14 августа 2015
BALTIC DIGITAL DAYS

Максимум текстовой релевантности сегодня: факторы, практические рекомендации

BalticDigitalDays

Пиксель Плюс
Интернет-агентство

Дмитрий Сева́льнев

О спикере

- ✓ Ведение и контроль более 150 сайтов по рекламе в интернете и SEO на постоянной основе.
- ✓ Ведущий передач «Практика SEO», «Аналитика SEO» и «Познай ТОП» на MegaIndex.tv.
- ✓ Участие в развитии более 550 проектов.
- ✓ Выступления на отраслевых конференциях: IBC Russia, РИФ+КИБ, SEO Conference, СПИК, SEO 2015, MegaIndex, Нетология, …

«Личный блог»

www.pixelplus.ru/samostoyatelno/

Пиксель Плюс
Интернет-агентство

Тезисы

1. Основные факторы текстового ранжирования.

а) Встречаемость слов и их вес.

- Значимость факторов.
- Рекомендации.

б) Фразовые соответствия.

- Типы соответствия.
- Использование в поиске и на практике.

в) Синонимы.

- Определение.
- Использование в тексте и на сайте.

2. Антиспам -VS- ранжирование.

3. Практические рекомендации по формированию ТЗ для копирайтера.

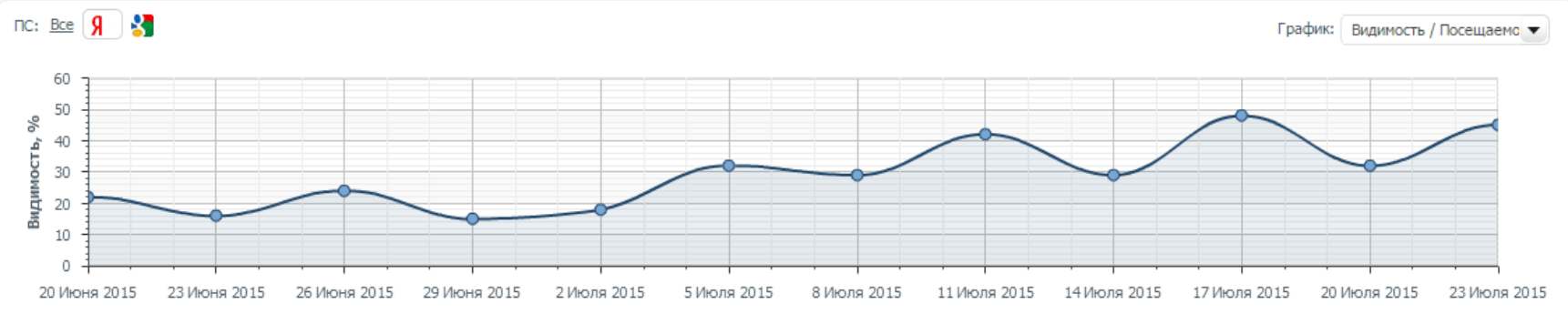
а) Формирование требований для копирайтера (ТЗ).

б) Основные ошибки, допускаемые оптимизаторами при формировании ТЗ.

4. Рекомендации.

Позиции вообще зависят от текста?

Текст и текстовый антиспам



- Продолжает наблюдаться повышенная значимость текстовых факторов ранжирования.
- Важность текстовых характеристик хоста.
- Основные изменения: рост аналитической базы в работе с текстовыми факторами.

Пиксель Плюс
Интернет-агентство

Текст и поисковые системы

Группы текстовых факторов

«Частотные» по отдельным словам	Взаимное расположение и позиции слов	Качество и общие параметры текста
Хостовые (уникальность, спамность)	Факторы антиспама	Соответствие языковой модели

Встречаемость слов и их вес

Не все слова одинаково полезны…

$$\text{Вес} \sim log\left(\frac{\text{Всего документов}}{1 + \text{Документы с нужным словом}}\right)$$

- Вес = IDF
- IDF фигурирует во множестве текстовых метрик всех поисковых систем
- Базовая характеристика слова

Пример вычисления IDF (Яндекс)

Типичные значения для «веса»

Буква, предлог, слово, число	1/Встречаемость	Вес	Нормировка на «в»
в	51	1,7	1,0
на	90	2,0	1,1
сайт	410	2,6	1,5
новости	572	2,8	1,6
форум	701	2,8	1,7
работа	707	2,8	1,7
москва	869	2,9	1,7
помощь	1 840	3,3	1,9
музыка	1 891	3,3	1,9
одноклассник	38 226	4,6	2,7
низ	132 190	5,1	3,0
зависть	155 142	5,2	3,0
гугл	182 181	5,3	3,1
маил	1 130 526	6,1	3,5
виндоус	2 055 716	6,3	3,7

