

СТАТ МОДУЛЬ 2.0

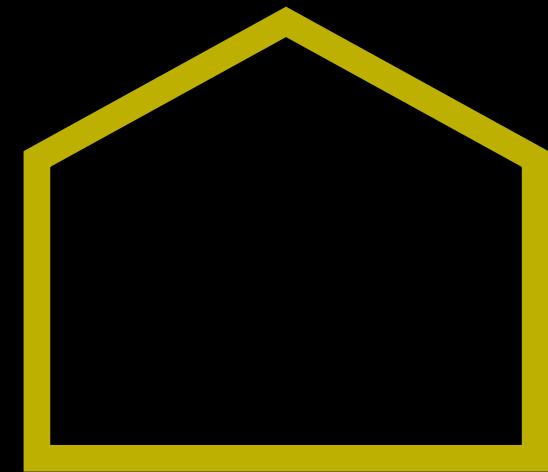
статистическая обработка выдачи
дифференциального корпуса

терзи владислав, ABBYY lab 2020

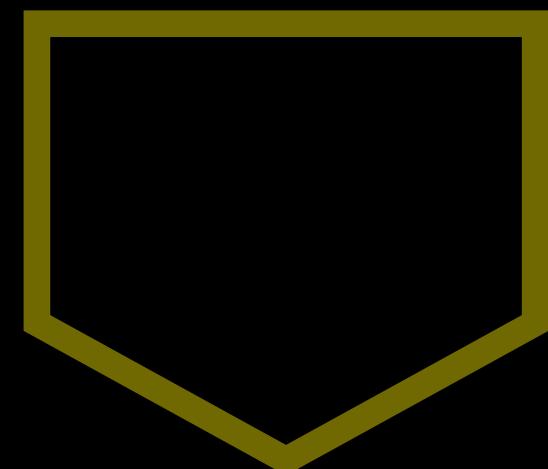
В ТЕОРИИ

1. диплом **куратова** “статистика в лингвистике”
2. диплом **гежеса** “арі генерации диф. частотников”
3. статья **шарова** “know the corpus!”
4. диплом **шлыкова** “сравнение диф. частотников”
5. диалог 2020 **деткова** “семантические скетчи”
6. книга **лагутина** “наглядные статьи”

пословный



1. IPM

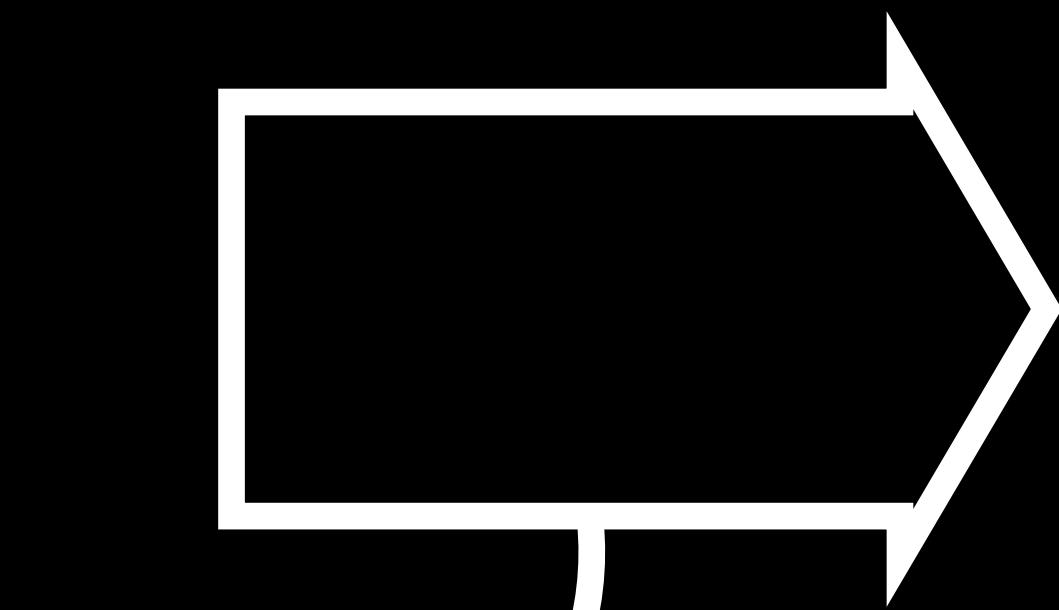


подокументный

МАГИЧЕСКАЯ КУРАТОВА

ЗАДАЧА:

СЛОВО



IPM

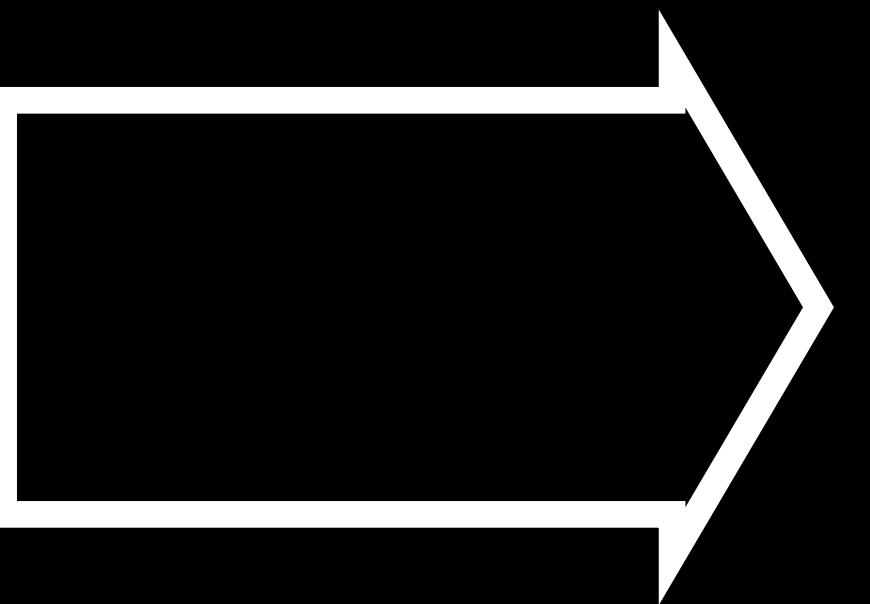
тут
есть
проблема

19,8 млрд

слов в гикря

ВЫХОД:

строим доверительный интервал с помощью
минимально возможного подкорпуса

100  [90, 110]

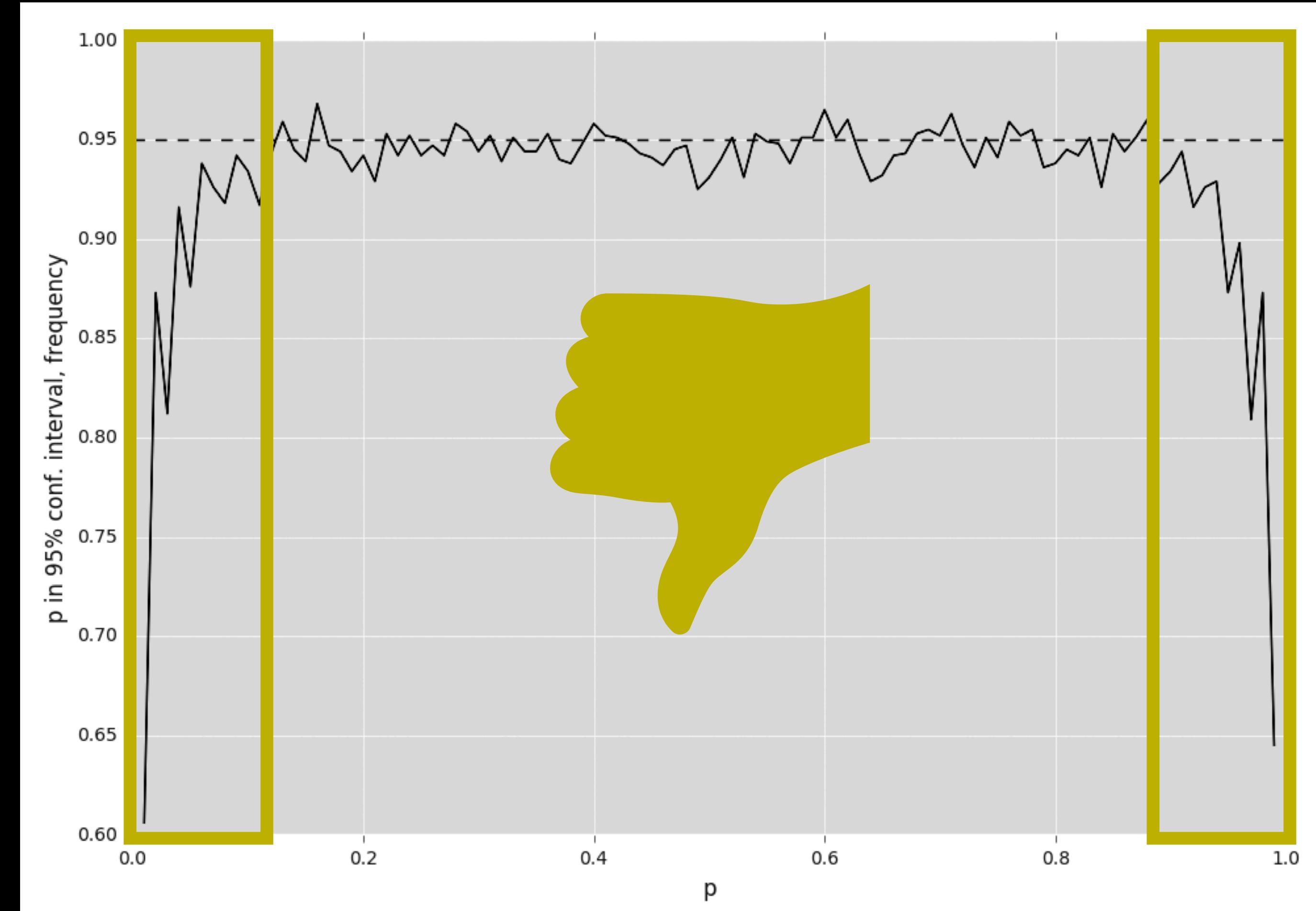
thumb up icon 95%

ПРОБЛЕМЫ

1.

