Тема:

Текстовое ранжирование и BM 25

SEO intellect

Докладчик:
Латыпов Артур
Email: artur@seointellect.ru
Тел.: +7 (495) 125-20-11

Анализ внутренних факторов

Wikipedia

BM25F — *модификация BM25, в которой документ рассматривается как совокупность нескольких полей (таких как, например, заголовки, основной текст, ссылочный текст), длины которых независимо нормализуются, и каждому из которых может быть назначена своя степень значимости в итоговой функции ранжировании.*

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})}$$

$$\text{IDF}(q_i) = \log \frac{N}{n(q_i)}$$

Внутренние факторы

- **Bm25 по зонам документа:** *title, исходящие анкоры в body, анкор-лист, фрагменты текста в body*

- **Учет различных вариантов вхождений**
  - Прямые
  - В леммах
  - В разном порядке
  - Синонимы
  - Частичные
  - Наличие всех слов из запроса

- **Месторасположение на документе или в зоне документа**

Текстовое ранжирование

$$Score = W_{single} + W_{pair} + k_1 * W_{AllWords} + k_2 * W_{Phrase} + k_3 * W_{HalfPhrase} + W_{PRF}$$

- встречаемость слов из запроса в документе (W_{single})
- встречаемость пар слов из запроса в документе (W_{pair})
- встречаемость текста запроса целиком (W_{Phrase})
- наличие всех слов запроса в документе ($W_{AllWords}$)
- количество предложений с наличием многих слов запроса в одном предложении ($W_{HalfPhrase}$)

Текстовое ранжирование

$$W_{single} = \log(p) * (TF_1 + 0.2 * TF_2)$$

$$TF_1 = \frac{TF}{TF + k_1 + k_2 * DocLength}, k_1 = 1, k_2 = 1/350$$

$$TF_2 = \frac{Hdr}{1 + Hdr}$$

$$p = 1 - \exp(-1.5 * \frac{CF}{D})$$

Hdr – сумма весов слова за форматирование

CF – число вхождений леммы в коллекцию

D – число документов в коллекции

Текстовое ранжирование

$$W_{pair} = 0.3 * (\log(p_1) + \log(p_2)) * \frac{TF}{1 + TF}$$

Учет пар слов:

- слова запроса встречаются в тексте подряд - 1 ,
- через слово или в обратном порядке - 0.5
- слова из трехсловных запросов через слово идут подряд – 0.1

Текстовое ранжирование

$$W_{AllWords} = 0.2 * \sum \log(p_i) * 0.03^{N_{miss}}$$

N_{miss} – количество отсутствующих в документе слов запроса

Бонус за наличие всех слов в документе.

Что влияет на релевантность

- *Есть все слова где-то в документе*
- *Все слова подряд в документе*
- *Точное вхождение в первом предложении*
- *Если текстовая релевантность мала и ссылок мало – бонус для вхождения фраз в анкоры ссылок*
- *Совпадение леммы слова*
- *Простой BM25 по тексту (+ с учетом синонимов)*
- *Наличие пары слов из запроса в точной форме (+ синонимы)*
- *Наличие слов в заголовке (+ с учетом синонимов)*

Что влияет на релевантность

- *BM25 только по предложением в которых есть ключ*
- *BM25 только по заголовкам*
- *BM25 по разным тэгам выделений (<b> <strong и другие)*
- *Количество предложений с словами из ключа (+ синонимы)*
- *Качество текста – именно как фактор*
- *BM25 с приоритетом начала документа и загловков*
- *Релевантность участков (plain-text , head , title)*
- *Вероятность встретить запрос в документе – тематическая близость*

