НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Анализ русскоязычных текстов в СУБД MongoDB

Дубов Михаил Сергеевич

Научно-учебная лаборатория
интеллектуальных систем и структурного анализа, НИУ ВШЭ

План

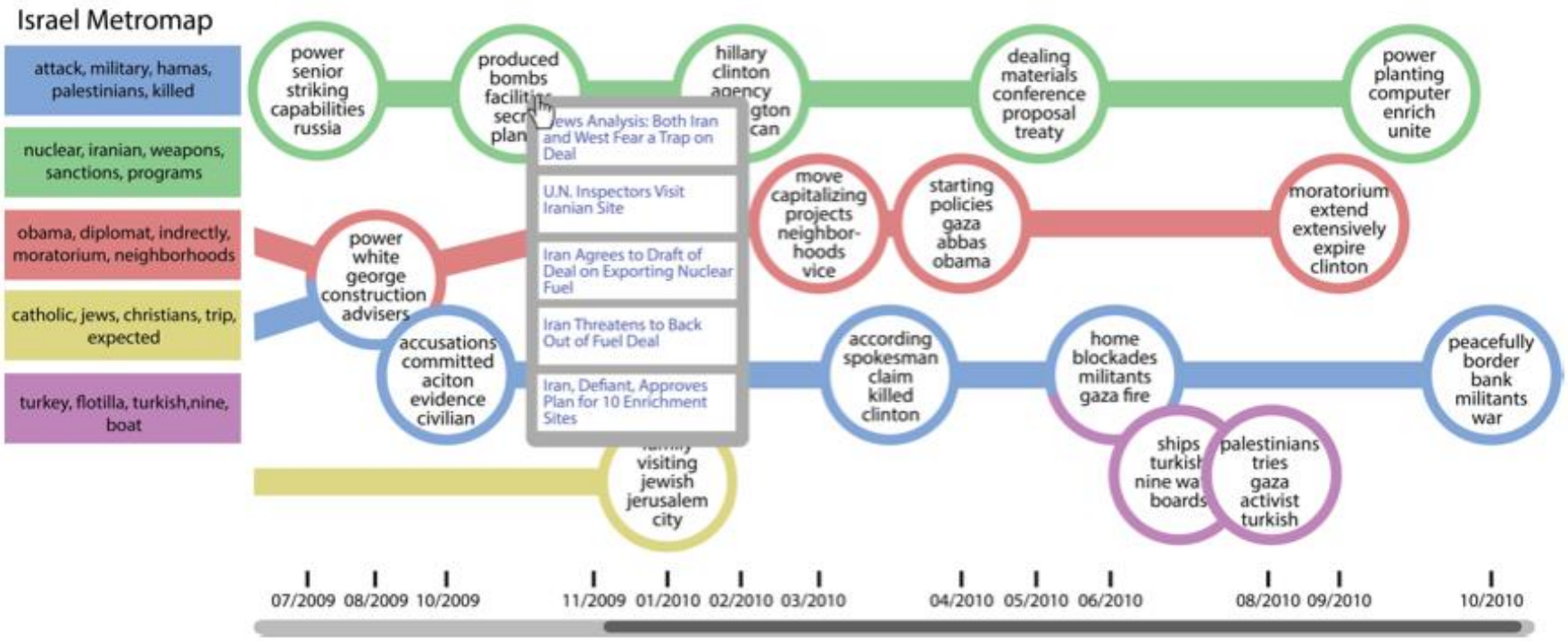
1. Визуализация текстов
2. Корпус газетных статей RuNeWC
3. Система LMMonitor
4. MongoDB для хранения корпуса
5. Оптимизация структуры коллекций MongoDB
6. Q & A

