

Бег в мешке

Ускорение поиска на основе мешка слов



Олег Ануфриев





О себе

- Ануфриев Олег
- Сервис Data Science / NLP
- Работаю в 2ГИС с 2017
- До этого писал на JAVA и немного — на Python



Вот про эту ссылочку...

СпортЕВ 
Магазин товаров для спорта и отдыха
★★★★★ 4.1 62 оценки

Спортинг. Тренажёры.
Спортинг. Велосипеды. Экипировка

[Посмотреть](#)

[←](#) [Инфо](#) [Отзывы 62](#) [Цены 12148](#) [Фото 34](#)

 Улица Геодезическая, 23
цокольный этаж
Ленинский район, Новосибирск,
630087

[Показать вход](#)

 Ежедневно с 10:00 до 19:00
Закрето. Откроется завтра в
10:00

 +7(383)3 [Показать телефон](#)

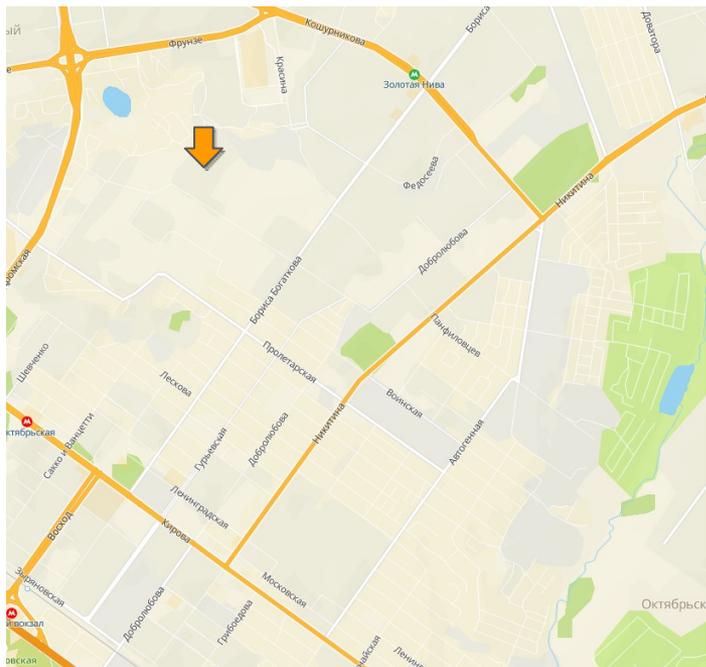
Площадь Маркса
Вход на станцию метро
Башня ж/м, Ленинский район, Новосибирск

 Ленинская линия



Задача дедупликации

- Есть товары в базе, их много (миллионы)