Проверка влияния BM25 на релевантность

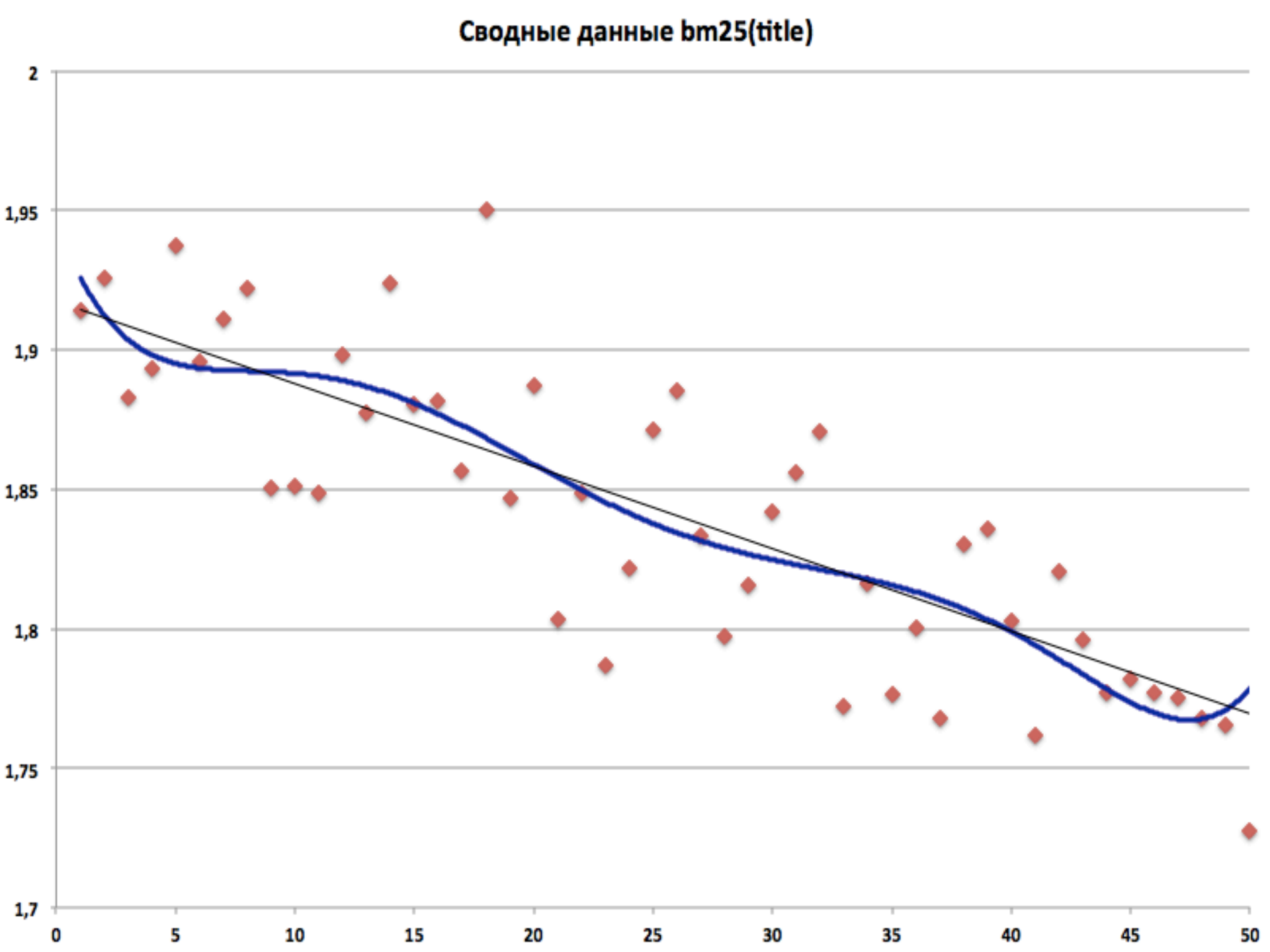