В помощь: tools.promosite.ru/old/weight.php

Факторы на базе TF-IDF

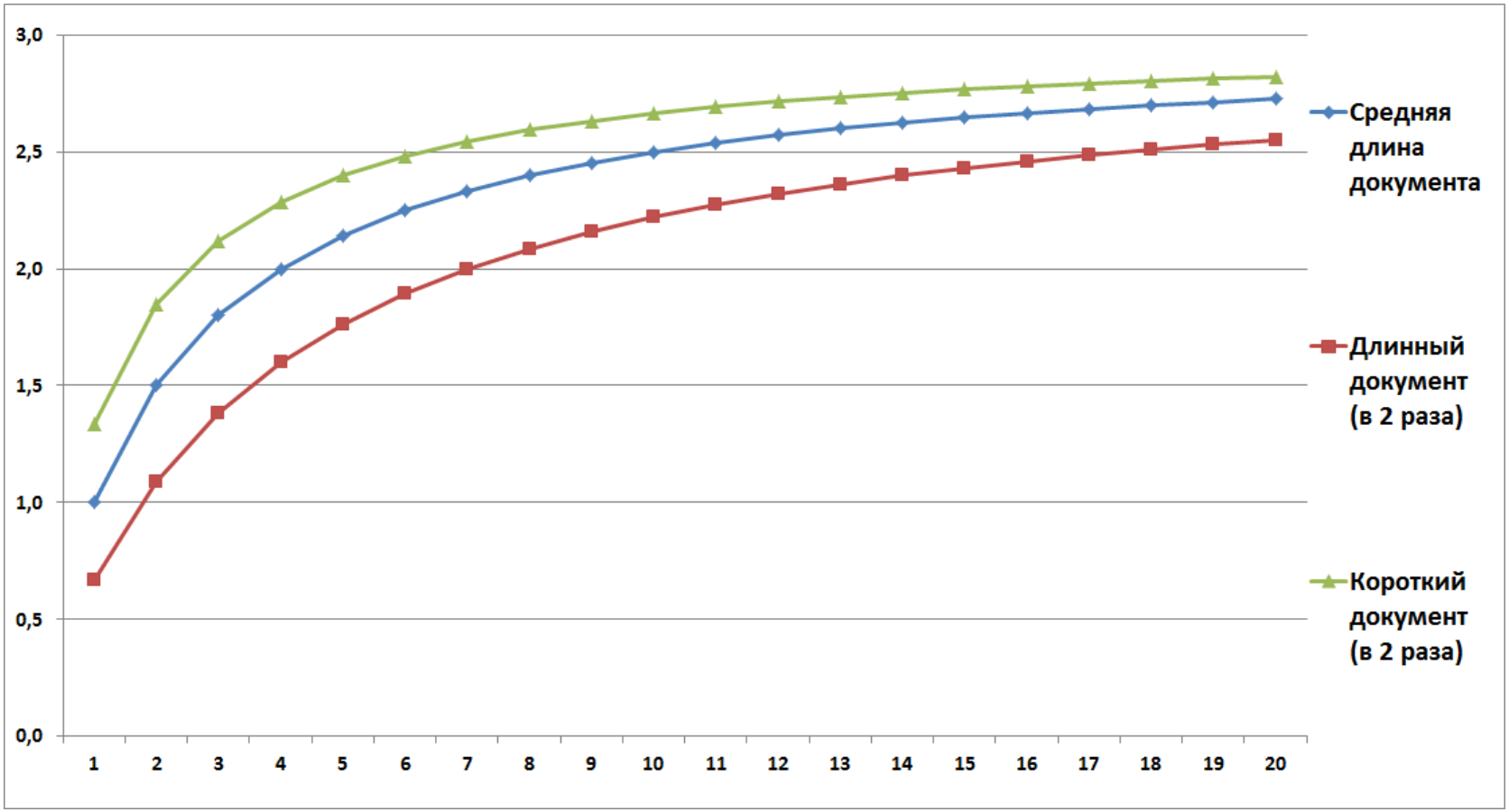
Значимые факторы данного типа

Фактор	Текст	Title
BM25	+	+
Все слова из запроса	+	+
BM25 + ГЕО	+	+

Но данного описания текста совсем недостаточно… Почему?

Насыщение BM25 при росте TF

Зависимость Score от TF (вхождений)



Рекомендации по группе факторов

Простые правила для SEO

1. Больше вхождений — больше релевантность.
2. Насыщение после **6-9 вхождений** (зависит от объема текста, ниже расчет для средней длины документа).
3. Ставка на слова с большим весом.
4. Вхождение всех слов во все зоны документа.

Процентный рост BM25 от числа вхождений

Число вхождений	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Рост в процентах	-	50,0%	20,0%	11,1%	7,1%	5,0%	3,7%	2,9%	2,3%	1,9%	1,5%	1,3%	1,1%	1,0%	0,8%	0,7%	0,7%	0,6%	0,5%	0,5%

Необходимость в прочих факторах

Для ранжирования важно учитывать

1. Взаимное расположение слов из запроса в тексте (относительно друг друга).
2. Словоформу (совпадение формы слова).
3. Позицию слов в документе (относительно начала документа).

Используемая модель «bag-of-words» («мешок слов») не позволяет это учесть. Следовательно — важными становятся и прочие группы факторов.