моделирование
вероятности
слова

=
схема бернулли



зависимость частоты попадания истинной вероятности p
95 % доверительный интервал от частоты p

2.

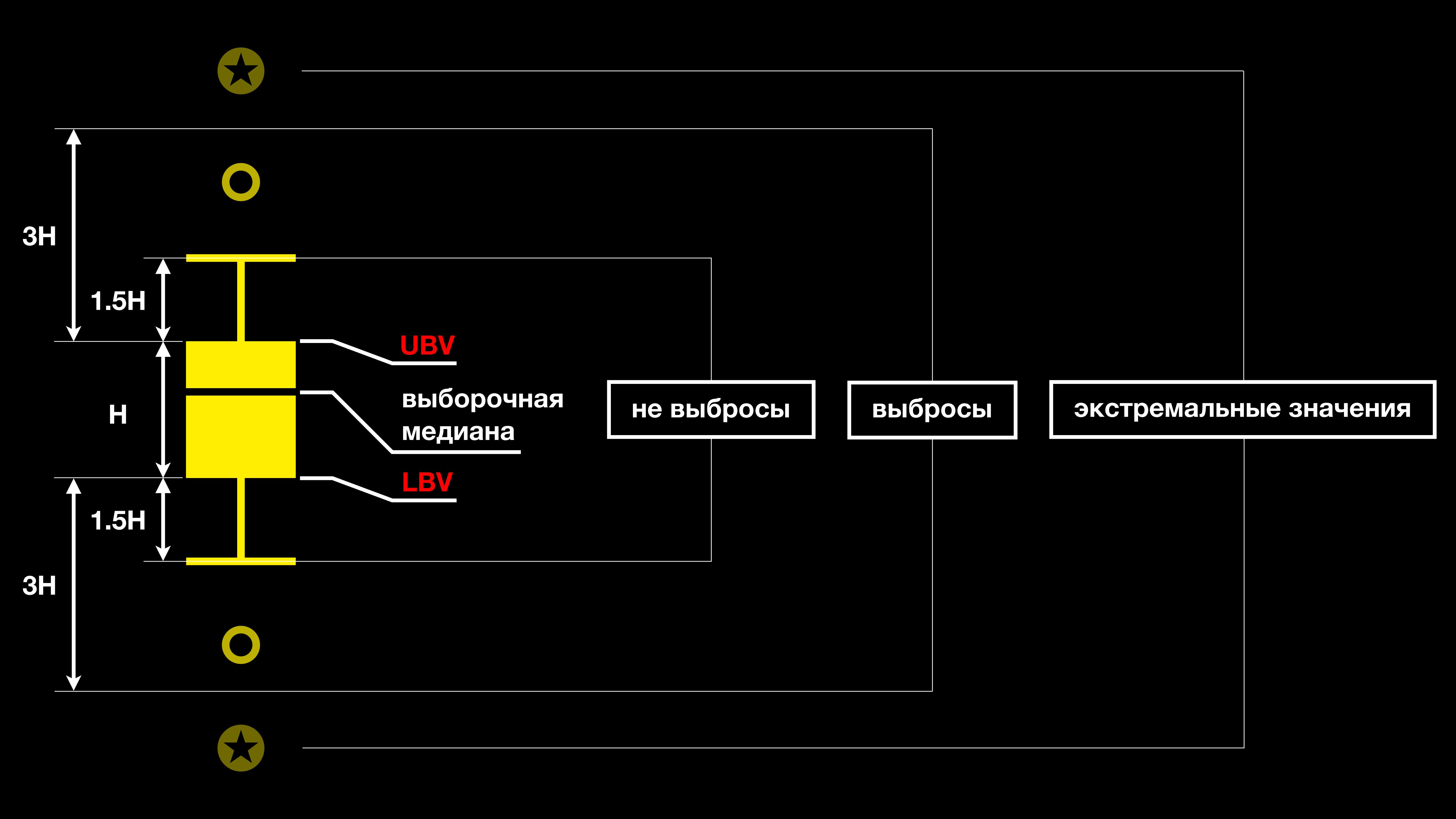
**два появления одного и того же слова в
одном документе – два зависимых события**

МЕТОДИКА:

$p \rightarrow \bar{p} = \sum_i p_i \omega_i$, где p_i – частота в i -м документе, ω_i – вес текста

$$Var \rightarrow \sqrt{\frac{n}{n-1} \sum_i \omega_i (p_i - \bar{p})^2}$$

ВЫБРОСЫ



МЕДИАНА MED^+

1. составляем вариационные ряды:

$p_{(1)} \leq \dots \leq p_{(n)}$ и соответствующий ему $\omega_{(1)}, \dots, \omega_{(n)}$

2. $m = \min_{i \leq n} \left\{ \sum_{k=1}^i \omega_{(k)} \geq \frac{1}{2} \right\}, y = \sum_{i=1}^{m-1} \omega_{(i)}$

3. $MED^+ = p_{(m-1)} + \frac{1}{\omega_m} \left(p_{(m)} - p_{(m-1)} \right) \cdot \left(\frac{1}{2} - y \right)$

ВЕРХНЯЯ КВАРТИЛЬ UBV^+

1. составляем вариационные ряды:

$p_{(1)} \leq \dots \leq p_{(n)}$ и соответствующий ему $\omega_{(1)}, \dots, \omega_{(n)}$

2. $m = \min_{i \leq n} \left\{ \sum_{k=1}^i \omega_{(k)} \geq \frac{3}{4} \right\}, y = \sum_{i=1}^{m-1} \omega_{(i)}$

3. $UBV^+ = p_{(m-1)} + \frac{1}{\omega_m} \left(p_{(m)} - p_{(m-1)} \right) \cdot \left(\frac{3}{4} - y \right)$

ПРЕДПОЛОЖЕНИЕ:

$\frac{H}{2} \approx UBV^+ - MED^+ \Rightarrow p_i$ является выбросом если оно оказалось больше, чем
 $UBV^+ + 6(UVB^+ - MED^+)$

БАЗОВАЯ МЕТОДИКА

1. случайно набираем тексты из корпуса, пока не наберем m_0 текстов, в которых встретилось слово (пусть для этого понадобилось n_0 текстов — наша выборка)
2. очищаем выборку от выбросов (по предположению), n'_0 — очищенная от выбросов выборка

БАЗОВАЯ МЕТОДИКА

3. вычисляем грубые оценки \bar{p}_0 и S_0 , по ним записываем следующие соотношения:

$$\delta \approx \delta_0 \approx \frac{S_0}{\sqrt{n'_0}} \Rightarrow n = \left(\frac{S_0}{\delta_0} \right)^2 - \text{реальный размер выборки}$$

4. берем выборку размера n , очищаем ее от выбросов – n' , находим более точные значения \bar{p} и S , строим 95% доверительный интервал:

$$p \in \left(\bar{p} - 2 \frac{S}{\sqrt{n'}}, \bar{p} + 2 \frac{S}{\sqrt{n'}} \right)$$

МЕТОДИКА СО СТРАТИФИКАЦИЕЙ

1. тексты разбиваются на K страт, каждая имеет долю q_i , $i \in 1, \dots, K$
2. случайно набираем тексты, пока не наберем m_0 таких, что в них встретилось слово (n_0 – наша выборка)
3. очищаем выборку и вычисляем грубые оценки \bar{p}_0 и S_0
4. выборка разделяется на подвыборки по стратам, для каждой считается \bar{p}_{0j} и S_{0j} , если текстов в j -й страте меньше 5 или $S_{0j} > S_0$, тогда

$$S_{0j} := S_0$$

МЕТОДИКА СО СТРАТИФИКАЦИЕЙ

5. вычисляем \tilde{p}' и \tilde{S}' :

$$\tilde{p}' = \sum_{i=1}^K q_i p_i, \quad \tilde{S}' = \sqrt{\sum_{i=1}^K q_i S_i^2}$$

6. как и в прошлой методике вычисляем размер выборки: $n = \left(\frac{\tilde{S}'}{\tilde{\delta}'} \right)^2$

7. пополняем n_0 до n , очищаем от выбросов.