Анализ внутренних факторов

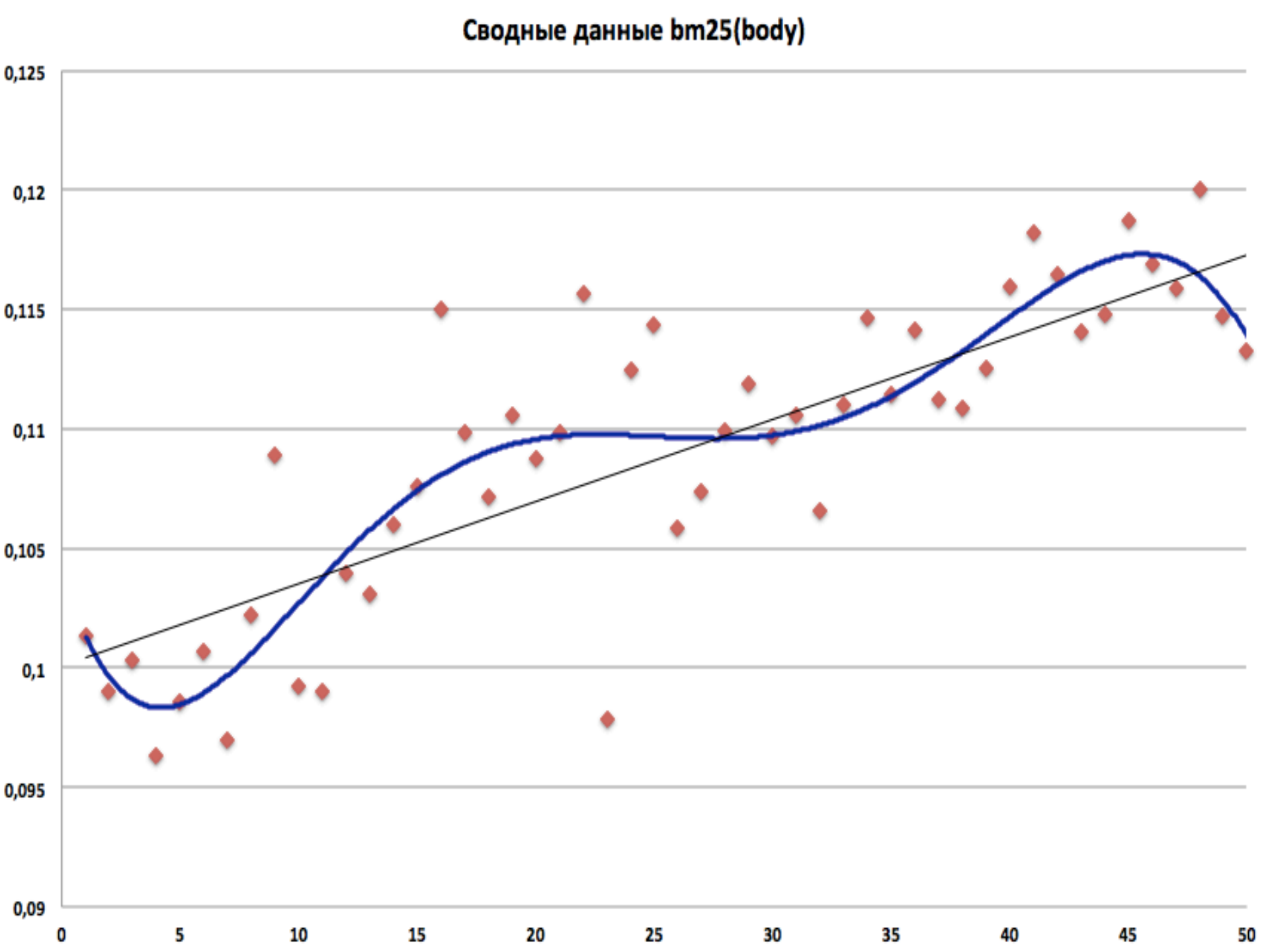
Расчет bm25 для 2-х зон документа:

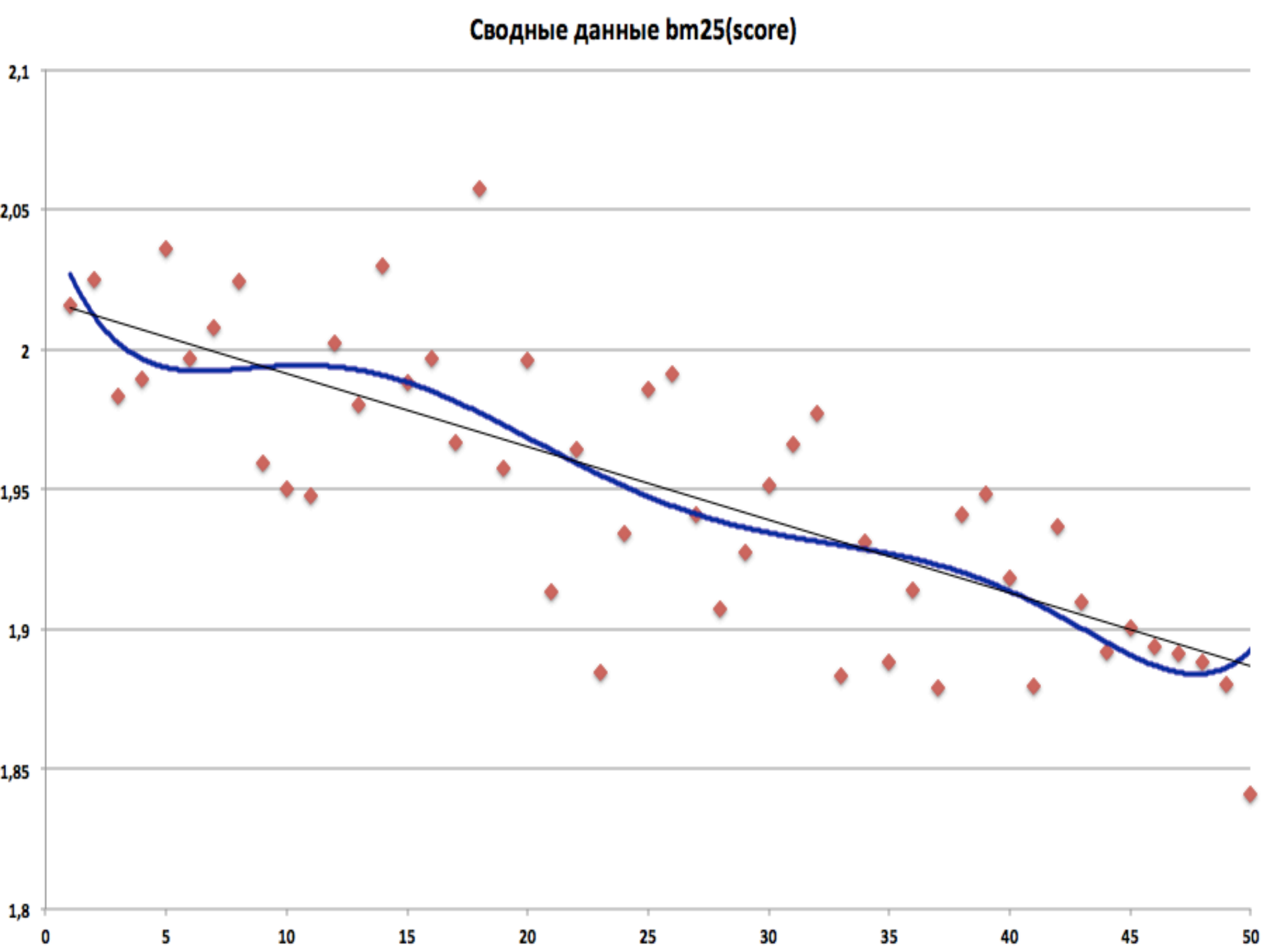
- Title
- Body (без разбиения на фрагменты)
- Score по title и body