Фразовые соответствия

Три типа соответствия

1. **Phrase** — все слова из поискового запроса встречаются в документе подряд.
2. **Strict** — все слова из поискового запроса есть в документе с учётом контекстных ограничений.
3. **All** — все остальные найденные документы.

Эволюционно, потребность в соответствии типа «phrase» менялась с числом найденных по нему результатов (в Яндексе).

Учёт близости слов

Значимые факторы данного типа

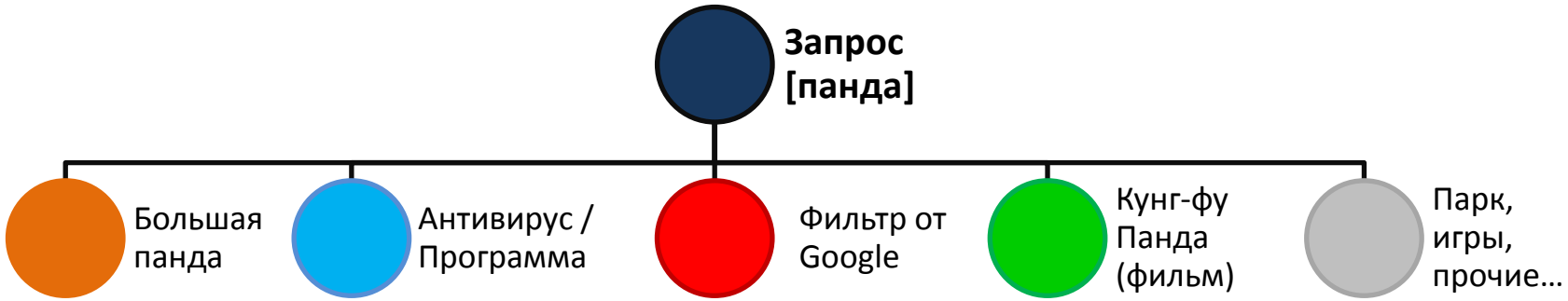
Фактор	Повышенная значимость
Все слова подряд	Устойчивые сочетания, корректные фразы (с точки зрения русского)
Близость к началу документа	Первое предложение/пассаж, первые 15-20% текста
Пары слов (фрагменты запроса)	Для запросов длинной в три и большее число слов

Базовая статья: Алгоритм текстового ранжирования Яндекса на РОМИП-2006 (download.yandex.ru/company/03_yandex.pdf)

Синонимы

Запрос, введенный с поисковую систему

1. **«Переколдовывается»**
2. **Определяются факторы запроса** (в том числе «СНСС» / «без СНСС», язык, геозависимость, основные поисковые интенты и т.д.)