Визуализация текстов. Облака тегов

- **Облака тегов**

Визуализация текстов. Metro Maps

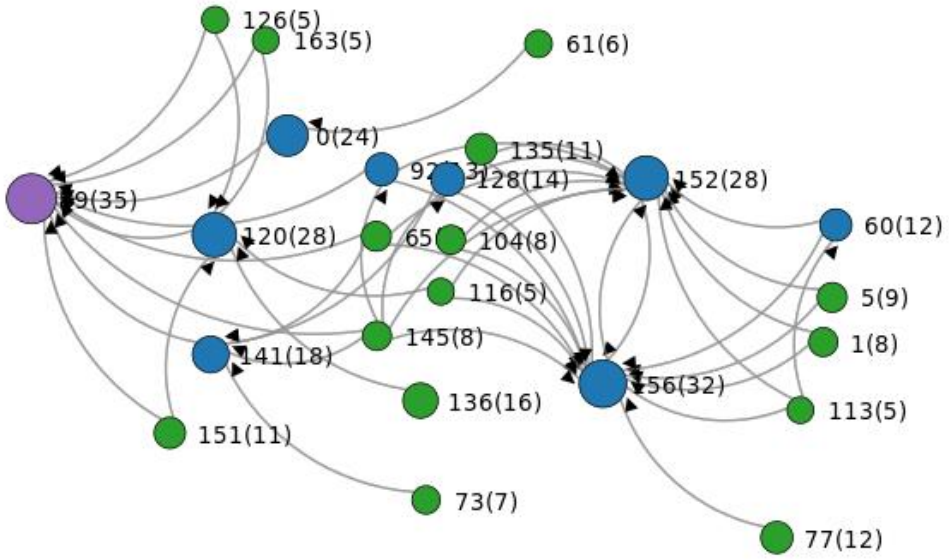
- **Metro Maps** *(Shahaf, Yang, Suen, Jacobs, Wang, Lescoves, 2013)*



- Визуализация коллекций текстов, собранных за некоторый период времени
- Каждая линия метро – скрытая тема, каждая остановка – ключевые слова

Визуализация текстов. Граф референций

- **Reference Graph** *(Mirkin, Chernyak, 2012)*



Id	Name
0	Президент Обама
1	Америка
5	Политическое положение в мире
60	Снижение зарплат
61	Заседание ЕС
65	Инфляция в России
73	Кризис
77	Международная обстановка

- Ключевое слово A отсылает к B, если Relevance(A;B) / Relevance(A) больше заданного порога

Корпус

- **Лингвистический корпус** – представительная совокупность текстов, подобранных определённым образом и снабжённых лингвистической информацией*.
- **Корпусный менеджер** – поисковая система по корпусу.

*http://opencorpora.org/doc/presentations/2011_Witology.pdf

Примеры корпусов

Название	Язык	Год	Объем	Особенности
Brown Corpus	амер.	1963	1 млн.	500 фрагментов за 1961 г.
BNC	брит.	1991	100 млн.	письмо и речь конца 20 в.
НКРЯ	рус.	2004	500 млн.	письмо, речь, мультимедиа; от сер. 18 до нач. 20 в.; параллельные подкорпусы
Открытый корпус	рус.	2009	1,6 млн.	создание сообществом; открытый код
ruTenTen11	рус.	2011	19 млрд.	тексты из Интернет; открытый

Представление корпуса

1. Набор файлов

- Разметка – в XML-подобном формате

```
<sentence id="52183">
  <source>— продолжил преподаватель.</source>
  <tokens>
    <token id="949346" text="—"><tfr t="—"><v><l id="0" t="—"><g v="PNCT"/></l></v></tfr></token>
    <token id="949347" text="продолжил"><tfr t="продолжил"><v><l id="279233" t="продолжил"><g v="VE
    "sing"/><g v="past"/><g v="indc"/></l></v></tfr></token>
    <token id="949348" text="преподаватель"><tfr t="преподаватель"><v><l id="267948" t="преподавате
    "sing"/><g v="nomn"/></l></v></tfr></token>
    <token id="949349" text="."><tfr t="."><v><l id="0" t="."><g v="PNCT"/></l></v></tfr></token>
  </tokens>
</sentence>
```