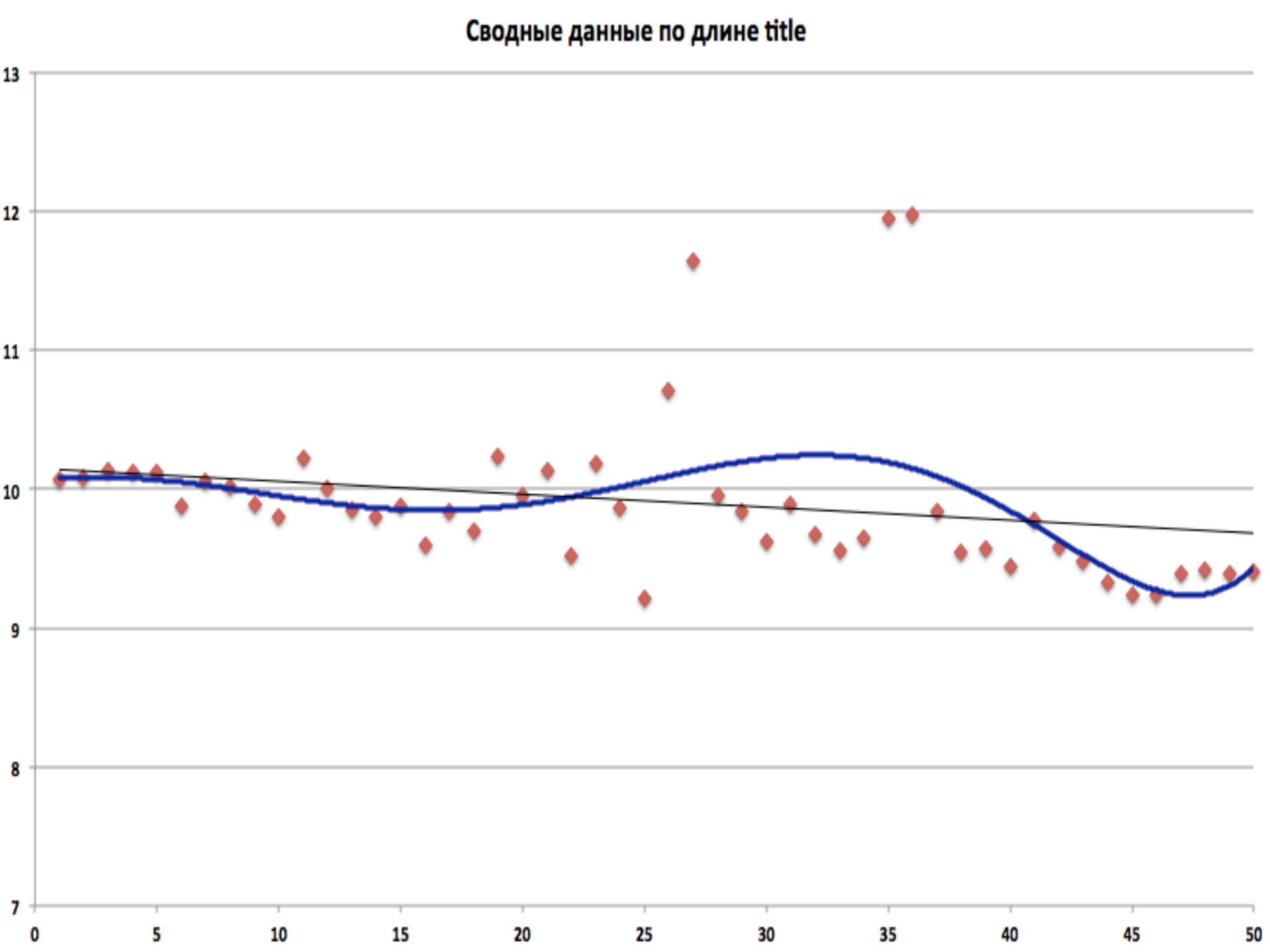
- Количество запросов - 9915
- Длина запросов - 1-3 слова
- Тип – коммерческие