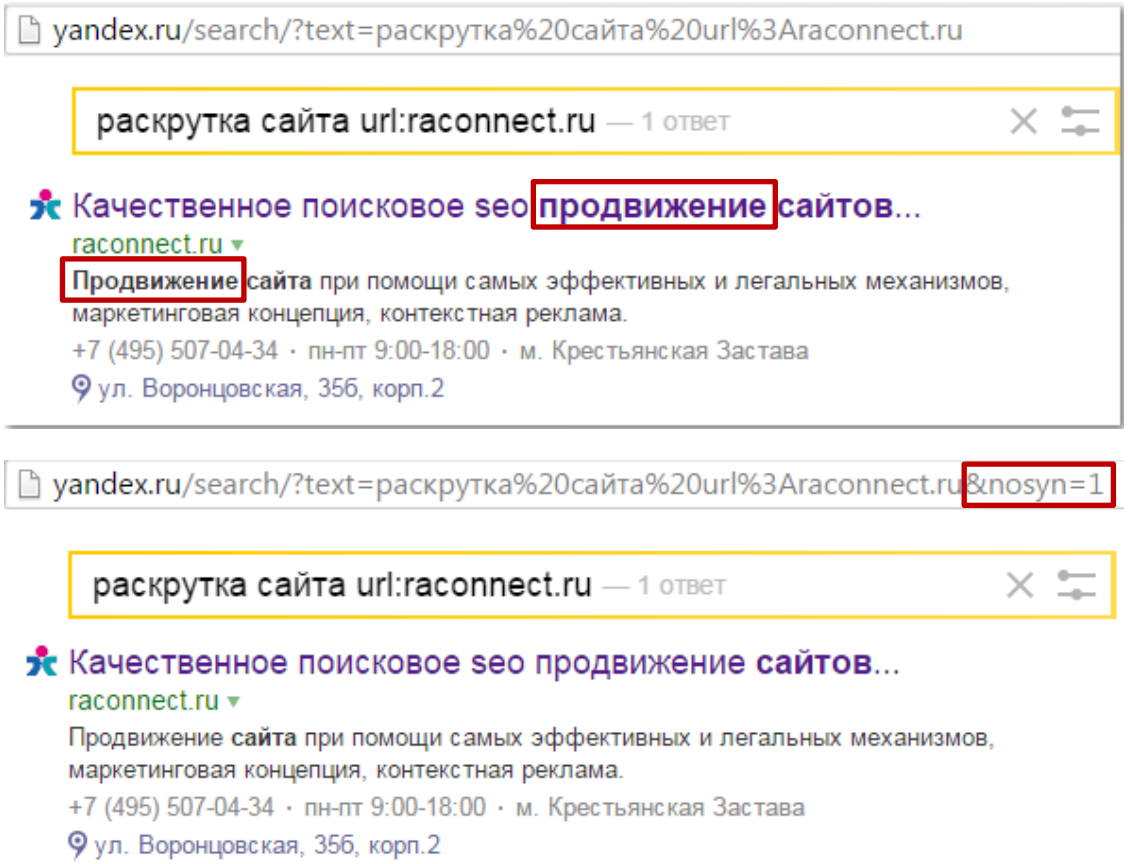


В помощь: pixelplus.ru/samostoyatelno/video/konferencii/analiz-mekhanizma-obrabotki-zaprosa-v-poiskovykh-sistemakh.html

Пиксель Плюс
Интернет-агентство

Синоним или СПЕКТР?

GET-параметр «nosyn=1»



1. Задаем запрос с GET-параметром
2. Пропадает подсветка синонимов
3. Ряд документов исчезает из выдачи (не проходят кворум)
4. СПЕКТР остается

В помощь: slideshare.net/staspolomar/ibc-14

Пиксель Плюс
Интернет-агентство

Учёт синонимов

Значимые факторы данного типа

Фактор	Текст	Title
BM25 с синонимами	+	+
Все слова из запроса (с учётом синонимов)	+	+

Достоверная проверка синонима

1. **Поиск** документа **без слов запроса**, но с **предполагаемым синонимом** [продвижение сайта ~~ раскрутка ~~ seo ~~ москва]
2. **Поиск** по исходному запросу с оператором «url» [раскрутка сайта url:site.ru/dir/] (исключаем «СНСС» с помощью ~~ ываывоатыват)
3. **Находится ли документ?**

Пиксель Плюс
Интернет-агентство

Статические языковые модели

Ответы на вопросы

1. Какова вероятность, что после фразы «Лето это маленькая…» будут идти слова «жизнь», «смерть», «Калининград», …?

Используется для **распознавания речи, переводов**.

2. Если слова «панда» и «поиск» встретились в тексте **три** и **один** раз соответственно, то какова вероятность, что текст посвящён тематикам: «Зоология», «SEO», «Кино», «Софт», …?

Используется в **информационном поиске и тематической классификации документов**.

Тезисы

1. Основные факторы текстового ранжирования.

а) Встречаемость слов и их вес.

- Значимость факторов.
- Рекомендации.

б) Фразовые соответствия.

- Типы соответствия.
- Использование в поиске и на практике.

в) Синонимы.

- Определение.
- Использование в тексте и на сайте.

2. Антиспам -VS- ранжирование.

3. Практические рекомендации по формированию ТЗ для копирайтера.

а) Формирование требований для копирайтера (ТЗ).

б) Основные ошибки, допускаемые оптимизаторами при формировании ТЗ.

4. Рекомендации.

Задачи текстового антиспама

На кого нацелены?

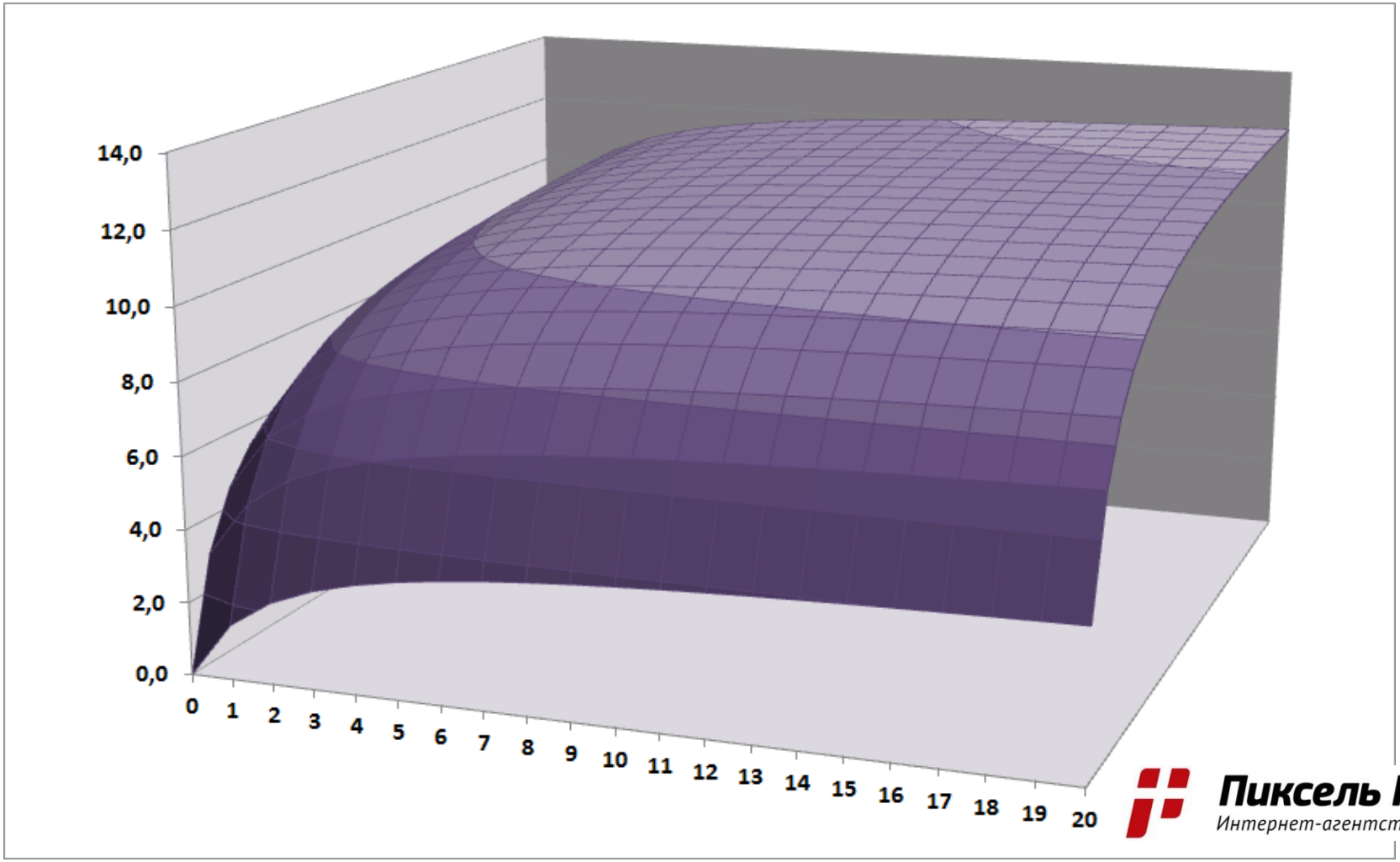
1. **Выявление переоптимизированных текстов** и применение текстовых антиспам пост-фильтров.
2. **Выявление откровенного спама:** машинописный текст, синомайзеры, автоматический перевод, …
3. **Поиск спамных хостов:** неуникальные тексты, …