МЕТОДИКА СО СТРАТИФИКАЦИЕЙ

8. делим на страты, считаем более точные \bar{p}_j и S_j

9. пересчитываем уточненные значения для \tilde{p} и \tilde{S}

10. строим 95% доверительный интервал:

$$p \in \left(\tilde{p} - 2 \frac{\tilde{S}}{\sqrt{\tilde{n}}}, \tilde{p} + 2 \frac{\tilde{S}}{\sqrt{\tilde{n}}} \right)$$

**МОЖНО ЕЩЕ НАЙТИ
МОМЕНТ СТАБИЛИЗАЦИИ**

**ЭТО МОЖЕТ ПОМОЖЕТ ДОСТИЧЬ ТОЧНЫХ РЕЗУЛЬТАТОВ,
НО ПРИ ЭТОМ СИЛЬНО БЫСТРЕЕ**

ПОДЫТОГ:

на популярных словах **хорошо работает**, на не очень популярных – смысла нет проверять (статистической обработке они все равно не поддаются)

БУЛЬБАЗАВРСКАЯ ГЕЖЕСА

ЧАСТОТНЫЙ СЛОВАРЬ

$$(w_i, f_i)_{i=1}^n$$

уникальное слово

лемма

морфема

частота

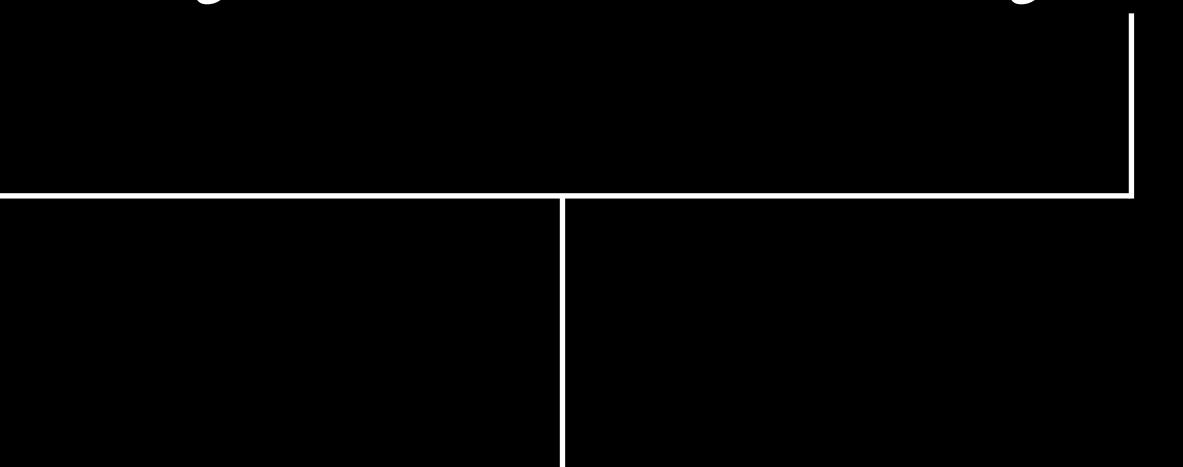
доля вхождений слова в корпус

доля документов, содержащих слово

доля авторов, использующих слово

ДИФФЕРЕНЦИАЛЬНЫЙ ЧАСТОТНИК

$$(w_i, a_i, \dots, z_i, f_i)_{i=1}^n$$



соответствующие
значения набора
параметров

РАЗНОСТНЫЙ ЧАСТОТНИК

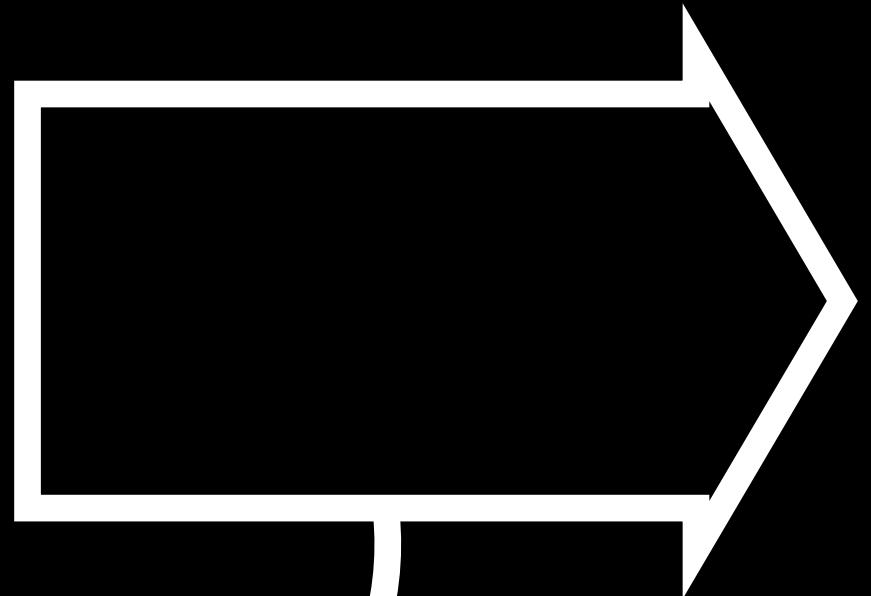
$$(w_i, \underbrace{f_i^1, f_i^2}_{\text{частоты в первом и втором словарях соответственно}}, d_i)_{i=1}^n$$

частоты в первом
и втором словарях
соответственно

разностная мера

ЗАДАЧА 1:

запрос



набор
параметров и
значений

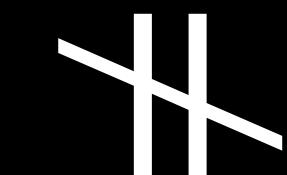
дифф
частотник

n самых
“популярных”
слов

ЗАДАЧА 2:

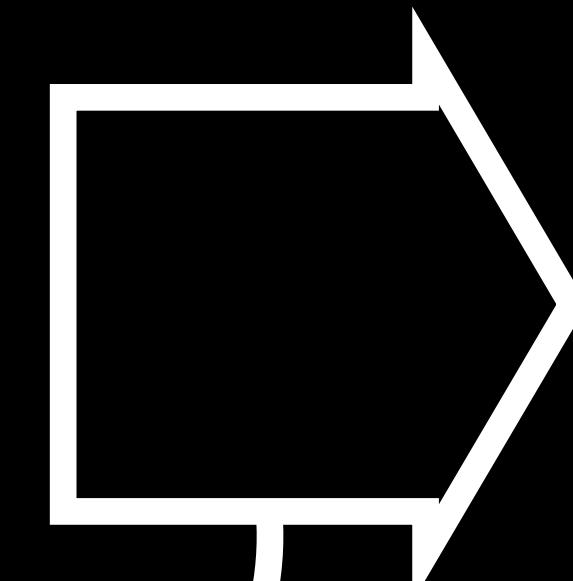
подкорпус 1

запрос 1



подкорпус 2

запрос 2



разностный

частотник

n самых
“различающихся”
слов

ПРОМЕЖУТОЧНЫЙ СЛОВАРЬ

$$(w_i, p_1, v_1, \dots, p_n, v_n, f)_{i=1}^n$$

p_i – имя параметра

v_i – его значение

количество вхождений

НЕМНОГО ПРО MAP-REDUCE

MAP

```
#include "mapreduce/mapreduce.h"

// User's map function
class WordCounter : public Mapper {
public:
    virtual void Map(const MapInput& input) {
        const string& text = input.value();
        const int n = text.size();
        for (int i = 0; i < n; ) {
            // Skip past leading whitespace
            while ((i < n) && isspace(text[i]))
                i++;
            // Find word end
            int start = i;
            while ((i < n) && !isspace(text[i]))
                i++;
            if (start < i)
                Emit(text.substr(start,i-start), "1");
        }
    }
};

REGISTER_MAPPER(WordCounter);
```

REDUCE

```
// User's reduce function
class Adder : public Reducer {
    virtual void Reduce(ReduceInput* input) {
        // Iterate over all entries with the
        // same key and add the values
        int64 value = 0;
        while (!input->done()) {
            value += StringToInt(input->value());
            input->NextValue();
        }
        // Emit sum for input->key()
        Emit(IntToString(value));
    }
};

REGISTER_REDUCER(Adder);
```