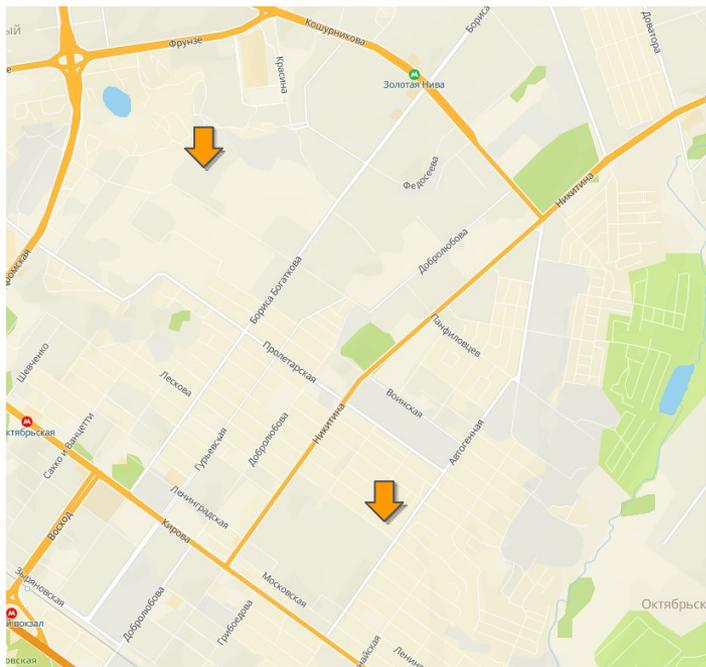


id	name
2489753	Аквилон (финишная планка) СтарыйДуб ЦМК-20x40x1250
2332017	Насос вибрацион. Ручеек-1 15м г.Могилев верх.забор
2074387	Герметик Момент силиконовый для окон и стекол, белый (280 мл)
1957561	Душевой набор GROHE Power&Soul 4 режима, хром (27742000)
1966102	Смеситель Hansgrohe Ecostat 1001 CL 13211000
2305617	Смеситель для кухни Milardo Amur
2489755	Уралкерамика Петра ПГЗПЕ404 Напольная плитка 41,8x41,8
2489756	CERESIT CE 40 ГРУППА №1 эластичная затирка для швов с антигрибковым эффектом
2489757	Паркетная доска Barlinek Multi Дуб Tiramisu 2200x207x14 мм
2489758	Винт 2,5x30 DIN 7985 TORX с полукруглой головкой и с шестирядиусным шлицем (TORX)
2489759	Клавиша смыва Migliore OTTONE CLASSIC золото, отверстие для ручки 145x217мм, модель 27.053.DO
2489760	Душевая стойка Timo Helmi SX-1070/00 412
1934372	Душевой набор 130 9 5 л мин Grohe Power&Soul (27742000)
2489762	SE Unica Беж Регулятор тёплого пола 10А, с датчиком (без функц. откл.)
2332034	Дрель Hyundai D 350
2489763	Штуцер МастерПроф ИС.070656
2332037	Проточный водонагреватель Electrolux Smartfix 2.0 TS (6,5 kW) - кран+душ
2489764	Полка Am.Pm Bliss A5534764
2489765	Полотенцесушитель Флагман "Лесенка" изогнутая с полочкой 600x600 нержавеющая сталь, подкл. D
2489766	Грунт-эмаль по ржавчине Parade Z1 Грунт-эмаль по ржавчине гладкая PARADE Z1 RAL Вишневый
2332046	Уголок крепежный равносторонний 80x80x40x2
2489767	RAVAK Дверь душевая Ravak BLDP4 - 140 белый + Транспарент
2489768	Смеситель для кухни Билл Билла Смарт-электрический с сенсорным управлением



Задача дедупликации

- Есть товары в базе, их много (миллионы)
- Нужно узнать, если ли у нас такой товар



id	name
2489753	Аквилон (финишная планка) СтарыйДуб ЦМК-20x40x1250
2332017	Насос вибрацион. Ручеек-1 15м г.Могилев верх.забор
2074387	Герметик Момент силиконовый для окон и стекол, белый (280 мл)
1957561	Душевой набор GROHE Power&Soul 4 режима, хром (27742000)
1966102	Смеситель Hansgrohe Ecostat 1001 CL 13211000
2305617	Смеситель для кухни Milardo Amur
2489755	Уралкерамика Петра ПГЗПЕ404 Напольная плитка 41,8x41,8
2489756	CERESIT CE 40 ГРУППА №1 эластичная затирка для швов с антигрибковым эффектом
2489757	Паркетная доска Barlinek Multi Дуб Tiramisu 2200x207x14 мм
2489758	Винт 2,5x30 DIN 7985 TORX с полукруглой головкой и с шестирядиусным шлицем (TORX)
2489759	Клавиша смыва Migliore OTTONE CLASSIC золото, отверстие для ручки 145x217мм, модель 27.053.DO
2489760	Душевая стойка Timo Helmi SX-1070/00 412
1934372	Душевой набор 130 9 5 л мин Grohe Power&Soul (27742000)
2489762	SE Unica Беж Регулятор тёплого пола 10А, с датчиком (без функц. откл.)
2332034	Дрель Hyundai D 350
2489763	Штуцер МастерПроф ИС.070656
2332037	Проточный водонагреватель Electrolux Smartfix 2.0 TS (6,5 kW) - кран+душ
2489764	Полка Am.Pm Bliss A5534764
2489765	Полотенцесушитель Флагман "Лесенка" изогнутая с полочкой 600x600 нержавеющая сталь, подкл. D
2489766	Грунт-эмаль по ржавчине Parade Z1 Грунт-эмаль по ржавчине гладкая PARADE Z1 RAL 7016 Вишневый
2332046	Уголок крепежный равносторонний 80x80x40x2
2489767	RAVAK Дверь душевая Ravak BLDP4 - 140 белый + Транспарент
2489768	Смеситель для ванны Grohe Power&Soul 4 режима поворотный хромированный Grohe Power&Soul 4 режима поворотный хромированный