При проведении работ, SEO-специалисты чаще всего сталкиваются с фильтрами за переоптимизацию контента

В помощь: pixelplus.ru/samostoyatelno/stati/prodvizhenie-saytov/sanktsii-poiskovykh-sistem.html

BM25 в случае с двумя термами

Зависимость Score от TF_1 и TF_2



Антиспам и значение BM25

NN	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	0%	12%	18%	22%	24%	26%	27%	29%	29%	30%	31%	31%	31%	32%	32%	32%	33%	33%	33%	33%	33%
1	24%	37%	43%	46%	49%	51%	52%	53%	54%	54%	55%	55%	56%	56%	57%	57%	57%	57%	57%	58%	58%
2	37%	49%	55%	59%	61%	63%	64%	65%	66%	67%	67%	68%	68%	68%	69%	69%	69%	69%	70%	70%	70%
3	44%	56%	62%	66%	68%	70%	71%	73%	73%	74%	75%	75%	75%	76%	76%	76%	77%	77%	77%	77%	77%
4	49%	61%	67%	71%	73%	75%	76%	77%	78%	79%	79%	80%	80%	81%	81%	81%	81%	82%	82%	82%	82%
5	52%	65%	71%	74%	77%	79%	80%	81%	82%	82%	83%	83%	84%	84%	84%	85%	85%	85%	85%	86%	86%
6	55%	67%	73%	77%	79%	81%	82%	84%	84%	85%	86%	86%	86%	87%	87%	87%	88%	88%	88%	88%	88%
7	57%	69%	75%	79%	81%	83%	85%	86%	86%	87%	88%	88%	88%	89%	89%	89%	90%	90%	90%	90%	90%
8	59%	71%	77%	81%	83%	85%	86%	87%	88%	89%	89%	90%	90%	90%	91%	91%	91%	91%	92%	92%	92%
9	60%	72%	78%	82%	84%	86%	87%	89%	89%	90%	91%	91%	91%	92%	92%	92%	93%	93%	93%	93%	93%
10	61%	73%	79%	83%	86%	87%	89%	90%	90%	91%	92%	92%	93%	93%	93%	93%	94%	94%	94%	94%	94%
11	62%	74%	80%	84%	86%	88%	90%	91%	91%	92%	93%	93%	93%	94%	94%	94%	95%	95%	95%	95%	95%
12	63%	75%	81%	85%	87%	89%	90%	91%	92%	93%	93%	94%	94%	95%	95%	95%	95%	96%	96%	96%	96%
13	64%	76%	82%	86%	88%	90%	91%	92%	93%	94%	94%	95%	95%	95%	96%	96%	96%	96%	97%	97%	97%
14	64%	76%	82%	86%	89%	90%	92%	93%	93%	94%	95%	95%	96%	96%	96%	97%	97%	97%	97%	97%	97%
15	65%	77%	83%	87%	89%	91%	92%	93%	94%	95%	95%	96%	96%	96%	97%	97%	97%	98%	98%	98%	98%
16	65%	77%	84%	87%	90%	91%	93%	94%	95%	95%	96%	96%	97%	97%	97%	98%	98%	98%	98%	98%	99%
17	66%	78%	84%	88%	90%	92%	93%	94%	95%	96%	96%	97%	97%	97%	98%	98%	98%	98%	99%	99%	99%
18	66%	78%	84%	88%	90%	92%	93%	95%	95%	96%	97%	97%	97%	98%	98%	98%	99%	99%	99%	99%	99%
19	66%	79%	85%	88%	91%	93%	94%	95%	96%	96%	97%	97%	98%	98%	98%	99%	99%	99%	99%	100%	100%
20	67%	79%	85%	89%	91%	93%	94%	95%	96%	97%	97%	98%	98%	98%	99%	99%	99%	99%	100%	100%	100%