ПО ФАКТУ

MAP

$(w_1, 1)$

$(w_1, 1)$

\vdots

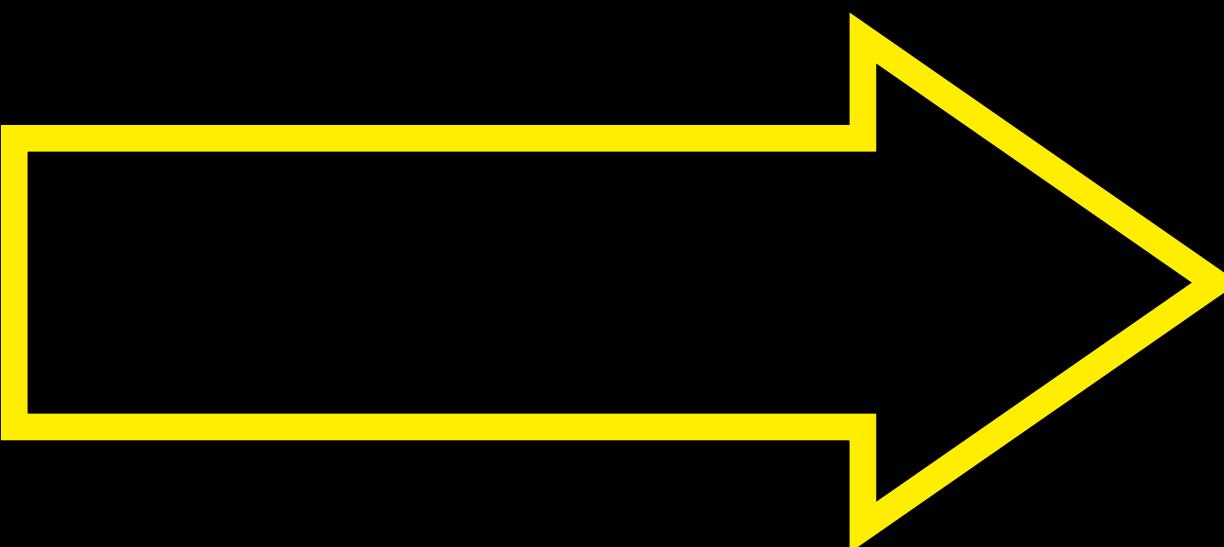
$(w_2, 1)$

\vdots

$(w_n, 1)$

\vdots

$(w_n, 1)$

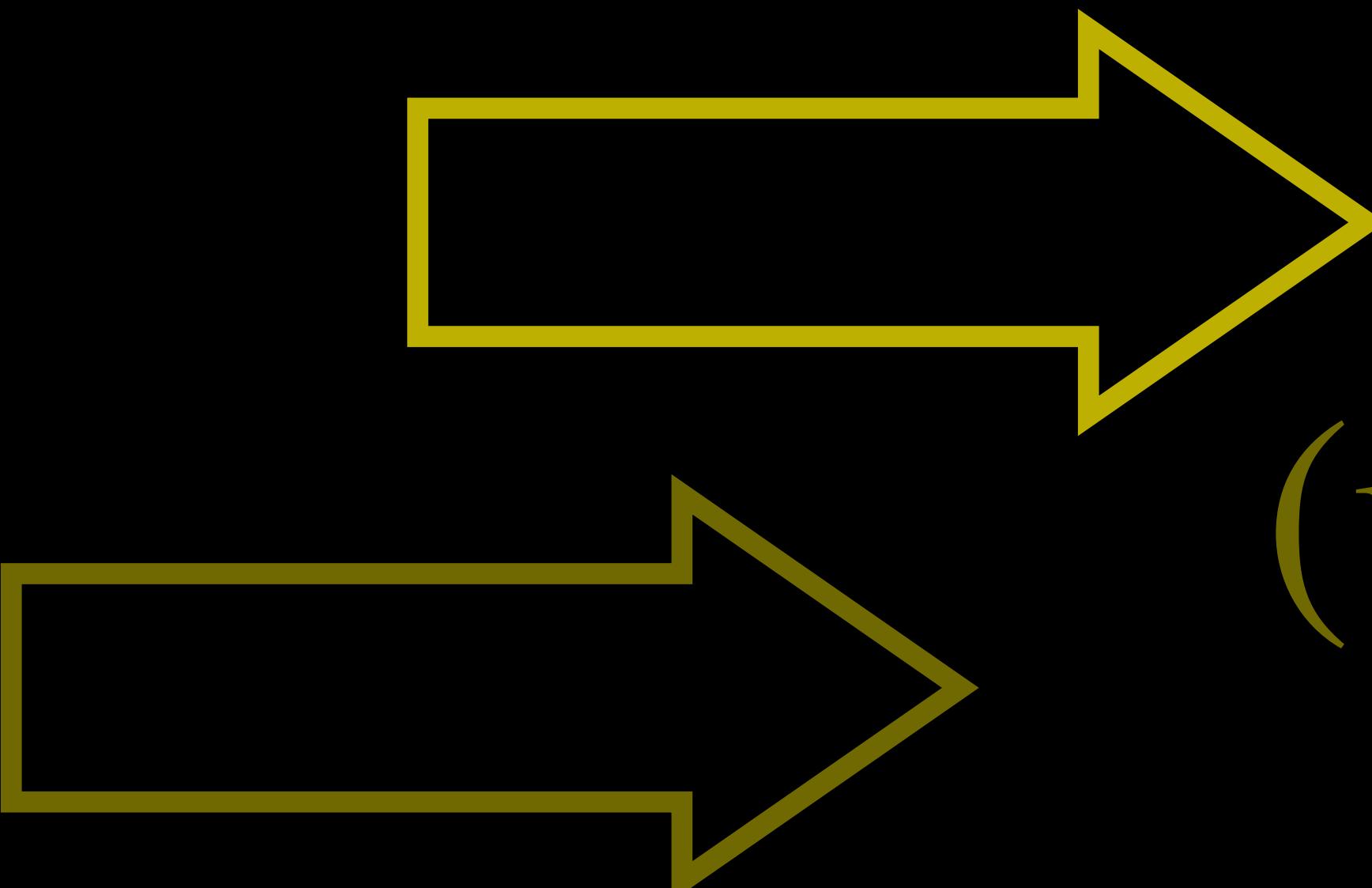


REDUCE

(w_1, S_1)

\vdots

(w_n, S_n)

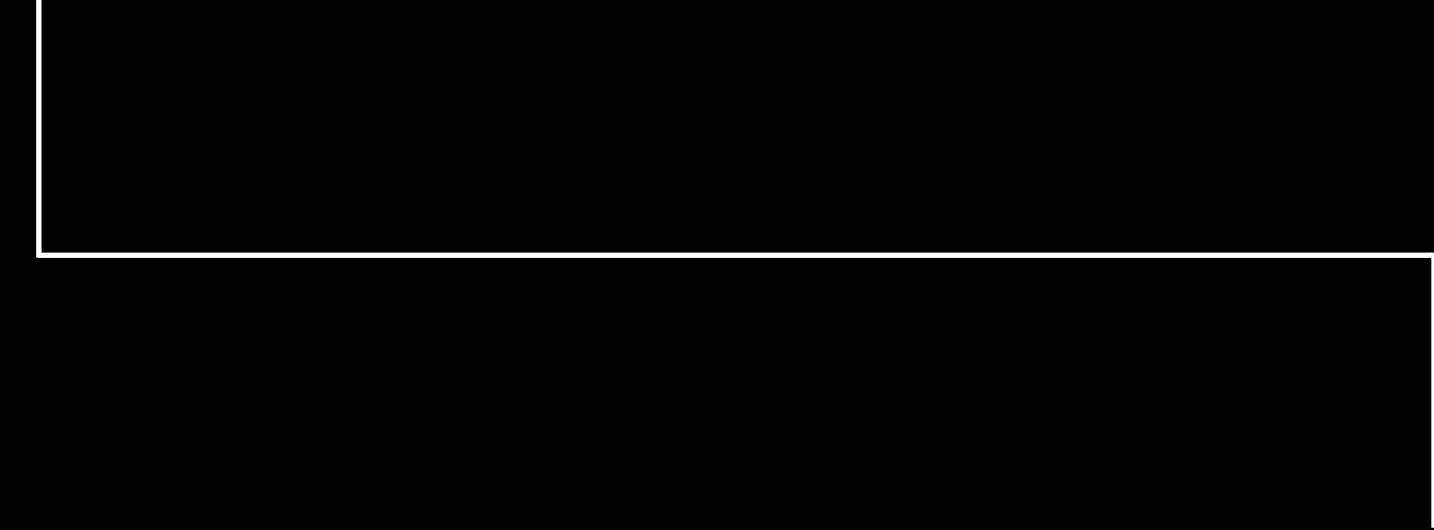


ЗАЧЕМ?

можем строить промежуточные словари с разным набором параметров **параллельно**, а значит **быстро** при наличии **вычислительных мощностей**

MAP ЭТАП

$(w_i, p_1, v_1, \dots, p_n, v_n, 1)_{i=1}^n$



только слова

REDUCE ЭТАП

$$(w_i, p_1, v_1, \dots, p_n, v_n, 1)_{i=1}^n$$

суммирование по ключу

$$\sum_{\substack{w_i \\ \downarrow \\ k_n}} (w_i, p_1, v_1, \dots, p_n, v_n, k_n)_{i=1}^n$$

ИТОГОВЫЙ СЛОВАРЬ

$$(w_i, p_1, v_1, \dots, p_n, v_n, f)_{i=1}^n$$

сортируем по необходимым
параметрам промежуточный словарь

MAP

$$(w_i, f_i)_{i=1}^n \Rightarrow \sum_{w_i} \Rightarrow (w_i, f)_{i=1}^n$$

REDUCE

ЗАМЕЧАНИЕ

$$(w_i, p_1, v_1, \dots, p_n, v_n, f)_{i=1}^n$$

для слов с $f \leq 4$ оценка статистически не достоверна
отсечение таких слов приводит к уменьшению
итогового словаря в несколько раз