Задача дедупликации

- Есть товары в базе, их много (миллионы)
- Нужно узнать, если ли у нас такой товар
- Записаны названия товаров могут быть по-разному

Volle 13-01-223 VOLLE умывальник 585*370*125мм, накладной 13-01-223

Раковина Volle 59 см 13-01-223

Умывальник Волле 13-01-223, 585*370*125 мм, цвет белый

Умывальник прямоугольный накладной VOLLE 13-01-223,Испания

раковина VOLLE 58,5 13-01-223



Задача дедупликации

- Есть товары в базе, их много (миллионы)
- Нужно узнать, если ли у нас такой товар
- Записаны названия товаров могут быть по-разному
- Таких запросов много
- Есть модель сравнения...

10^7

...

10^0



Задача дедупликации

- Есть товары в базе, их много (миллионы)
- Нужно узнать, если ли у нас такой товар
- Записаны названия товаров могут быть по-разному
- Таких запросов много
- Есть модель сравнения... качественная, но медленная

10^7

...

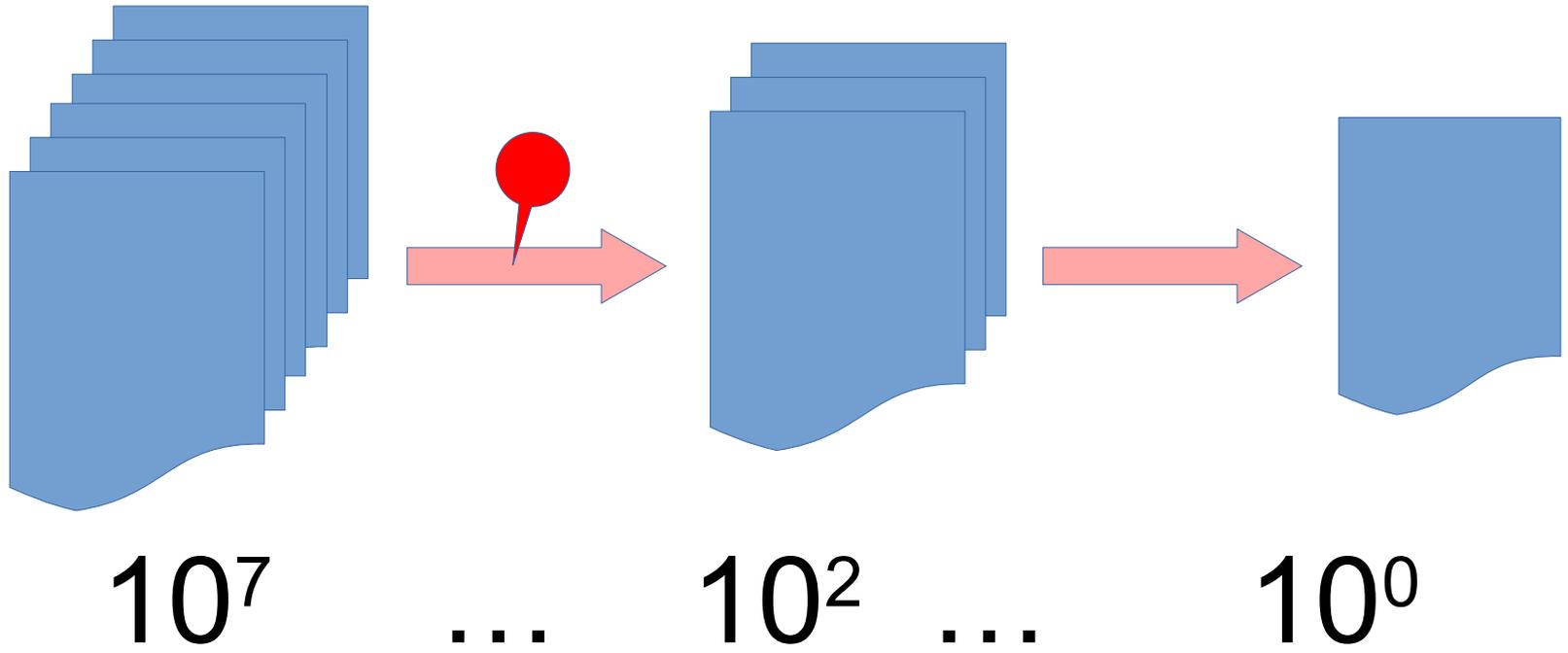
10^2

...