BM25 для двух слов и соотношения IDF_1/IDF_2=2

Сдвиг границ антиспама

NN	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	0%	12%	18%	22%	24%	26%	27%	29%	29%	30%	31%	31%	31%	32%	32%	32%	33%	33%	33%	33%	33%
1	24%	37%	43%	46%	49%	51%	52%	53%	54%	54%	55%	55%	56%	56%	57%	57%	57%	57%	57%	58%	58%
2	37%	49%	55%	59%	61%	63%	64%	65%	66%	67%	67%	68%	68%	68%	69%	69%	69%	69%	70%	70%	70%
3	44%	56%	62%	66%	68%	70%	71%	73%	73%	74%	75%	75%	75%	76%	76%	76%	77%	77%	77%	77%	77%
4	49%	61%	67%	71%	73%	75%	76%	77%	78%	79%	79%	80%	80%	81%	81%	81%	81%	82%	82%	82%	82%
5	52%	65%	71%	74%	77%	79%	80%	81%	82%	82%	83%	83%	84%	84%	84%	85%	85%	85%	85%	86%	86%
6	55%	67%	73%	77%	79%	81%	82%	84%	84%	85%	86%	86%	86%	87%	87%	87%	88%	88%	88%	88%	88%
7	57%	69%	75%	79%	81%	83%	85%	86%	86%	87%	88%	88%	88%	89%	89%	89%	90%	90%	90%	90%	90%
8	59%	71%	77%	81%	83%	85%	86%	87%	88%	89%	89%	90%	90%	90%	91%	91%	91%	91%	92%	92%	92%
9	60%	72%	78%	82%	84%	86%	87%	89%	89%	90%	91%	91%	91%	92%	92%	92%	93%	93%	93%	93%	93%
10	61%	73%	79%	83%	86%	87%	89%	90%	90%	91%	92%	92%	93%	93%	93%	93%	94%	94%	94%	94%	94%
11	62%	74%	80%	84%	86%	88%	90%	91%	91%	92%	93%	93%	93%	94%	94%	94%	95%	95%	95%	95%	95%
12	63%	75%	81%	85%	87%	89%	90%	91%	92%	93%	93%	94%	94%	95%	95%	95%	95%	96%	96%	96%	96%
13	64%	76%	82%	86%	88%	90%	91%	92%	93%	94%	94%	95%	95%	95%	96%	96%	96%	96%	97%	97%	97%
14	64%	76%	82%	86%	89%	90%	92%	93%	93%	94%	95%	95%	96%	96%	96%	97%	97%	97%	97%	97%	97%
15	65%	77%	83%	87%	89%	91%	92%	93%	94%	95%	95%	96%	96%	96%	97%	97%	97%	98%	98%	98%	98%
16	65%	77%	84%	87%	90%	91%	93%	94%	95%	95%	96%	96%	97%	97%	97%	98%	98%	98%	98%	98%	99%
17	66%	78%	84%	88%	90%	92%	93%	94%	95%	96%	96%	97%	97%	97%	98%	98%	98%	98%	99%	99%	99%
18	66%	78%	84%	88%	90%	92%	93%	95%	95%	96%	97%	97%	97%	98%	98%	98%	99%	99%	99%	99%	99%
19	66%	79%	85%	88%	91%	93%	94%	95%	96%	96%	97%	97%	98%	98%	98%	99%	99%	99%	99%	100%	100%
20	67%	79%	85%	89%	91%	93%	94%	95%	96%	97%	97%	98%	98%	98%	99%	99%	99%	99%	100%	100%	100%