2. С использованием СУБД

- Удобно пользоваться встроенным функционалом

*http://opencorpora.org/doc/presentations/2011_Witology.pdf

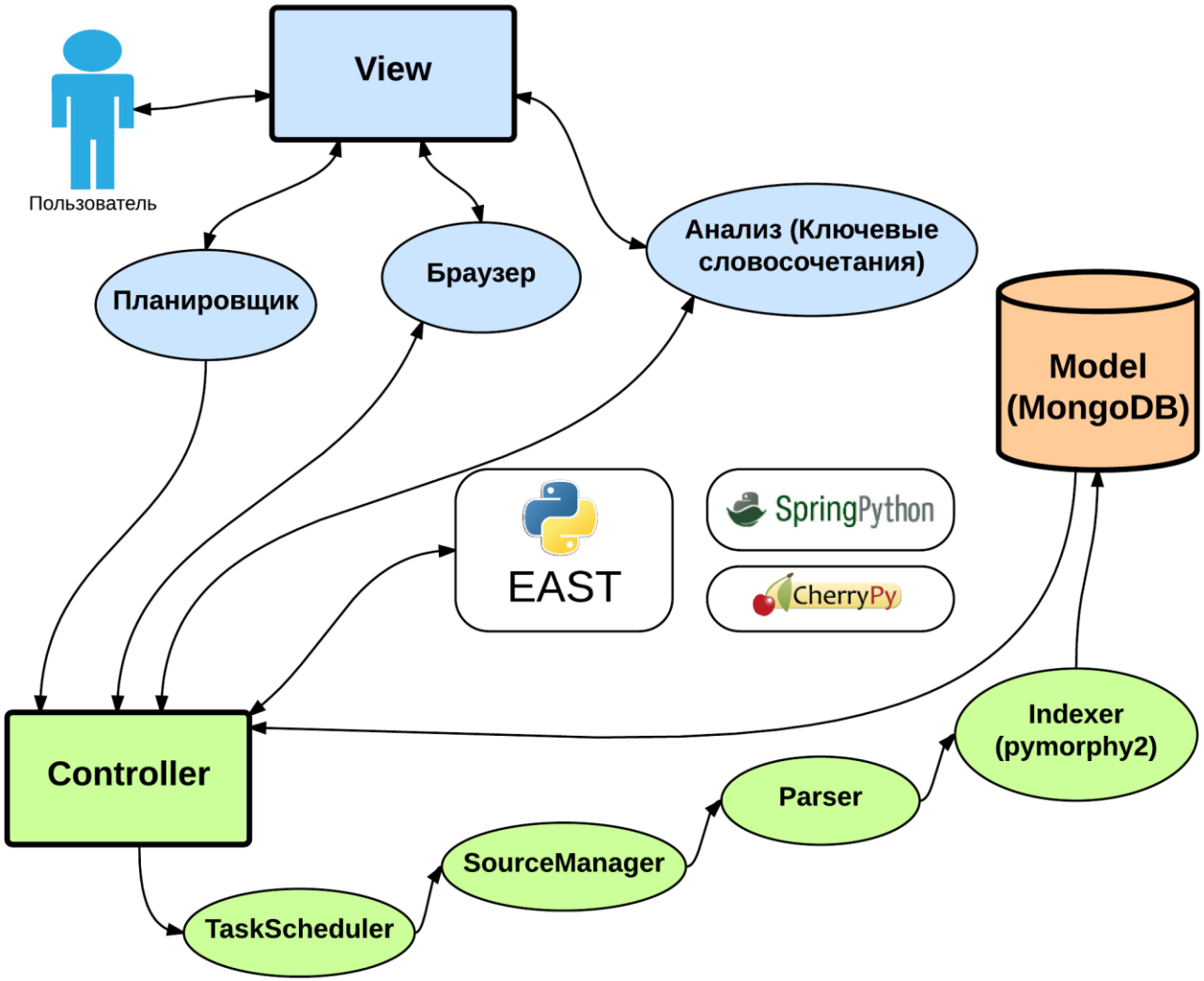
Корпус RuNeWC

- **RuNeWC – Russian Newspaper Web Corpus**
- 5 российских газет за 2014-2015 гг. по теме «Экономика»