10^0



Задача дедупликации





Имеющиеся данные

- Они есть
- Данных довольно много
- Очень качественно размечены
- Довольно репрезентативны
- Содержат разнообразные (во всех смыслах) строки и токены
- Есть формы слова, сокращения, транслит, ошибки и опечатки



Первые шаги

- Метрика - попадание верного варианта в top100
- Типа kaggle
- Мешок слов
- TF/IDF и libMinHashGPU





Первые шаги

- Метрика - попадание верного варианта в top100
- Типа kaggle
- Мешок слов
- TF/IDF и libMinHashGPU

Лампа 99Вт E27	
лампа накаливания E27 99Вт	
40Вт лампа E27 Горсветторг	
...	...





Первые шаги

- Метрика - попадание верного варианта в top100
- Типа kaggle
- Мешок слов
- TF/IDF и libMinHashGPU

Лампа 99Вт E27	
лампа накаливания E27 99Вт	0.80
40Вт лампа E27 Горсветторг	0.55
...	...





Начальный вариант (самый первый)

- Токенизация и нормализация
Красноголовка Pepsi 0.25 с анисом



Начальный вариант (самый первый)

- Токенизация и нормализация
Красноголовка Pepsi 0.25 с анисом
- Обогащение (транслит)
красноголовка pepsi **пепси** 0 25 анисом



Начальный вариант (самый первый)

- Токенизация и нормализация

Красноголовка Pepsi 0.25 с анисом

- Обогащение (транслит)

красноголовка pepsi **пепси** 0 25 анисом

- Веса токенов (почти TF/IDF)

красноголовка:**0.8** pepsi:0.7 пепси:0.75 0:0.1 25:0.3 анисом:**0.9**



Начальный вариант (самый первый)

- Токенизация и нормализация

Красноголовка Pepsi 0.25 с анисом

- Обогащение (транслит)

красноголовка pepsi **пепси** 0 25 анисом

- Веса токенов (почти TF/IDF)

красноголовка:**0.8** pepsi:0.7 пепси:0.75 0:0.1 25:0.3 анисом:**0.9**

- Слияние списков, ограничения (over 9000 → /dev/null) вместо стоп-слов

- **98% попадания в top100**



Начальный вариант (самый второй)

- Токенизация и нормализация - с морфологией
красноголовка pepsi 0.25 с анисом



Начальный вариант (самый второй)

- Токенизация и нормализация - с морфологией префиксы
красноголовка pepsi 0.25 с анисом



Начальный вариант (самый второй)

- Токенизация и нормализация - с морфологией префиксы
красноголовка pepsi 0.25 с анисом
- Обогащение (транслит, до префикса)
красн pepsi **пепси** 0 25 с анисом



Начальный вариант (самый второй)

- Токенизация и нормализация - с морфологией префиксы
красноголовка pepsi 0.25 с анисом
- Обогащение (транслит, до префикса и цифры)
красн pepsi **пепси** 0 25 **250** с анисом



Начальный вариант (самый второй)

- Токенизация и нормализация - с морфологией префиксы
красноголовка pepsi 0.25 с анисом
- Обогащение (транслит, до префикса и цифры)
красн pepsi **пепси** 0 25 **250** с анисом
- Веса - рус лат циф - разные; для 1 символа - тоже



Начальный вариант (самый второй)

- Токенизация и нормализация - с морфологией префиксы
красноголовка pepsi 0.25 с анисом
- Обогащение (транслит, до префикса и цифры)
красн pepsi **пепси** 0 25 **250** с анисом
- Веса - рус лат циф - разные; для 1 символа - тоже
- Веса с учётом совместности
Tuti обл/плитка бежевая (TGM 011 D)
Tuti Бежевый TGM 011 D 25 x 35



Начальный вариант (самый второй)

- Токенизация и нормализация - с морфологией префиксы
красноголовка pepsi 0.25 с анисом
- Обогащение (транслит, до префикса и цифры)
красн pepsi **пепси** 0 25 **250** с анисом
- Веса - рус лат циф - разные; для 1 символа - тоже
- Веса с учётом совместности
Tuti обл/плитка **бежевая** (**TGM 011 D**)
Tuti Бежевый **TGM 011 D** 25 x 35
- **99.8%** попадания в top100



Несколько месяцев* счастливого неведения

- Документация написана, примеры, помощь в реализации
- Другая команда пишет реализацию
- Мы пилим другое, улучшая и размножая модель дедупликации
- Датасеты собираются
- **Возникла задача повышения качества кандидатов**

Что бы могло пойти не так?