IPM

1. в том же проходе считаем суммарную частоту вхождений всех слов S

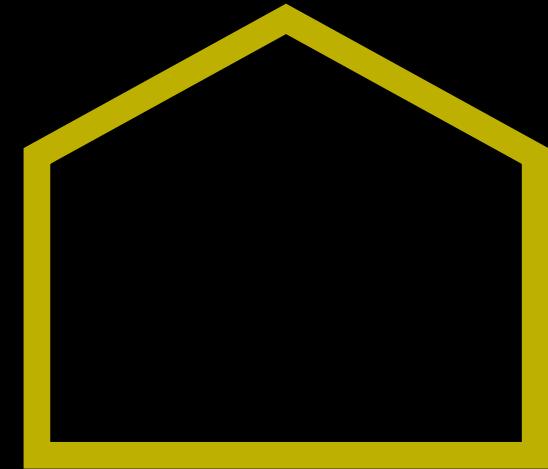
2. далее считаем интересующую нас относительную частоту $p = \frac{f}{S}$

3. строим доверительный интервал $p \pm v$, где

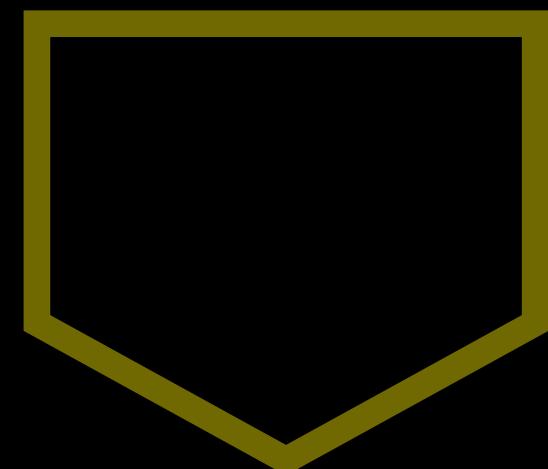
$$v = z \sqrt{\frac{p(1-p)}{S}},$$

где z – z-score желаемого уровня доверия

пословный

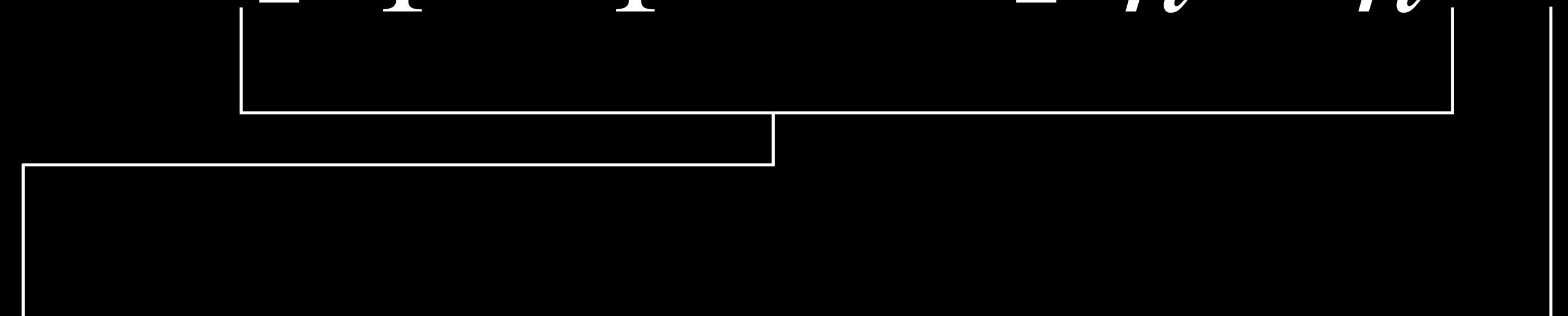


1. IPM



подокументный

ПРОМЕЖУТОЧНЫЙ СЛОВАРЬ

$$(w, p_1, v_1, \dots, p_n, v_n, d_i)_{i=1}^n$$


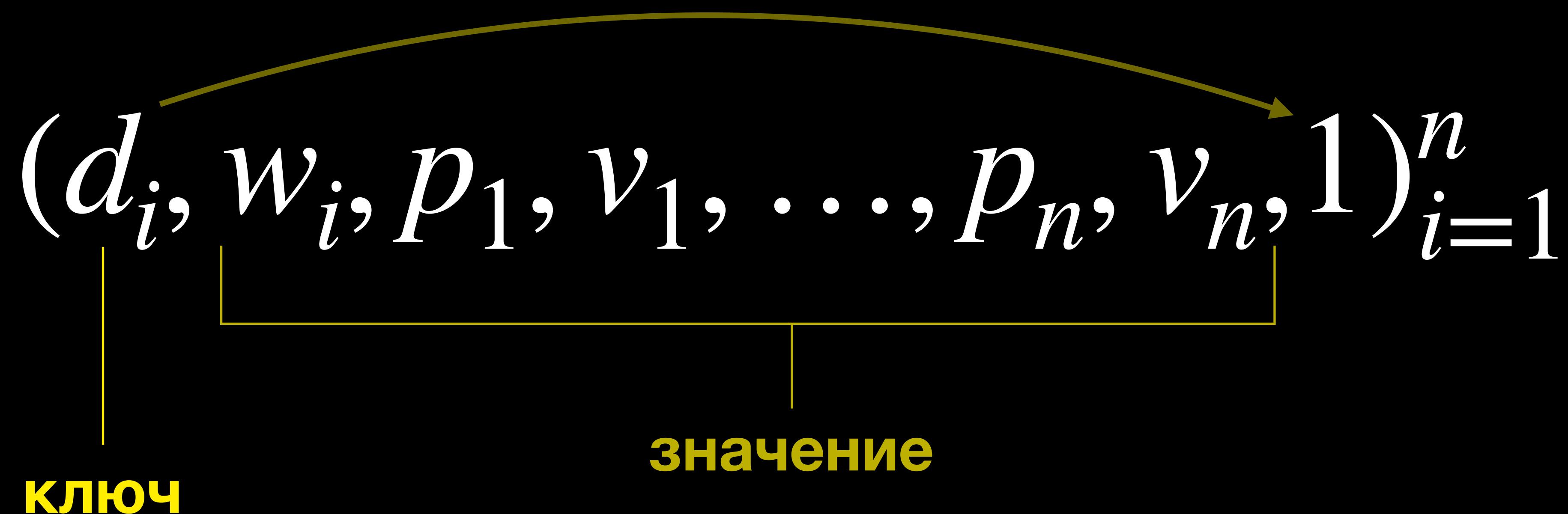
p_i – имя параметра

v_i – его значение

идентификатор

документа

MAP ЭТАП



REDUCE ЭТАП

$$(d_i, w_i, p_1, v_1, \dots, p_n, v_n, 1)_{i=1}^n$$

\downarrow

$$\sum_{w_i}$$

\downarrow

$$(d_i, w_i, p_1, v_1, \dots, p_n, v_n, f)_{i=1}^n$$

**количество документов,
содержащих слово**

IPM

1. считается **аналогично предыдущему**, за исключением того, что мы не можем просто суммировать частоты
2. необходимо отдельно завести список документов со строками вида
$$(d, p_1, v_1, \dots, p_n, v_n),$$
где d – идентификатор документа, уникальный в списке
3. далее просто **нормируем** на количество документов в списке из 2

2. СРАВНЕНИЕ КОРПУСОВ

АКА ПОИСК СМЕЩЕНИЙ

1. КЛЮЧЕВЫЕ СЛОВА

1. оба словаря фильтруются от **местоимений**
2. фильтруются от **обращений**
3. фильтруются от **часто используемые глаголов**, которые применяются для структуры предложения, а не для передачи смысла
4. из оставшихся берут **несколько самых частотных слов**

грубая, не количественная оценка



2. СТАТИСТИЧЕСКАЯ ГИПОТЕЗА

**гипотеза: наличие зависимости частоты
слова от подкорпуса**

	первый подкорпус	второй подкорпус	сумма
частота слова	f_1	f_2	$f_1 + f_2$
частота остатка	$S_1 - f_1$	$S_2 - f_2$	$S_1 + S_2 - f_1 - f_2$
сумма	S_1	S_2	$S_1 + S_2$

МЕТОДЫ ПРОВЕРКИ ГИПОТЕЗ

2.1. ХИ-КВАДРАТ

статистика для подсчета:

$$\chi^2 = \frac{(f_1 S_2 - f_2 S_1)^2}{(f_1 + f_2) (S_1 + S_2 - f_1 - f_2) S_1 S_2}$$



ненадежна при малых частотах,
что очень для нас критично

2.2. МАКНИМАРА

статистика для подсчета:

$$\chi^2 = \frac{(S_1 - f_2 - f_1)^2}{S_1 + f_2 + f_1}$$



не учитывает размер второго
корпуса, но прост в подсчете

2.3. ТОЧНЫЙ ТЕСТ ФИШЕРА

статистика для подсчета:

$$p = \binom{f_1 + f_2}{f_1} \binom{S_1 + S_2 - f_1 - f_2}{S_1 - f_1} / \binom{S_1 + S_2}{S_1}$$



сложен вычислительно, подсчёт
(!) очень долгий

2.3. LIKELYHOOD RATIO TEST

$$e_i = S_i \frac{f_1 + f_2}{S_1 + S_2} - \text{эффективность}$$

$$\mathcal{L} = 2 \sum_i \left(f_i \ln \frac{f_i}{e_i} \right) - \text{логарифм правдоподобия}$$

лишен вышеперечисленных недостатков



А ЧТО ЕСЛИ $f_1 = 0$?