- Собрано документов – 495750

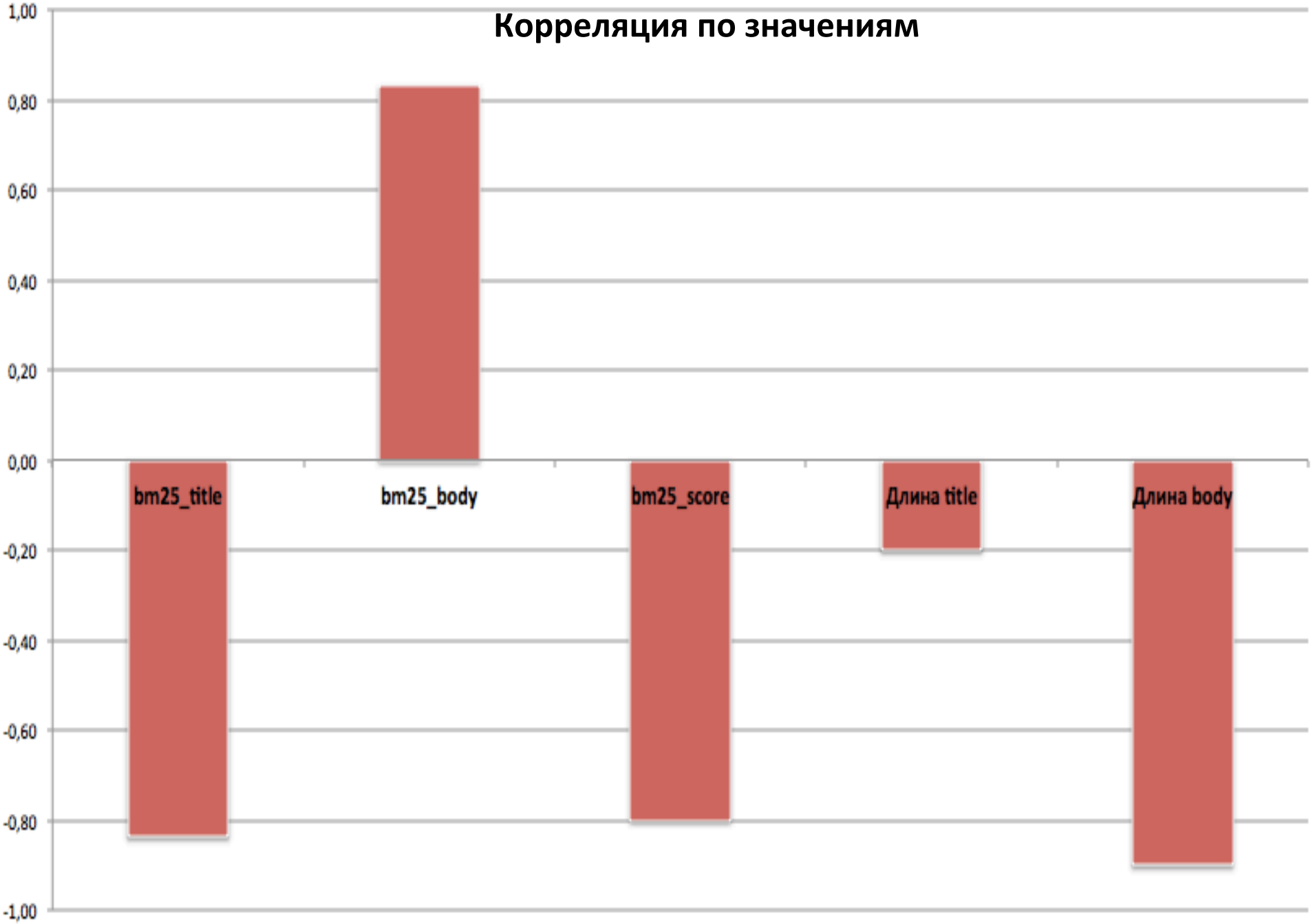

Сводные данные bm25(title)
2
1,95
1,9
1,85
1,8
1,75
1,7
0
5
10
15
20
25
30
35
40
45
50


Сводные данные bm25(body)
0,125
0,12
0,115
0,11
0,105
0,1
0,095
0,09
0
5
10
15
20
25
30
35
40
45
50


Сводные данные bm25(score)
2,1
2,05
2
1,95
1,9
1,85
1,8
0
5
10
15
20
25
30
35
40
45
50