*) примерно два года



Реализованная версия :: Первая кровь

- Реализовано не всё (версия 1)
 - Ест много памяти
 - Ест чрезмерно много памяти
 - Сложно проверить качество
- ...“1 запрос выполняется ~20-30 мин.”
- 30 млн. товаров не влезли в 30Гб





Реализованная версия :: Первая кровь

- Реализовано не всё (версия 1)
- Ест много памяти
- Ест чрезмерно много памяти
- Сложно проверить качество
- ...“1 запрос выполняется ~20-30 мин.”
- 30 млн. товаров не влезли в 30Гб
- Новый dataset - 30 000 000





Изменения и сомнения (пока не всерьёз)

- Откопали прототип - зачем, кстати?
- 1.5 RPS на 30млн - это серьёзно?
- Что с ним не так - в основном python
- Кстати, уберём ограничение на over 9000
- Перепишем алгоритм
(но это будет только рефакторинг)





Слияние списков

Ищем “Лампа 99Вт E27”

1	лампа 99Вт E27	0.85
2	лампа накаливания E27 99Вт	0.85*
3	40Вт лампа E27 Горсветторг	0.65
4	паяльник 99Вт криптотул	0.38
...

лампа	0.22
99	0.2
вт	0.18
е	0.1
27	0.15

лампа **1 2 3 7 12 44 43**

99 **1 2 4 7 11 72 101 121**

вт **1 2 3 4 12 43 121**

е **1 2 3 12 13 33 101 102**

27 **1 2 3 13 41 43 96 121**

$$1 - 0.22 + 0.2 + 0.18 + 0.1 + 0.15$$

$$2 - 0.22 + 0.2 + 0.18 + 0.1 + 0.15$$

$$3 - 0.22 + 0.18 + 0.1 + 0.16$$

$$4 - 0.2 + 0.18$$



Изменения (пока не всерьёз)

- Слияние списков итерация по спискам
- Сортировка построение top100
- Битовые индексы (для частотных токенов)
- Уникальные токены
- Выделим сущности, вспомним ООП