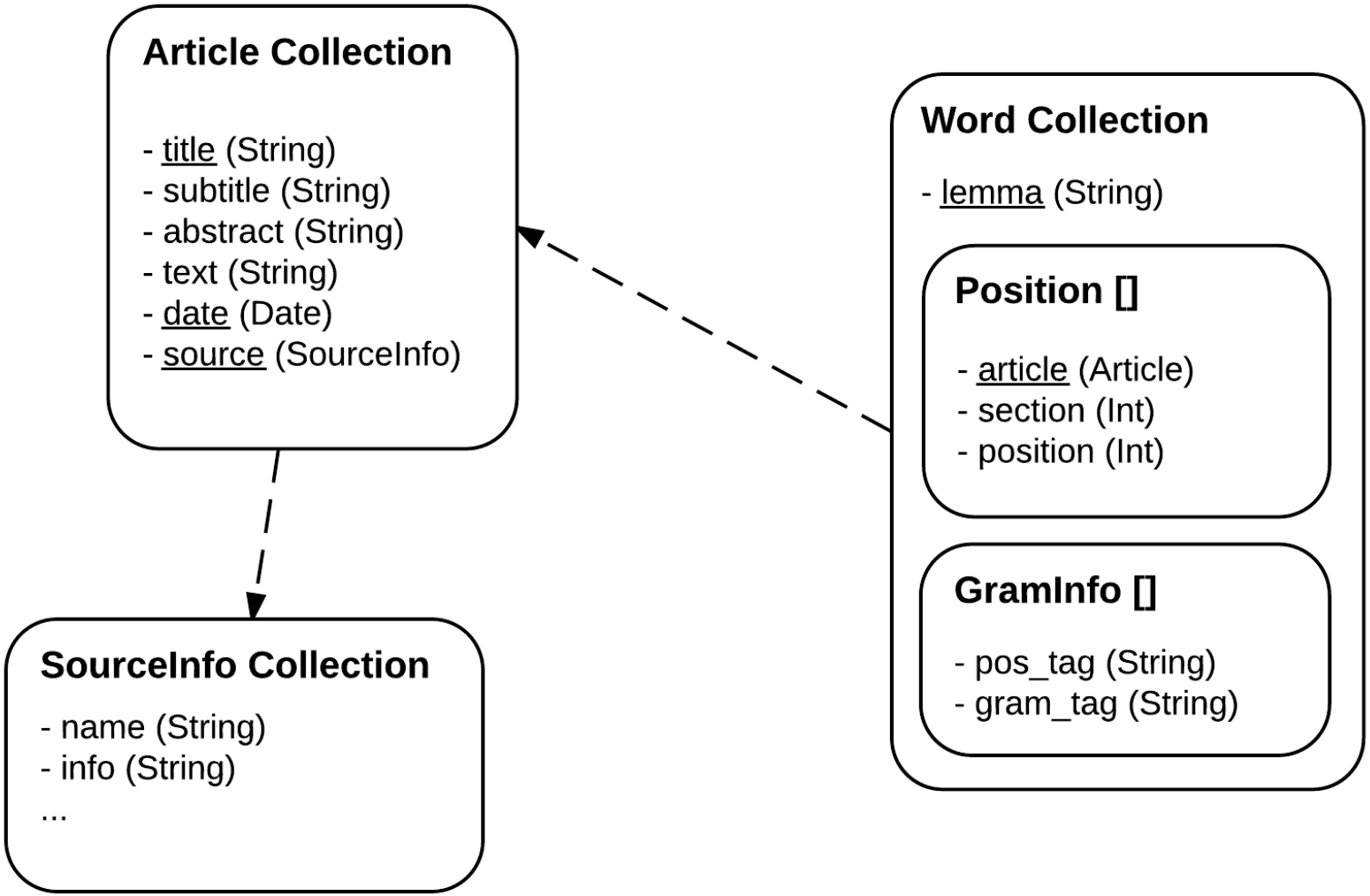
Статей	~ 4500
Словоупотреблений	~ 2,7 млн.
Различных слов	~ 65 000
Морфологическая разметка	есть
Синтаксическая разметка	нет