— не проблема!

$$\lim_{f_1 \rightarrow 0} f_1 \ln \frac{f_1}{S_1 \frac{f_1 + f_2}{S_1 + S_2}} = \lim_{f_1 \rightarrow 0} f_1 \frac{f_1 (S_1 + S_2)}{S_1 f_2} = \lim_{f_1 \rightarrow 0} \left(f_1 \ln \frac{S_1 + S_2}{f_2} + f_1 \ln f_1 \right) = 0$$

в случае, если слово есть только в одном словаре, используем формулу:

$$\mathcal{L} = 2f_2 \ln \frac{S_1 + S_2}{S_2}$$

ГЕНЕРАЦИЯ

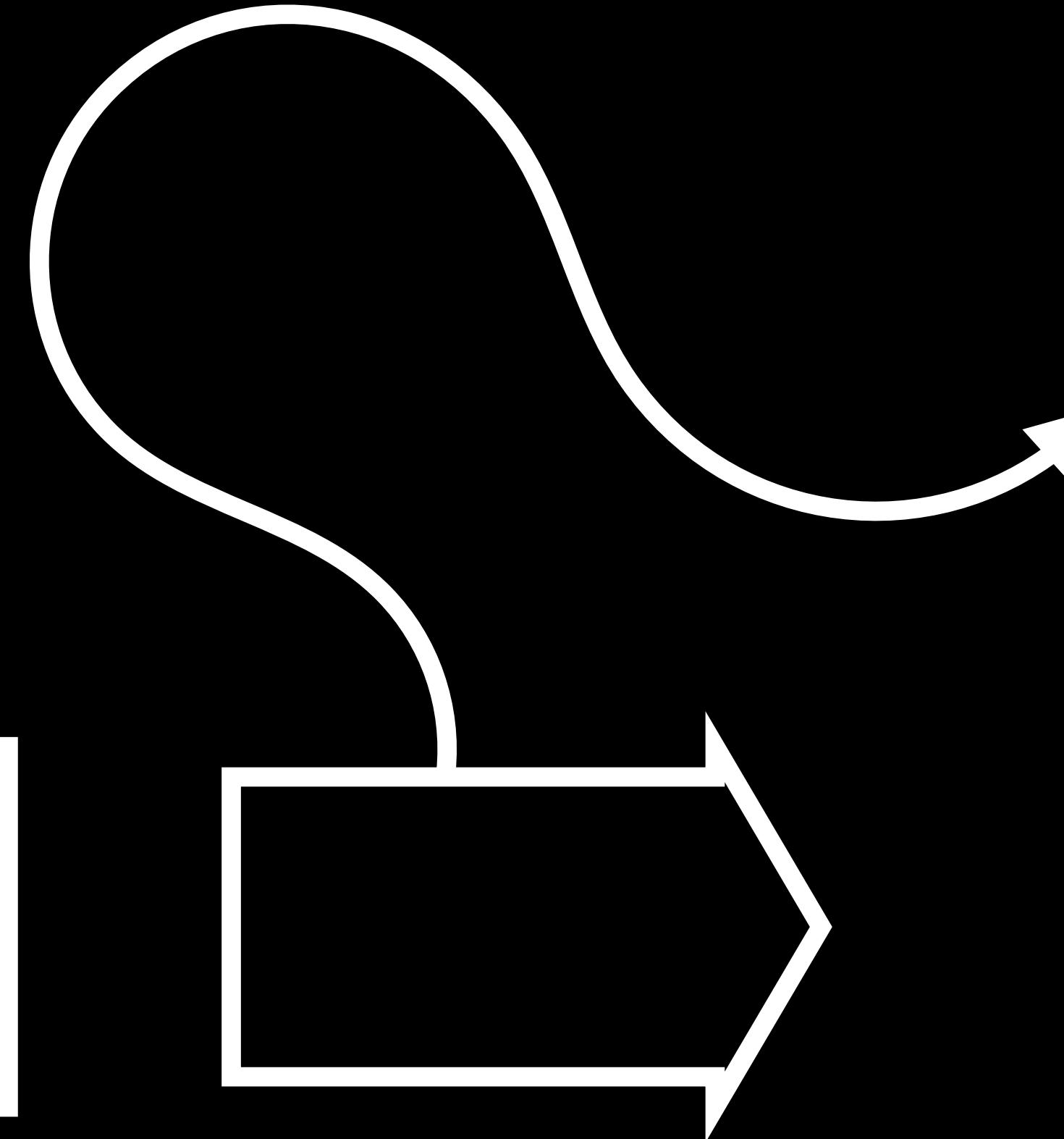
1. генерация двух дифференциальных словарей с соответствующими параметрами
2. сортируем оба словаря по словам и сливаем их сопоставляя “слово к слову”
3. если какое-то слово встречается только в одном словаре, то в другом оно имеет частоту 0

3. ВЫБРОСЫ

СТАТЬЯ ШАРОВА

ЗАДАЧА:

слово



тут
есть
проблема

IPM

не по корпусу,
а в реальном
языке

ПРОБЛЕМЫ

1.

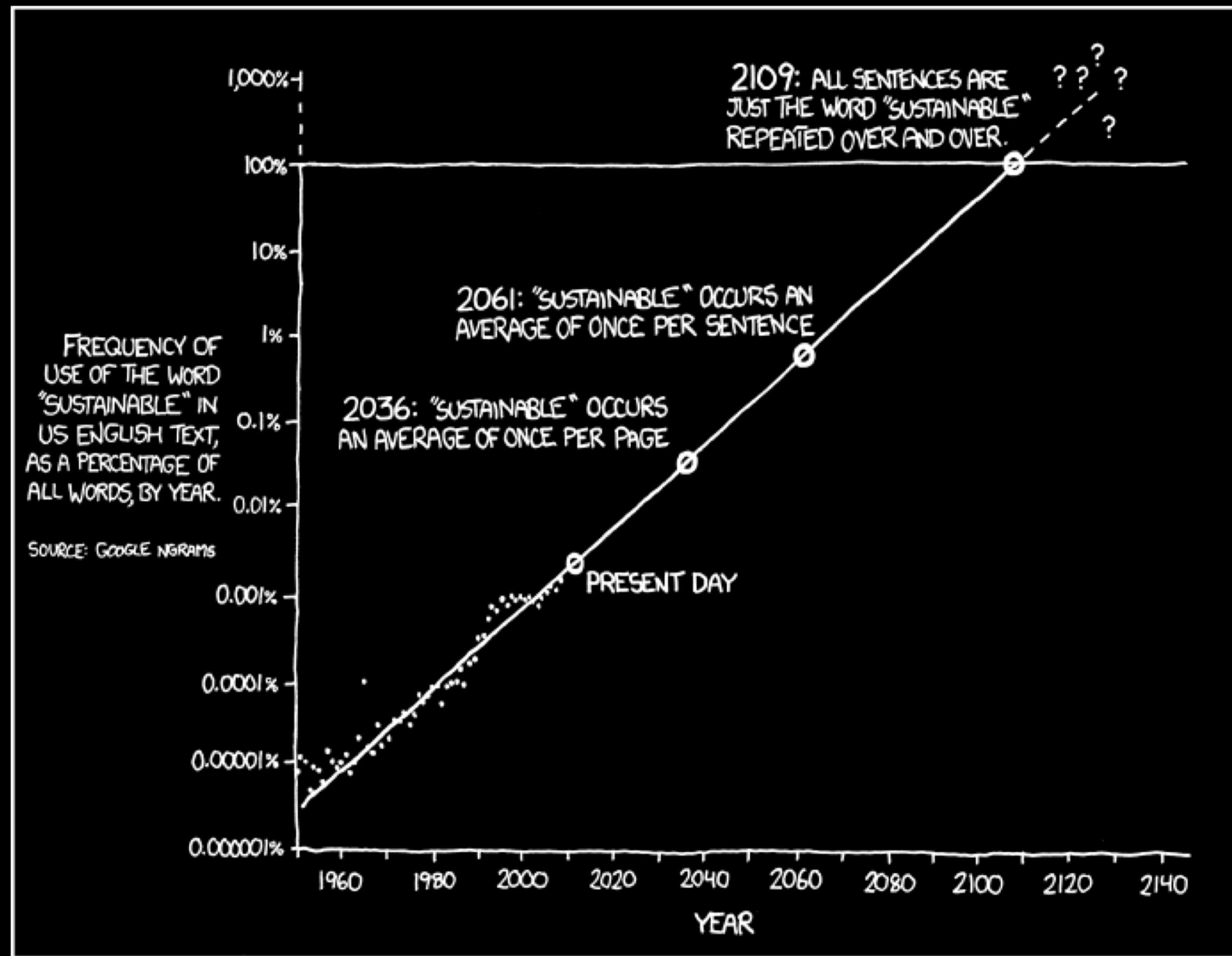
НКРЯ

“властелин колец” поднял слово “хоббит” в
1000 часто используемых

British National Corpus

moon, assert, crown, **gastric**₃₇₆₃, correct, lock, mutual, thoroughly
planner, evil, cage, **pylorus**₅₉₅₅, disguise, sulight, repay

2.



THE WORD "SUSTAINABLE" IS UNSUSTAINABLE.

