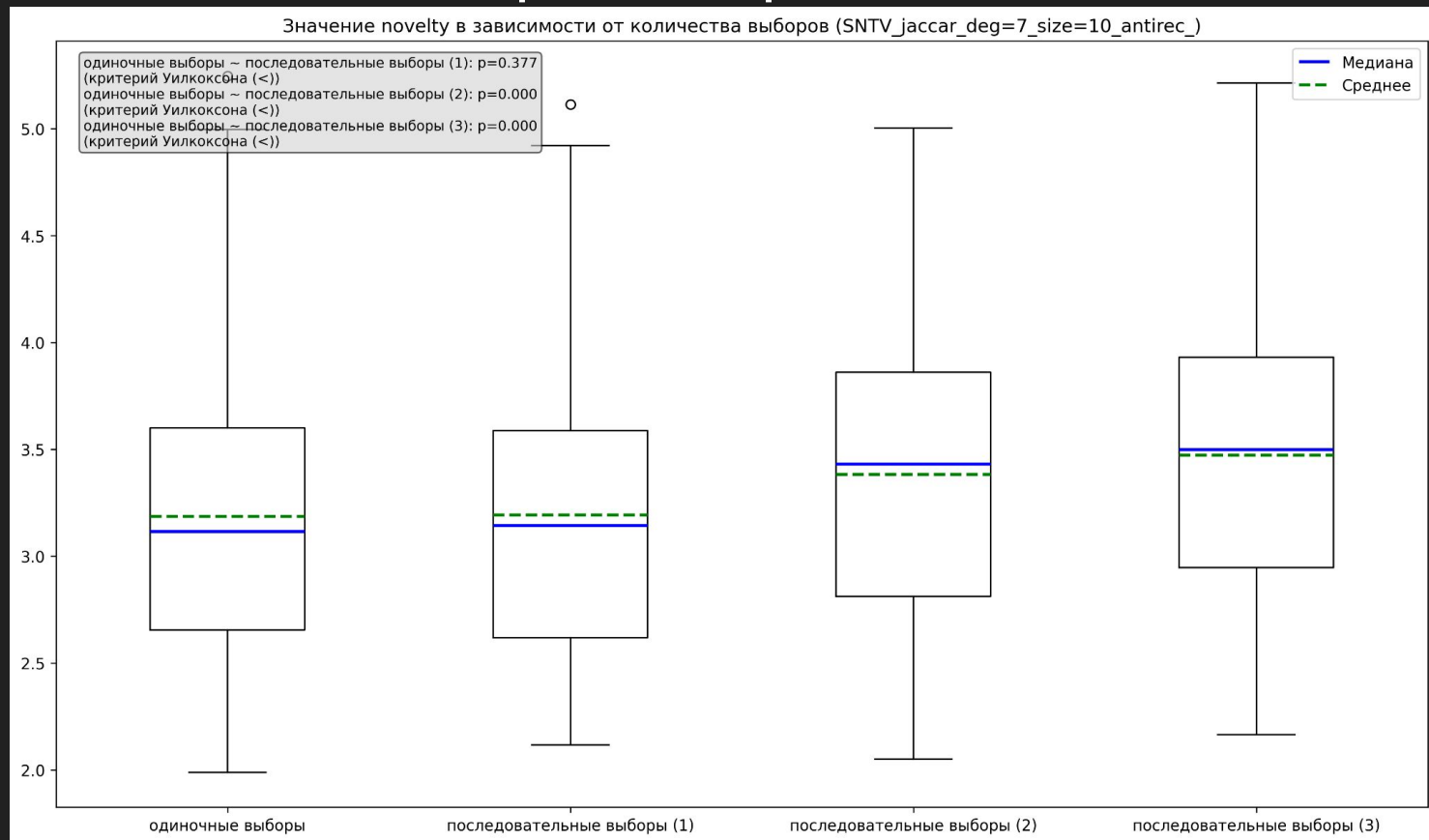
Влияние метода учёта оценок избирателей на разнообразие