Итерация по спискам

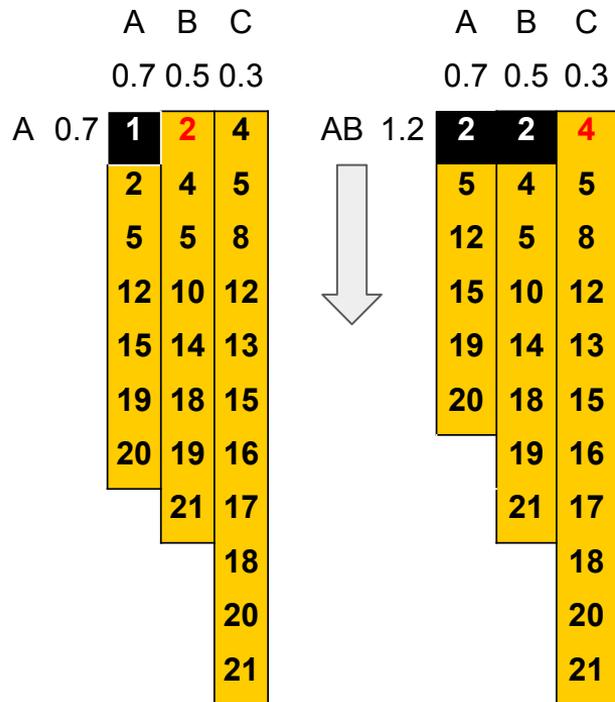
A B C
0.7 0.5 0.3

A	0.7	1	2	4
		2	4	5
		5	5	8
		12	10	12
		15	14	13
		19	18	15
		20	19	16
			21	17
				18
				20
				21

#	w	txt
1	0.7	A



Итерация по спискам



#	w	txt
2	1.2	AB
1	0.7	A

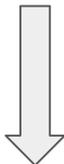


Итерация по спискам

	A	B	C
	0.7	0.5	0.3
A 0.7	1	2	4
	2	4	5
	5	5	8
	12	10	12
	15	14	13
	19	18	15
	20	19	16
		21	17
			18
			20
			21

	A	B	C
	0.7	0.5	0.3
AB 1.2	2	2	4
	5	4	5
	12	5	8
	15	10	12
	19	14	13
	20	18	15
		19	16
		21	17
			18
			20
			21

	A	B	C
	0.7	0.5	0.3
BC 0.8	5	4	4
	12	5	5
	15	10	8
	19	14	12
	20	18	13
		19	15
		21	16
			17
			18
			20
			21



#	w	txt
2	1.2	AB
4	0.8	BC
1	0.7	A



Итерация по спискам

A B C
0.7 0.5 0.3

A 0.7

1	2	4
2	4	5
5	5	8
12	10	12
15	14	13
19	18	15
20	19	16
	21	17
		18
		20
		21

A B C
0.7 0.5 0.3

AB 1.2

2	2	4
5	4	5
12	5	8
15	10	12
19	14	13
20	18	15
	19	16
	21	17
		18
		20
		21

A B C
0.7 0.5 0.3

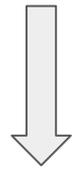
BC 0.8

5	4	4
12	5	5
15	10	8
19	14	12
20	18	13
	19	15
	21	16
		17
		18
		20
		21

A B C
0.7 0.5 0.3

ABC 1.5

5	5	5
12	10	8
15	14	12
19	18	13
20	19	15
	21	16
		17
		18
		20
		21



#	w	txt
5	1.5	ABC
2	1.2	AB
4	0.8	BC
1	0.7	A



Итерация по спискам

A 0.7

A	B	C
0.7	0.5	0.3
1	2	4
2	4	5
5	5	8
12	10	12
15	14	13
19	18	15
20	19	16
	21	17
		18
		20
		21

AB 1.2

A	B	C
0.7	0.5	0.3
2	2	4
5	4	5
12	5	8
15	10	12
19	14	13
20	18	15
	19	16
	21	17
		18
		20
		21

BC 0.8

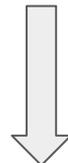
A	B	C
0.7	0.5	0.3
5	4	4
12	5	5
15	10	8
19	14	12
20	18	13
	19	15
	21	16
		17
		18
		20
		21

ABC 1.5

A	B	C
0.7	0.5	0.3
5	5	5
12	10	8
15	14	12
19	18	13
20	19	15
	21	16
		17
		18
		20
		21

C 0.3

A	B	C
0.7	0.5	0.3
12	10	8
15	14	12
19	18	13
20	19	15
	21	16
		17
		18
		20
		21



#	w	txt
5	1.5	ABC
2	1.2	AB
4	0.8	BC
1	0.7	A
8	0.3	C



Итерация по спискам - не искать ненужного

A B C D E F
 1.0 0.8 **0.4** 0.3 0.2 0.1

B..	0.6..	15	14	14	14	15	14
		18	15	25	19	16	15
		25	19	28	20	17	..
		29	21	52	24
		84	38	53	31
			69	65
			91	71
				78
				81
				90
			

#	w	txt
..
..
..
7	2.0	ABE
1	1.9	ABF
12	1.7	ACD
8	1.6	BCDEF





Итерация по спискам - не искать ненужного

A B C D E F
 1.0 0.8 **0.4** 0.3 0.2 0.1

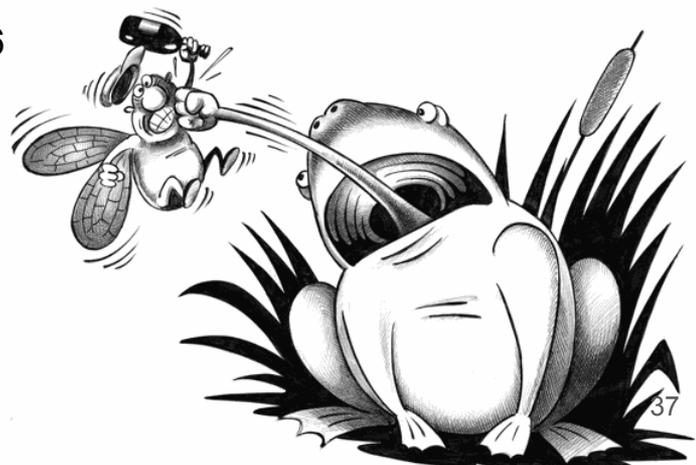
B..	0.6..	15	14	14	14	15	14
		18	15	25	19	16	15
		25	19	28	20	17	..
		29	21	52	24
		84	38	53	31
			69	65
			91	71
				78
				81
				90
			