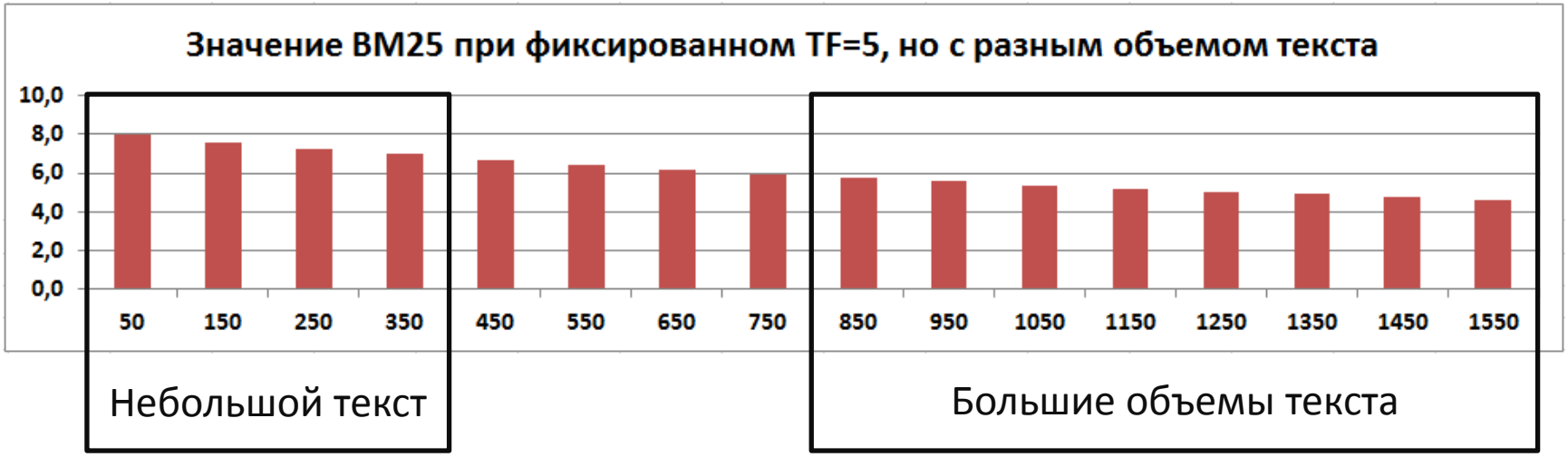
BM25 для двух слов и соотношения IDF_1/IDF_2=2

Пиксель Плюс
Интернет-агентство

BM25 и объем текста

Зависимость Score от объема текста

Объем текста (слов)	50	150	250	350	450	550	650	750	850	950	1050	1150	1250	1350	1450	1550
Падение в процентах	-	-4,6%	-4,4%	-4,2%	-4,1%	-3,9%	-3,8%	-3,6%	-3,5%	-3,4%	-3,3%	-3,2%	-3,1%	-3,0%	-2,9%	-2,8%



Фиксированная частота вхождений равная пяти

Пиксель Плюс
Интернет-агентство

Почему «портянки рулили»?

Базовый расчет для двух типов текста

Небольшой текст

- 2% вхождений ключа
- 350 слов ≈ 2 300 символов
- IDF = 3

Итоговое значение

Score = 7,45 (≈ -6%)

Классическая «портянка»

- 2% вхождений ключа
- 3000 слов ≈ 20 000 символов
- IDF = 3

Итоговое значение

Score = 7,86 (≈ +6%)

Если запрос из двух слов, то профит в факторе составит уже 12%

Рекомендации по анализу

Простые правила для SEO

1. Вхождение всех слов — обязательно.
2. «Подспамливать» лучше более «тяжелым» словом, это эффективней (таблица на предыдущем слайде).
3. При числе вхождений 6-9 — итоговые значения близки к оптимальным.
4. Можно сократить число необходимых вхождений путём урезания объема текста.
5. Пороги антиспам-фильтров постоянно сдвигаются — лучше немного недобрать, что немного перебрать.

Тезисы

1. Основные факторы текстового ранжирования.

а) Встречаемость слов и их вес.

- Значимость факторов.
- Рекомендации.

б) Фразовые соответствия.

- Типы соответствия.
- Использование в поиске и на практике.

в) Синонимы.

- Определение.
- Использование в тексте и на сайте.

2. Антиспам -VS- ранжирование.

3. Практические рекомендации по формированию ТЗ для копирайтера.

а) Формирование требований для копирайтера (ТЗ).

б) Основные ошибки, допускаемые оптимизаторами при формировании ТЗ.

4. Рекомендации.

Управления факторами

Основные факторы для управления

1. Объем текста
2. Процент вхождений каждого слова
3. Процент вхождений каждого из синонимов
4. Вхождение фраз и из морфология
5. Структура и оформление текста: наличие списков, картинок, подзаголовков и т.д.
6. Наличие специализированных терминов, задающих тематику и их количество
7. Распределение ключевых фраз по объему

Сбор данных для ТЗ

Данные на вход

- **Продвигаемые запросы**: [офисная мебель] + СЧ
- **Регион и стоп-слова** (служебные части речи)

Данные на выходе

Уникальные слова из запросов

```
офисный
мебель
магазин
интернет
купить
продажа
```

Слова из подсветки в выдаче

```
офис
москва
цена
у
покупать
заказ
б
каталог
бу
mebel
ofisnaya
доставка
прайс
заказывать
```

Слова, задающие тематику

```
компания
кресло
предлагать
персонал
недорого
производство
выбор
выгодный
качественный
качество
недорогой
кабинет
руководитель
производитель
```

Формирование задания

Основные параметры

1. **Объем текста:** от 800 знаков, достаточный для вхождения нужного количества ключевых фраз (без превышения порогов по спаму).
2. **Ограничение на использование слов «сверху»**: задаётся ограничение использования тех слов, которыми может быть переспамлен текст. Ограничение на уровне ≈ 2% от планируемого количества слов в тексте **(объем без пробелов / разделить на 6,5).**
3. **Минимальное использование слов**: ограничивается число вхождений каждого слова «снизу». Задавать его требуется с помощью ключевых фраз (затребовать использование нужного количества каждого из слов). В этом случае, все слова будут употреблены в нужном виде, последовательности.
4. **Структура текста**: требуется заранее определять основную мысль текста и его структуру. Опираться на: семантику, слова, задающие тематику.
5. **Пост-проверка текста**: проверка на основные ошибки и соблюдение ТЗ.