ПОДХОДЫ К РЕШЕНИЮ

1. ВЗРЫВАЕМОСТЬ

$p(k = 0) = p_0$ – вероятность отсутствия слова в тексте

$p(k = 1) = p_1$ – вероятность появления его один раз

$p(k \geq 2) = \sum_{r \geq 2} p_r$ – вероятность возникновения больше одного раза

$\alpha = 1 - p_0$ – доля текстов, содержащих слово

$\gamma = 1 - \frac{p_1}{1 - p_0}$ – доля текстов, где слово тематическое

$B = \frac{\sum r p_r}{\sum p_r} (r \geq 2)$ – параметр взрыва

можно было бы удалять слова с большой взрываемостью, но проблема в том, что такими словами еще оказываются слова типа **do**, **have** и **not**

не самая лучшая оценка, потому что нет явного интервал возможных значений



2. УСТОЙЧИВЫЕ ОЦЕНКИ

дисперсия

$MAD = b \times \text{median} |x_i - \text{median } x|$ – median absolute deviation, $b = 1.48$

$S_n = c \times \text{median}_i \left(\text{median}_j |x_i - x_j| \right)$ – оценка руссо, $c = 1.19$

коэффициенты подобраны так, чтобы подогнуться
под нормальное распределение

МЕДИАНА huber-M

1. $\mu_0 = \text{median}$
2. μ_{k+1} обновляется итеративно после того, как выкидываются значения, удовлетворяющие следующему неравенству:

$$|x_i - \mu_k| > 1.28 \times MAD$$

если использовать эти алгоритмы в лоб, то нам будут мешать около нулевые значения, которые все сведут к 0

применяем винзоризацию, то есть сначала разделяемся с ненулевыми значениями

БУЛЬБАЗАВРСКАЯ ШЛЫКОВА

в работе проверяются подходы, описанные в статье
шарова к гикря, но винзоризация и huberM не дает
удовлетворительных результатов



4. РЕАЛИЗАЦИЯ

ПРОСТОЙ ПОДСЧЕТ

обращение по индексу: $\frac{found}{viewed} \cdot 10^6$

The screenshot shows a web application window titled 'GICR' with the URL 'int.webcorpora.ru/artu/'. The interface includes a sidebar with a 'geekря' logo and a yellow rubber duck icon. The main area has tabs for 'Your search filter', 'Search setup', and 'Snippets', with 'Snippets' currently selected. A 'New query' section contains a blue button labeled '***'. Below it is an 'Enter your query' field containing the text 'привет' and a 'Request the number of words' field. To the right, a 'Results' section displays the following data:

Results number:	IPM:	Documents number:	IPM:
9796	~97.960064	8909	8598.507880

Below this, there are fields for 'Search among words:' (100000000) and 'Or find number of results:', and buttons for 'Add query', 'Plot a chart', 'Save', 'Delete', and 'Share with'. At the bottom, there are checkboxes for 'Stat. query', 'Test mode', and 'Base query'.

СТАТИСТИЧЕСКИЙ АНАЛИЗ

по базовой методике, описанной ранее

The screenshot shows a web browser window with the title 'GICR'. The URL in the address bar is 'int.webcorpora.ru/artu/'. The page displays search results analysis for the query 'привет'.

On the left, there is a sidebar with a 'New query' section containing the text 'привет' and a 'Run shift search' button. Below this is another section labeled 'Помада'.

The main content area has tabs for 'Search setup', 'Snippets', and 'Results analysis'. The 'Results analysis' tab is active, showing a table of results:

Segment	IPM	+-	Error, %	Documentwise IPM	Documentwise +-	Error, %
Живой Журнал - Кассандра	54.387	1.235	1.135	4759.791	135.299	1.421
TOTAL	54.387	1.235	1.135	4759.791	135.299	1.421

At the bottom of the table, there is a 'Run shift search' button.

	Attribute	Value	Number of words in results	Number of documents in results	Share of words in results, %	Share of words in the corpus, %	Deviation, %	Share of documents in search results, %	Share of documents in the corpus, %	Deviation, %	Chi-squared	p-value

***	birth	0000	7255	6605	74.061	76.271	2.898	74.139	72.309	2.530	14.897	0.000
	birth	1900	3	3	0.031	0.093	66.989	0.034	0.123	72.545	5.758	0.016
***	birth	1902	8	7	0.082	0.006	1250.858	0.079	0.010	658.285	40.007	0.000
***	birth	1905	7	7	0.071	0.015	382.555	0.079	0.017	358.093	19.598	0.000
	birth	1907	10	4	0.102	0.013	706.540	0.045	0.018	148.293	3.543	0.060
***	birth	1908	2	2	0.020	0.006	218.246	0.022	0.007	209.807	2.842	0.092
	birth	1910	7	5	0.071	0.011	571.476	0.056	0.009	515.907	21.609	0.000
***	birth	1912	2	2	0.020	0.007	172.582	0.022	0.008	174.021	2.210	0.137
***	birth	1917	1	1	0.010	0.031	66.942	0.011	0.034	66.546	1.324	0.250
	birth	1918	3	3	0.031	0.010	198.495	0.034	0.009	282.988	6.274	0.012
***	birth	1923	6	6	0.061	0.006	988.016	0.067	0.005	1267.396	70.486	0.000
	birth	1930	3	3	0.031	0.005	514.706	0.034	0.003	995.986	27.154	0.000
***	birth	1941	1	1	0.010	0.016	34.407	0.011	0.012	2.943	0.001	0.976
	birth	1942	1	1	0.010	0.008	26.665	0.011	0.005	115.273	0.617	0.432
	birth	1948	16	13	0.163	0.036	348.660	0.146	0.024	511.077	55.581	0.000
	birth	1951	2	2	0.020	0.037	44.541	0.022	0.030	25.152	0.169	0.681

как описано ранее

СМЕЩЕНИЯ

5. ПЛАНЫ

ПРОСТОЙ ПОДСЧЕТ

хочется чтобы он был тоже статистическим
кстати, мы переходим на новый индекс

The screenshot shows a web browser window titled 'GICR' with the URL 'int.webcorpora.ru/artu/'. The interface includes a sidebar with a 'geekря' logo and a yellow rubber duck, and a top bar with user information ('terzi.va@phystech.edu') and navigation links ('Exit the search', 'Language ▾', 'Start ▾'). The main content area has tabs for 'Your search filter', 'Search setup', and 'Snippets', with 'Snippets' currently selected. A 'New query' section contains a blue button labeled '***'. Below it is an 'Enter your query' field containing the text 'привет' and a 'Request the number of words' field. To the right, a summary table displays results:

Results number:	IPM:	Documents number:	IPM:
9796	~97.960064	8909	8598.507880

Below the table, there are fields for 'Search among words:' (100000000) and 'Or find number of results:', and buttons for 'Add query', 'Plot a chart', 'Save', 'Delete', and 'Share with'. At the bottom, there are checkboxes for 'Stat. query', 'Test mode', and 'Base query'.

подокументный подсчет

хочется чтобы он был тоже статистическим
по методике, описанной ранее

The screenshot shows a web application window titled "GICR" with the URL "int.webcorpora.ru/artu/". The interface includes a search bar with placeholder text "Enter your query" containing "привет", a word count input field, and a results summary table.

Documents number:	IPM:
9796	~97.960064
8909	8598.507880

Below the table, there are sections for "Your search filter", "Search setup", and "Snippets". A "New query" section is visible with a blue button labeled "***". At the bottom, there are buttons for "Add query", "Plot a chart", and "Save", along with checkboxes for "Stat. query", "Test mode", and "Base query".

СТАТИСТИЧЕСКИЙ АНАЛИЗ

хочется по стратифицированной методике

The screenshot shows a web browser window titled 'GICR' with the URL 'int.webcorpora.ru/artu/'. The interface includes a sidebar with 'geekря' and a yellow rubber duck icon, and a top bar with user information ('terzi.va@phystech.edu') and navigation links ('Exit the search', 'Language ▾', 'Start ▾'). The main content area has tabs for 'Your search filter', 'Search setup', 'Snippets', and 'Results analysis'. The 'Results analysis' tab is active, displaying a table for the query 'привет'.

Search results analysis

Segment	IPM	+-	Error, %	Documentwise IPM	Documentwise +-	Error, %
Живой Журнал - Кассандра	54.387	1.235	1.135	4759.791	135.299	1.421
TOTAL	54.387	1.235	1.135	4759.791	135.299	1.421

Below the table, there is a button 'Run shift search'.