#	w	txt
..
..
..
7	2.0	ABE
1	1.9	ABF
12	1.7	ACD
8	1.6	BCDEF

$$0.6 + 0.4 + 0.3 + 0.2 + 0.1 = 1.6$$

1.6 < **1.6** ? НЕТ!

НЕ ГОДИТСЯ!





Битовые индексы

1	2	4	5	6	11	12	14	15	18	20		
---	---	---	---	---	----	----	----	----	----	----	--	--

На каждое вхождение — 32 бита (или 64)

0000	0000	0001	0100	1101	1000	0110	0110		
------	------	------	------	------	------	------	------	--	--

На каждое вхождение — 1 бит



Битовые индексы

1	2	4	5	6	11	12	14	15	18	20		
---	---	---	---	---	----	----	----	----	----	----	--	--

На каждое вхождение — 32 бита (или 64)

0000	0000	0001	0100	1101	1000	0110	0110		
------	------	------	------	------	------	------	------	--	--

На каждое вхождение — 1 бит

- Вхождений * 32 > Документов — битовые индексы
- Вхождений * 32 < Документов — списки



Уникальные токены

- SQT ЛВК и т.д. - встречаются 1 раз
- Однозначно указывают на документ

Токен	doc_id
ЛВК	301
SQTV	414
HDXT	337
ОРХЦ	901



token_id	doc_id
3801	301
3972	414
4041	337
4392	901



Изменения (гоняемся за производительностью)

- Cython или скромное очарование противоестественного
- `array.array` лучше, чем `List`
- А для словаря возьмём `microdict`... ooops!
- Кто ещё ест память?
- Вышло порядка порядка по скорости (13 RPS)
- Добавление без перестроения индекса





В соседнем дворе тоже выживали как могли



- Staging ждать не может
- Версия на ElasticSearch
- Быстрее Cython (20 RPS на 30Млн)
- *Несколько* другой алгоритм
- ...и потом ещё TF/IDF
- медленно и неточно



Несовпадение результатов

(всё на top100)	10млн	30млн
ES простой	75.58	55.58
ES TF/IDF	95.58	91.26
Наше (over 9000)	96.86	93.15
Наше (+совместность)	97.95	95.17

Прототип	top1	top10	top100
Dataset1	78.44	98.03	99.95
Dataset2	79.97	97.41	99.85





Почему бы не C/C++

- Cython это не очень-то и быстро
- Cython - достаточно извращённая технология
(нет даже подсветки синтаксиса)
- GIL всё там же
- Код уже писался с низкоуровневыми штуками “в уме”



Результаты C++ (три порядка это три порядка)



- Да, это всё ещё рефакторинг!
- Код читается лучше
- GIL не влияет
- 250 RPS на 30Млн, 500 - на 10
- * 4 thread 750 и 1500 соответственно
- ...а парсинг (python) параллельно
- В 32 гига поместится до 300млн.
товаров



Код

```
void QueryBatch::join() {  
    if (!threads.size()){  
        return;  
    }  
    for (auto& thread: threads) {  
        thread.join();  
    }  
    threads.clear();  
}
```



Итоги

- Датасет был сфранкенштейнен
- Добавленная часть была “неверно размечена”
- Нельзя использовать число токенов (leak) – всё ещё не **TF/IDF**
- Качество этим датасетом не измерить, соответственно и не **повысить**



Итоги

- От порядка в сортировке зависит результат
- Некоторые ошибки повышали и качество и скорость
- Скорость зависит от качества (лучше разделяем подходящее от неподходящего – быстрее пропускаем лишнее)



Все рады новой версии

- Не такое надёжное?
- Update
- Delete
- Сервис





Спасибо за внимание

- Олег Ануфриев, 2ГИС
- ol.anufriev@2gis.ru



2ГИС

- Отдельное спасибо за иллюстрации - Сергей Корсун