Помощь: slideshare.net/SergeyYurkov1/ss-46831161

Основные ошибки

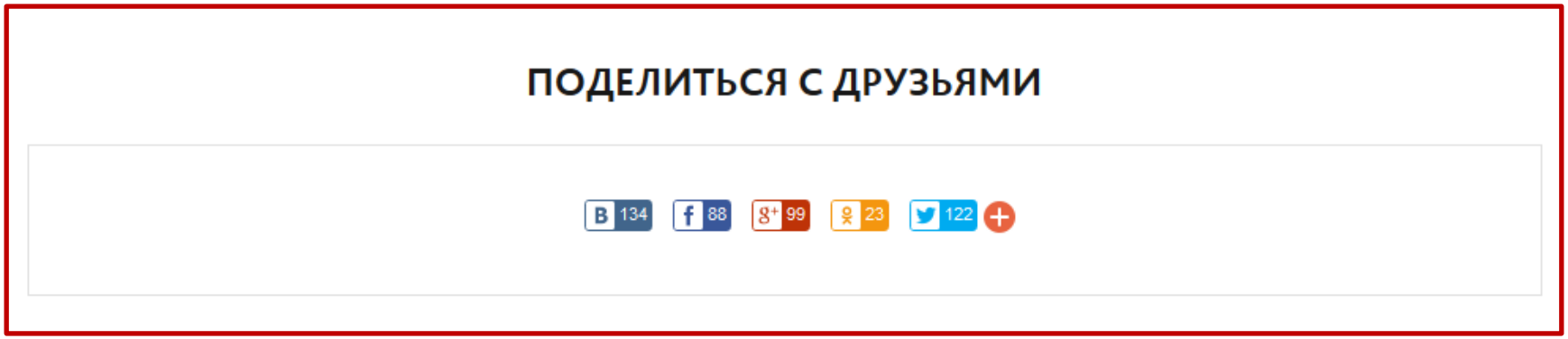
Копирайтер и/или поиск не будет вас любить, если

1. **Слишком много** ключевых фраз.
2. **Слишком мало** ключевых фраз (слова из продвигаемых запросов, встречаются менее 3 раз).
3. **Корявые** ключевые фразы.
4. **В ключевых фразах** отсутствию синонимы слов из запроса и топоним (для ГЗ фраз).
5. **Очень сложное** задание.
6. **Не заданы тематические** слова.
7. **Структура текста** не продумана.

Шаблон ТЗ для копирайтера

1. **Посещение раздела или детальной страницы: http://www.pixelplus.ru/samostoyatelno/**

2. **«Лайк» в любой социальной сети на выбор.**



3. **Скидываем принт-скрин на почту: seo@pixelplus.ru с фразой «BALTIC DIGITAL DAYS 2015» в теме письма.**

4. **Получаем ответным письмом: пример шаблона ТЗ для копирайтера с рекомендациями.**

Пиксель Плюс
Интернет-агентство

Тезисы

1. Основные факторы текстового ранжирования.

а) Встречаемость слов и их вес.

- Значимость факторов.
- Рекомендации.

б) Фразовые соответствия.

- Типы соответствия.
- Использование в поиске и на практике.

в) Синонимы.

- Определение.
- Использование в тексте и на сайте.

2. Антиспам -VS- ранжирование.

3. Практические рекомендации по формированию ТЗ для копирайтера.

а) Формирование требований для копирайтера (ТЗ).

б) Основные ошибки, допускаемые оптимизаторами при формировании ТЗ.

4. Рекомендации.

Рекомендации

1. Для вывода конкурентных запросов в ТОП требуется обязательно производить «тонкую юстировку» текстовых факторов.
2. При размещении текста и его написании — не забывайте про здравый смысл и пользователей.
3. Не превышайте ≈ 2% вхождений и не раздувайте объем (≤ 550 слов для коммерческих запросов), это важно, так как фиксируются регулярные сдвиги порогов срабатывания антиспам-фильтров.
4. Важно понимать, что поиск оперирует более чем 50 факторами связанными с текстом.
5. Текст требуется проверять на корректность и соответствие требованиям SEO.

14 августа 2015
BALTIC DIGITAL DAYS

Максимум текстовой релевантности сегодня: факторы, практические рекомендации


manager@pixelplus.ru, seo@pixelplus.ru


Отдел продаж: +7 (495) 989-53-11


Основной офис в Москве:
г. Москва, ул. Шаболовка, дом 34

BalticDigitalDays

Пиксель Плюс
Интернет-агентство