Хранится в MongoDB

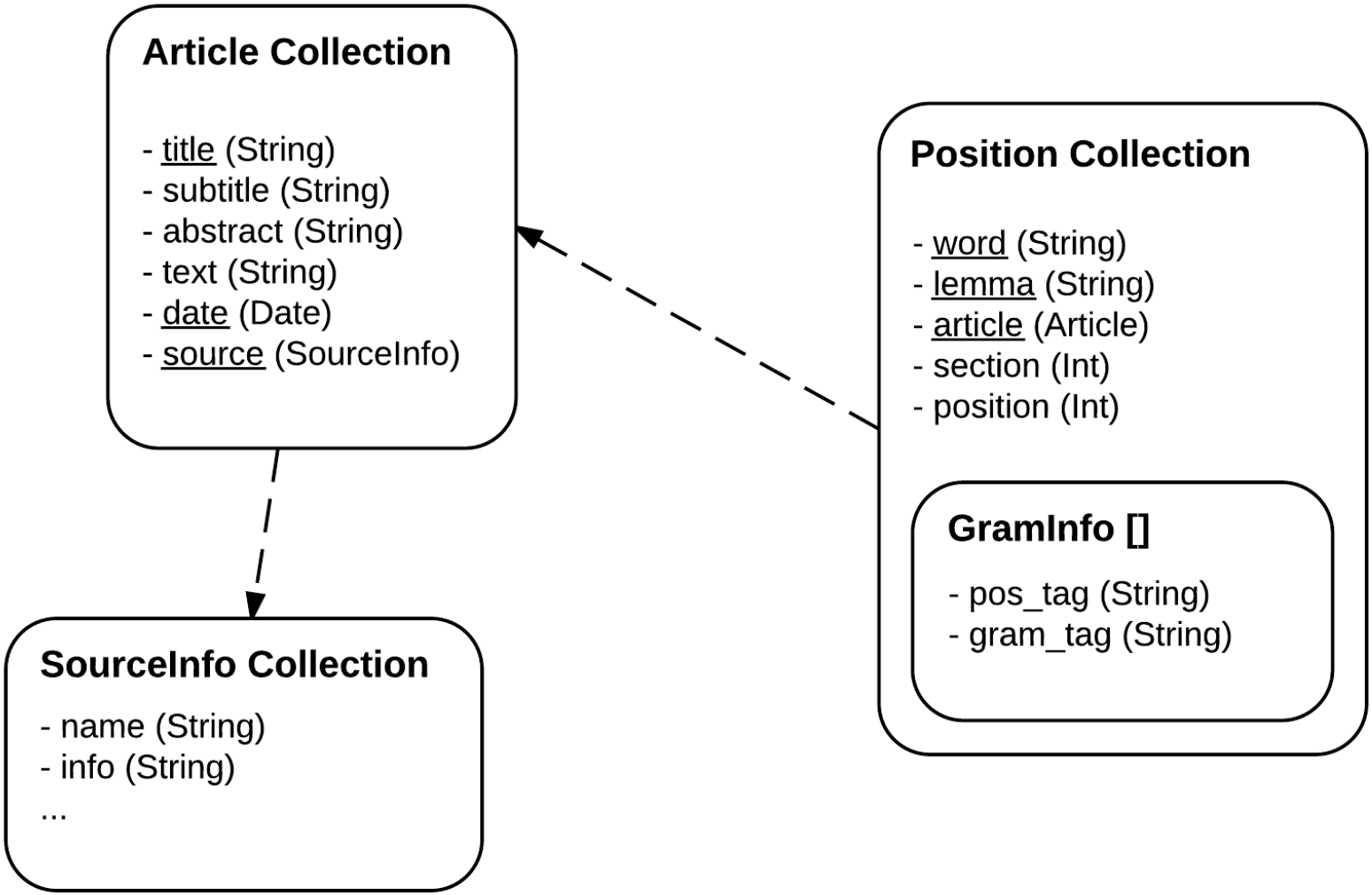
Система LM Monitor



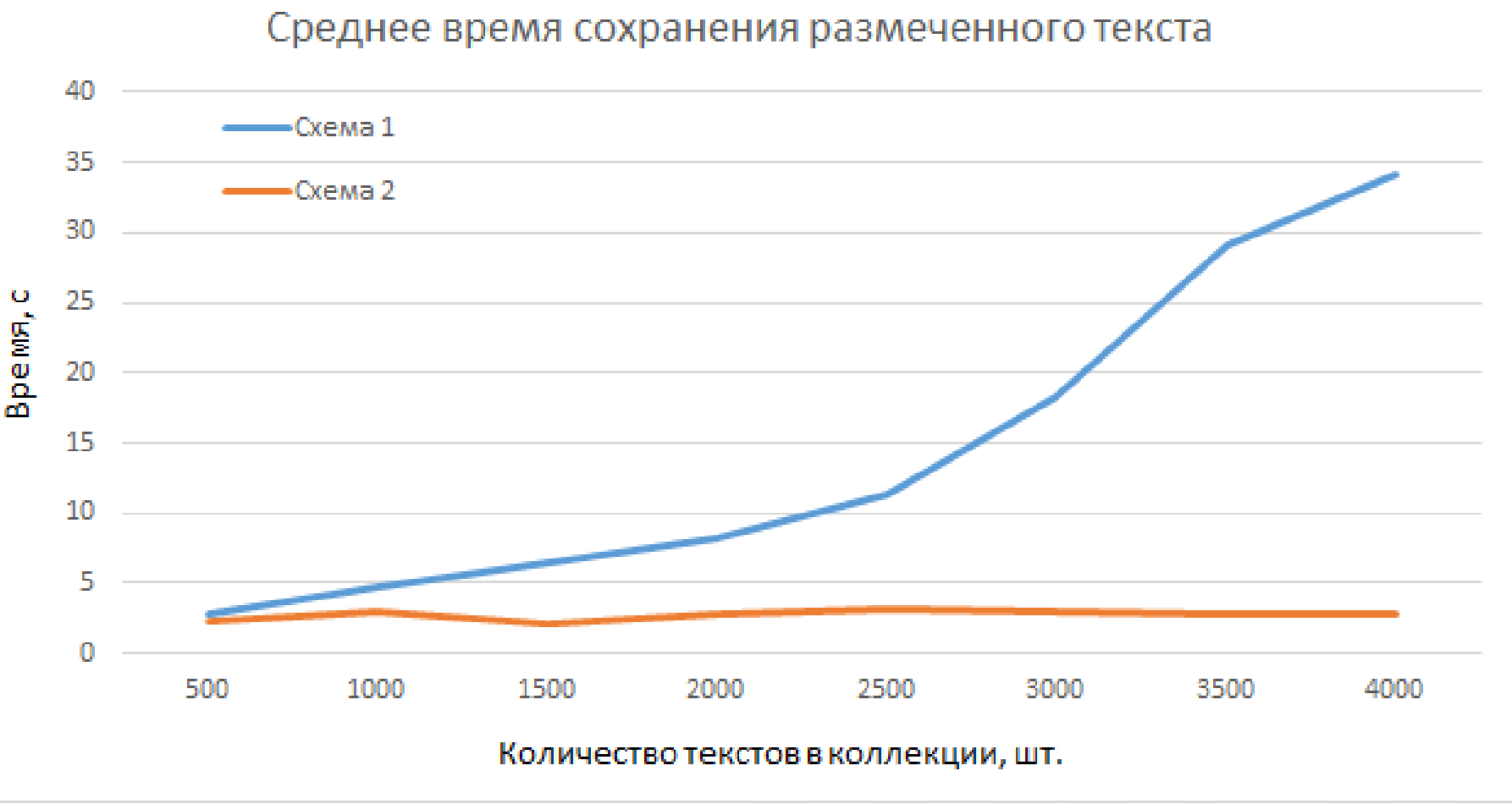
Коллекции MongoDB – Версия 1



Коллекции MongoDB – Версия 2



Сравнение производительности



Оптимизация MongoDB

- Embedded-массивы снижают производительность*:
  - Рост массива ⇒ рост содержащего документа ⇒ перемещение документа на диске
  - Индексированные поля в массиве ⇒ при работе с содержащим документом требуется больше операций

- Positions: длина пропорциональна количеству текстов
- GramInfo: длина < 5

* http://askasya.com/post/largeembeddedarrays

«… таких ориентиров три …»

{ "_id" : ObjectId("54e26479f80685a47ccc0f05"), **"lemma"** : "такой", **"word"** : ***"таких"***, **"position"** : 14, "article" : ObjectId("54e26479f80685a47ccc0eb9"), "section" : "article", **"gram_info"** : [{ **"pos_tag"** : "ADJF", **"gram_tag"** : "Apro,plur,gent" }, { **"pos_tag"** : "ADJF", **"gram_tag"** : "Apro,plur,loct" }, { **"pos_tag"** : "ADJF", **"gram_tag"** : "Apro,anim,plur,accs" }] }