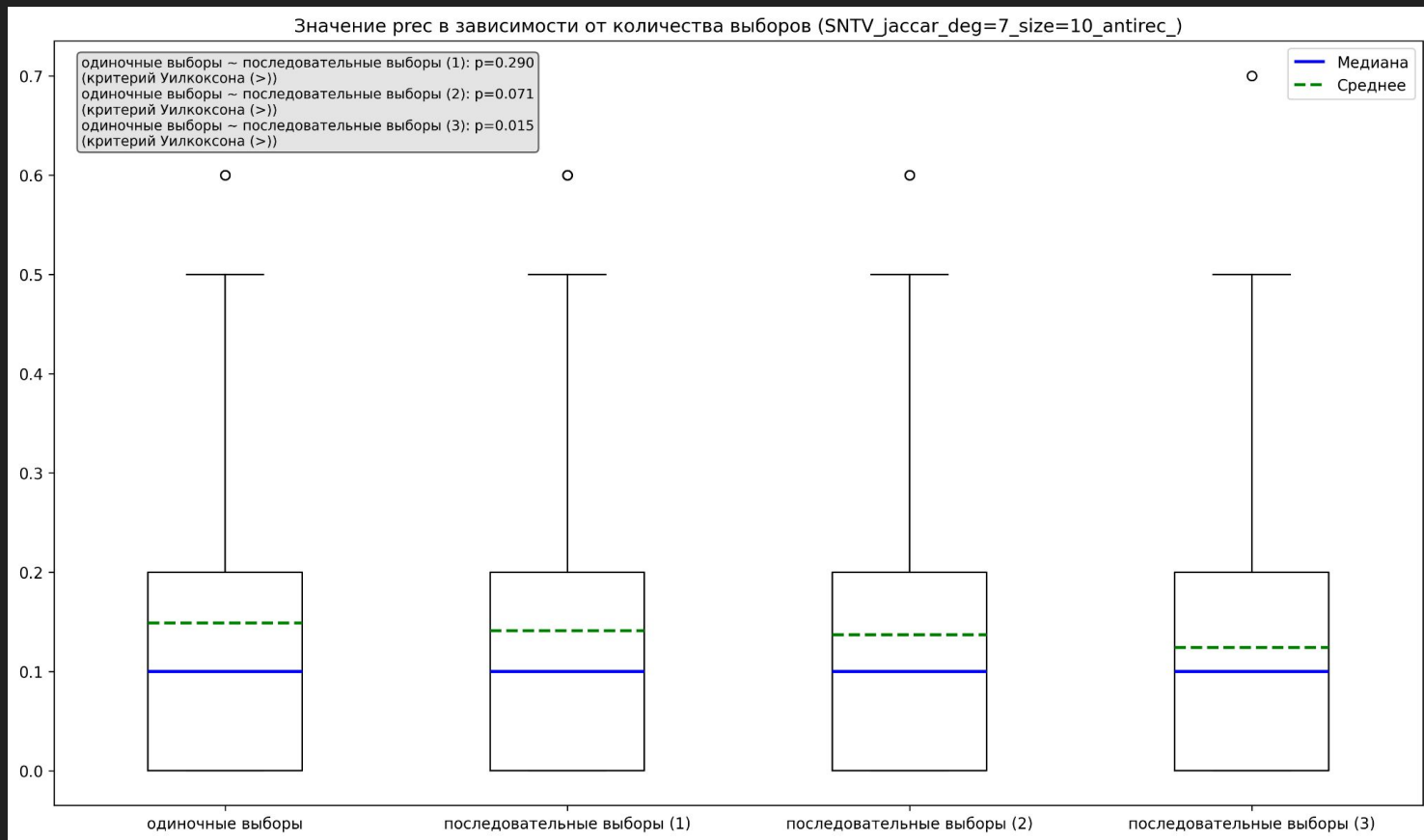
Влияние метода учёта оценок избирателей на разнообразие



Влияние метода учёта оценок избирателей на разнообразие



Влияние метода учёта оценок избирателей на точность



Примеры задачи составления фильмотеки

Составление фильмотеки

Пример: Выбор некоторого количества фильмов для полёта в самолёте

Пример: Выбор ограниченного числа локаций для постройки некоторых зданий среди возможных

Избиратель оценивает комитет по единственному его члену