Помада

привет: Живой Журнал - Кассандра

Attribute	Value	Number of words in results	Number of documents in results	Share of words in results, %	Share of words in the corpus, %	Deviation, %	Share of documents in search results, %	Share of documents in the corpus, %	Deviation, %	Chi-squared	p-value
-----------	-------	----------------------------	--------------------------------	------------------------------	---------------------------------	--------------	---	-------------------------------------	--------------	-------------	---------

Attribute	Value	Number of words in results	Number of documents in results	Share of words in results, %	Share of words in the corpus, %	Deviation, %	Share of documents in search results, %	Share of documents in the corpus, %	Deviation, %	Chi-squared	p-value

birth	0000	7255	6605	74.061	76.271	2.898	74.139	72.309	2.530	14.897	0.000
birth	1900	3	3	0.031	0.093	66.989	0.034	0.123	72.545	5.758	0.016

birth	1902	8	7	0.082	0.006	1250.858	0.079	0.010	658.285	40.007	0.000

birth	1905	7	7	0.071	0.015	382.555	0.079	0.017	358.093	19.598	0.000

birth	1907	10	4	0.102	0.013	706.540	0.045	0.018	148.293	3.543	0.060

birth	1908	2	2	0.020	0.006	218.246	0.022	0.007	209.807	2.842	0.092

birth	1910	7	5	0.071	0.011	571.476	0.056	0.009	515.907	21.609	0.000

birth	1912	2	2	0.020	0.007	172.582	0.022	0.008	174.021	2.210	0.137

birth	1917	1	1	0.010	0.031	66.942	0.011	0.034	66.546	1.324	0.250

birth	1918	3	3	0.031	0.010	198.495	0.034	0.009	282.988	6.274	0.012

birth	1923	6	6	0.061	0.006	988.016	0.067	0.005	1267.396	70.486	0.000

birth	1930	3	3	0.031	0.005	514.706	0.034	0.003	995.986	27.154	0.000

birth	1941	1	1	0.010	0.016	34.407	0.011	0.012	2.943	0.001	0.976

birth	1942	1	1	0.010	0.008	26.665	0.011	0.005	115.273	0.617	0.432

birth	1948	16	13	0.163	0.036	348.660	0.146	0.024	511.077	55.581	0.000

birth	1951	2	2	0.020	0.037	44.541	0.022	0.030	25.152	0.169	0.681

хочется задавать **параметры смещений руками + добавить необходимые подсказки**

СМЕЩЕНИЯ

6. ИНТЕРФЕЙС

СЕЙЧАС

ГИКРЯ

int.webcorpora.ru/artu/

geekря 🐥

terzi.va@phystech.edu Выйти из системы Язык ▾ Пуск ▾

Фильтр по Вашим поискам Настройка поиска Сниппеты Анализ результатов

Новый поиск

Ведите ваш запрос
привет

Запрос количества слов

Результатов: IPM: 9796 Документов: IPM: ~97.960064 8909 8598.507880

Добавить запрос Построить график

Искать среди слов: 100000000 или найти столько результатов:

Имя запроса: *** Сохранить Удалить Поделиться с

Стат. запрос Тестовый режим Базовый запрос

Помада

Включить? Сегмент Слов: balanceHeader-template

<input checked="" type="checkbox"/>	Живой Журнал - Кассандра	8720 млн.	<div style="width: 8720px;"></div>
<input type="checkbox"/>	ВКонтакте - Кассандра	9820 млн.	<div style="width: 9820px;"></div>
<input type="checkbox"/>	Новости - Кассандра	851 млн.	<div style="width: 851px;"></div>
<input type="checkbox"/>	Журнальный Зал Кассандра	313 млн.	<div style="width: 313px;"></div>

Запустить Остановить

БУДЕТ

ГИКРЯ  int.webcorpora.ru/artu/   

ГЕЕКРЯ

 – инструкция [somebody@mailbox.com](#) [Меню](#)  [Выход](#)

[▶ Фильтровать поиски](#) [Настройки поиска](#) [Сниппеты](#) [Анализ результатов](#)

Новый поиск
Поиск 1
Поиск 2
привет
Поиск 4
Поиск 5
Поиск 6
Поиск 7
Поиск 8
Поиск 9

Ваш запрос
привет 
+ Построить график
искать среди слов или найти столько слов
имя запроса   

Запрос количества слов
Результатов: IPM:
Документов: IPM:

Сегмент	Слов (млн)	Баланс корпусов	нормировать
Живой журнал	8720		
ВК	9820		
Новости	851		
Журнальный зал	313		

[Остановка](#) **Запуск**

ОСНОВНЫЕ МОМЕНТЫ

ОСНОВНЫЕ МОМЕНТЫ

1. нужны отступы от кнопок
2. некоторые названия стоит поменять
3. редактируемые колонки в таблицах
4. более удобный выбор корпусов
5. более понятные подсказки
6. добавить инструкцию
7. (переделать инструкцию)
8. диалоговое окно не вылезает за экран
9. вывод IPM

ГИКРЯ

int.webcorpora.ru/artu/

geekря 🐥

terzi.va@phystech.edu Выйти из системы Язык ▾ Пуск ▾

Фильтр по Вашим поискам Настройка поиска Сниппеты Анализ результатов

Новый поиск

Ведите ваш запрос
привет

Запрос количества слов

Результатов: IPM: 9796 Документов: IPM: ~97.960064 8909 8598.507880

Добавить запрос Построить график

Искать среди слов: 100000000 или найти столько результатов:

Имя запроса: *** Сохранить Удалить Поделиться с

Стат. запрос Тестовый режим Базовый запрос

Помада

balanceHeader-template

Включить?	Сегмент	Слов:
<input checked="" type="checkbox"/>	Живой Журнал - Кассандра	8720 млн.
<input type="checkbox"/>	ВКонтакте - Кассандра	9820 млн.
<input type="checkbox"/>	Новости - Кассандра	851 млн.
<input type="checkbox"/>	Журнальный Зал Кассандра	313 млн.

Запустить Остановить

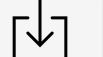
ГИКРЯ  int.webcorpora.ru/artu/   

ГЕЕКРЯ

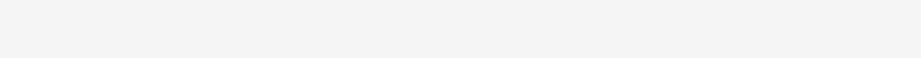
 – инструкция [somebody@mailbox.com](#) [Меню](#)  [Выход](#)

[Фильтровать поиски](#) [Настройки поиска](#) [Сниппеты](#) [Анализ результатов](#)

Новый поиск
Поиск 1
Поиск 2
привет
Поиск 4
Поиск 5
Поиск 6
Поиск 7
Поиск 8
Поиск 9

Ваш запрос
привет 
+ Построить график
искать среди слов или найти столько слов
имя запроса   

Запрос количества слов
Результатов: IPM:
Документов: IPM:

Сегмент	Слов (млн)	Баланс корпусов	нормировать
Живой журнал	8720		
ВК	9820		
Новости	851		
Журнальный зал	313		

[Остановка](#) **Запуск**

ГИКРЯ

int.webcorpora.ru/vlad/

geekря 🐥

Фильтр по Вашим поискам Настройка поиска Сниппеты

Новый поиск

Результат поиска

Слева Результат Справа

	U.	Слева	Результат	Справа
1	htt...	Бартаев (14.40.05 29/09/2010) Кать , привет	о , уезжаю , приеду , напишу . -) Бартаев (14.40.05 29/09/2010)	
2	htt...	наклоняюсь к ней , - ты откуда ? кто ? " я , - говорит , - Весна , - говорит , - привет	! я , - говорит , - с собой не совсем в ладу . вот - го	
3	htt...		о , привет	, Син !))
4	htt...	жизнь , сначала театральные друзья и подруги , а потом еще и " привет	из прошлого " , рушат все планы главного героя . М	
5	htt...	дувало легким матерком ... Пешеходный переезд . Порнометражный фильм	Привет участникам естественного отбора ! Придумают же л	
6	htt...		Всем привет	! Давненько меня здесь не было . Мои перерывы в ж
7	htt...		Привет	, мой самый преданный комментатор ! :) Да , я тож д
8	htt...		Всем привет	. Как выходные провели ? Удивительно , но второй у
9	htt...		Привет	! Наверное меня здесь уже все забыли)) Полтора м
10	htt...		Всем привет	! Мы вернулись !!! Неделька свободы была зачетная
11	htt...		Всем привет	:) И все - таки я человек , который очень быстро " п
12	htt...		Привет	, а я Катя . Живу в Москве . Работаю корреспондент
13	htt...	же тебя ... побьют !! Надпись на парте на филфаке : Мальчики филфака , привет	! Ниже этим же почерком (видимо , позже приписан	
14	htt...		привет	, девчёнки . недавно в ЖЖ .Хочу написать свою исто
15	htt...		Аня , привет	! Рад твоему ЖЖивому отклику ! Знаешь , я по уст
16	htt...		Привет	! Я уезжал . Вот вернулся , отвечаю :) Каким бы забс
17	htt...		Привет	. Как обычно : у меня под вечер , но как всегда)
18	htt...	на пару постов о " кухне " , ну и о кухне наверно)))	Привет	, что ли))))))
19	htt...		Привет	! Неделя за неделей проходят незаметно . Я на онел

Удалить выбранные результаты Создать выгрузку Сравнить с прямым результатом Запустить анализ результатов

Выберите колонки для группировки:

- ??SnippetWidgetFieldDescrTooltip
- Author
- Author's year of birth
- Document ID
- Document size

ГИКРЯ

int.webcorpora.ru/vlad/

Выберите колонки для группировки:

??SnippetWidgetFieldDescrTooltip
Author
Author's year of birth
Document ID
Document size

Показать запрос

Выполнить запрос

Value	Count
Живой Журнал - Ка...	9796

Создать выгрузку результатов группировки

Выбрать кластер

Отображать все кластеры

Применить

ОЧЕНЬ ХОЧЕТСЯ

нормальную и удобную регистрацию в корпус