Сводные данные по длине title
13
12
11
10
9
8
7
0
5
10
15
20
25
30
35
40
45
50


Сводные данные по длине body
1200
1100
1000
900
800
700
600
0
5
10
15
20
25
30
35
40
45
50


Корреляция по значениям
1,00
0,80
0,60
0,40
0,20
0,00
-0,20
-0,40
-0,60
-0,80
-1,00
bm25_title
bm25_body
bm25_score
Длина title
Длина body

Анализ внутренних факторов

Выводы

Существует корреляционная зависимость между позицией документа и формулой текстовой релевантности bm25

По зоне документа (body) – больший bm25 не значит лучше

Сильно влияет больший bm25 в зоне документа (title) и score по всему документу

Нужно рассчитывать по разным зонам документа

ВАЖНО При оптимизации ВЧ и СЧ запросов учитывать для каждого запроса (есть запросы, где зависимость не прослеживается)

***ВАЖНО* Расчет может быть в виде рекомендаций по кол-ву вхождений и объему зоны документа**

BM25 в цифрах

BM25 в цифрах

Велотренажер housefit купить в интернет магазине Москва

Велотренажер housefit

Купить велотренажер housefit kinetic

Велотренажер housefit купить велотренажер в интернет магазине в Москве

Велотренажер housefit купить housefit в интернет магазине в Москве

Фраза “**велотренажер housefit**”

Где текст релевантней???

BM25 в цифрах

	всего слов
Велотренажер housefit купить в интернет магазине Москва	7
Велотренажер housefit	2
Купить велотренажер housefit kinetic	4
Велотренажер housefit купить велотренажер в интернет магазине в Москве	9
Велотренажер housefit купить housefit в интернет магазине в Москве	9

кол-во слов		в корпусе		TF	
велотренажер	housefit	велотренажер	housefit	велотренажер	housefit
1	1	4000000	500000	0,142857143	0,142857143
1	1	4000000	500000	0,5	0,5
1	1	4000000	500000	0,25	0,25
2	1	4000000	500000	0,222222222	0,111111111
1	2	4000000	500000	0,111111111	0,222222222

BM25 в цифрах

	док 1	
	велотренажер	housefit
общее кол-во документов в коллекции	5000000000	5000000000
кол-во документов содержащих слово	4000000	500000
длинна документа	7	7
частота слова	0,142857143	0,142857143
средняя длина документа в коллекции	500	500
k1	2	2
b	0,75	0,75
формула	0,645577792	0,645577792
idf	3,096910013	4
score	1,999296329	2,582311168
Велотренажер housefit	**4,581607497**	

BM25 в цифрах

	bm25
Велотренажер housefit купить в интернет магазине Москва	4,58
Велотренажер housefit	10,58
Купить велотренажер housefit kinetic	6,99
Велотренажер housefit купить велотренажер в интернет магазине в Москве	4,85
Велотренажер housefit купить housefit в интернет магазине в Москве	5,18

BM25 в цифрах

Более простое решение – анализ топа.

Запросы	Прямое вхождение в <title>	Прямое вхождение в <body>	Всего в <body>	Всего в <a>	Объём контента
велотренажеры housefit	есть	1	2	12	4216