{ "_id" : ObjectId("54e26479f80685a47ccc0f06"), "lemma" : "ориентир", "word" : ***"ориентиров"***, "position" : 15, "article" : ObjectId("54e26479f80685a47ccc0eb9"), "section" : "article", "gram_info" : [{ "pos_tag" : "NOUN", "gram_tag" : "inan,masc,plur,gent" }] }

{ "_id" : ObjectId("54e26479f80685a47ccc0f09"), "lemma" : "три", "word" : ***"три"***, "position" : 16, "article" : ObjectId("54e26479f80685a47ccc0eb9"), "section" : "article", "gram_info" : [{ "pos_tag" : "NUMR", "gram_tag" : "inan,accs" }, { "pos_tag" : "NUMR", "gram_tag" : "nomn" }, { "pos_tag" : "VERB", "gram_tag" : "impf,tran,sing,impr,excl" }] }

Примеры запросов

- ***Число словоупотреблений:***

```
> db.position.count()
2669362
```

- ***Из них не требуют снятия морфологической омонимии:***

```
> db.position.count({"gram_info": {$size: 1}})
1023684
```

- ***Длинные слова:***

```
> db.position.find({$where: "this.word.length > 20"}, {lemma: 1, _id: 0})
{ "lemma" : "программно-технический" }
{ "lemma" : "среднестатистический" }
{ "lemma" : "садовод-питомниковод" }
...
```

Полнотекстовый поиск

(до версии 2.6 – в экспериментальном режиме)

```
$ mongod --dbpath /var/lib/mongodb/ --setParameter textSearchEnabled=true
...
> db.article.ensureIndex({title: "text", subtitle: "text", abstract: "text", text: "text"})
> db.article.runCommand("text", {search: "китайская йена", project: {"title": 1}})


{ ... "results" : [
    {... "title" : "Русгидро и китайская корпорация China Three Gorges Corporation создадут совместное предприятие"},
    {... "title" : " Сила Сибири: Россия и Китай достигли окончательной договоренности о цене "},
    {... "title" : " Китайцы помогут Калуге восстановить аэропорт "},
    ...
},
```

Q & A