Подсветка
москве, москва

http://www.trenazer.ru/velotrenazhery/Housefit 4216

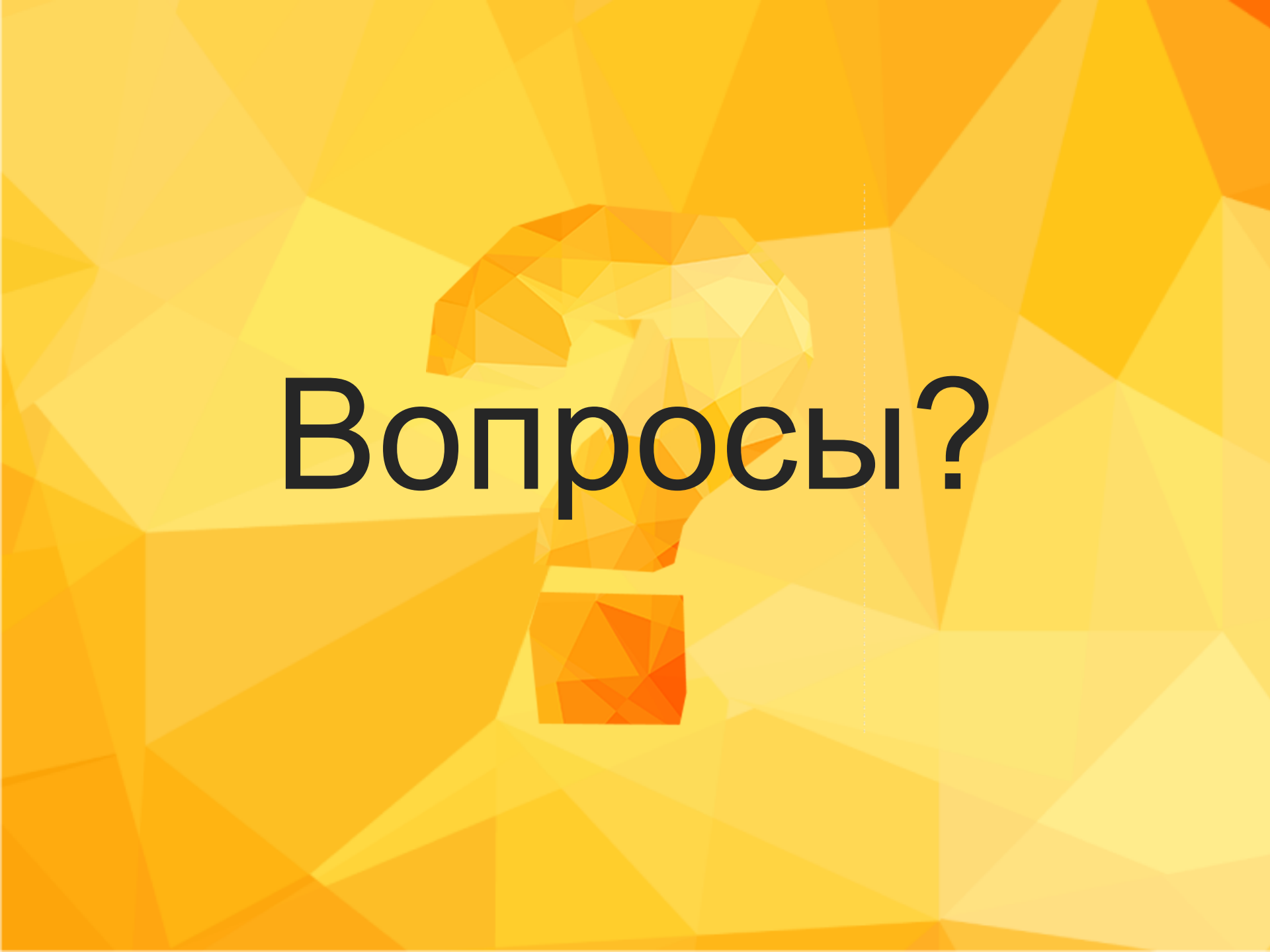
Слова из запросов	Непрямые вхождения	Слова из запросов	Непрямые вхождения
велотренажеры	7	housefit	4

Рекомендации для http://www.trenazer.ru/velotrenazhery/Housefit

Запросы	Прямое вхождение в <title>	Прямое вхождение в <body>	Всего в <body>	Всего в <a>	Объём контента
велотренажеры housefit	есть	1	1 (+ 1)	23 (- 11)	3207

+ автоматический сбор тематически близких слов. Из Engine:

вертикальный вес доставка занятий кг купить магазина магнитная модель нагрузки оборудование отзывов пользователей профессионального система спортом уровня


Вопросы?

SEO intellect

- **Продвижение сайтов**
- **SEO консалтинг, аудиты**
- **Контекстная реклама**
- **Изготовление сайтов**

Сайт: http://seointellect.ru

Докладчик:
Латыпов Артур
Email: artur@seointellect.ru
Тел.: +7 (495) 125-20-11