

Software Engineering Conference Russia 2018

October 12-13, Moscow



**Разработка кроссплатформенной библиотеки
морфологического анализа текстов на русском
языке для использования в промышленных
системах**

Екатерина Полицына, МАИ

Сергей Полицын, МАИ

Александр Поречный, МАИ

ЗАЧЕМ НУЖНА БИБЛИОТЕКА МОРФОЛОГИИ?

- Информационные системы используются повсеместно и оперируют большими объемами данных, в первую очередь текстовых.
- Для расширения возможностей систем документооборота, электронной коммерции и др. применяются средства автоматической обработки текстовых данных.
- Инструменты морфологического анализа являются наиболее востребованными во многих промышленных системах, работающих с большими объемами текстовых данных.



ПОЧЕМУ БИБЛИОТЕКА? КАКАЯ?

- Минимизация зависимости от разработчиков инструмента:
 - ✓ Открытая разработка
 - ✓ Не веб-API
- Доступность автономно, т.е. не в составе другой крупной системы
- Кроссплатформенность
- Поддержка русского языка

КАКИЕ ЕСТЬ БИБЛИОТЕКИ?

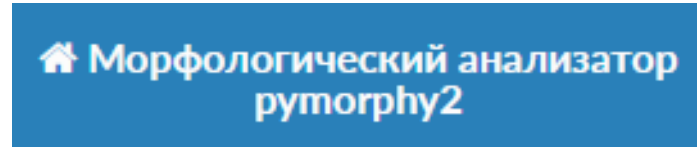


LingPipe



ABBYY

Russian Morphology for Lucene



Яндекс

MyStem

CrossMorphy

КАКАЯ НУЖНА БИБЛИОТЕКА?

- Возможность получения морфологических характеристик слова.
- Возможность получения начальной формы слова.
- Возможность получения словоформы по строковому представлению начальной формы и заданным морфологическим характеристикам.
- Кроссплатформенность.
- Высокая производительность.
- Наличие классов для работы с предложениями и словами.

АРХИТЕКТУРА

Файл с
морфологическими
характеристиками

- Связь словоформ с начальной формой
- Морфологические характеристики словоформ

JMorfSdk

- Получение морфологических характеристик слова
- Получение начальной формы слова
- Получение слова по его начальной форме и морфологическим характеристикам

ОПТИМИЗАЦИЯ БИБЛИОТЕКИ

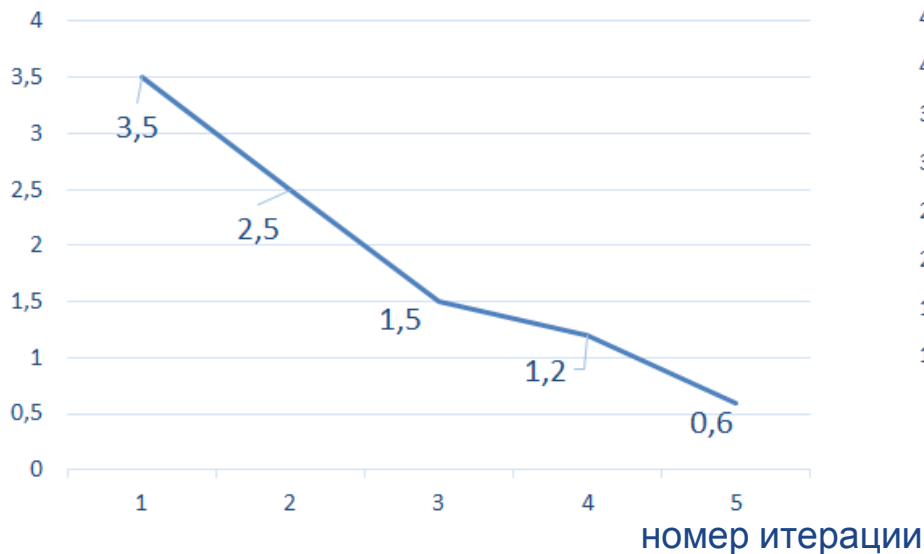
Было:

- Словарь в ПЗУ – 500МБ
- Библиотека в ОЗУ 3.5ГБ

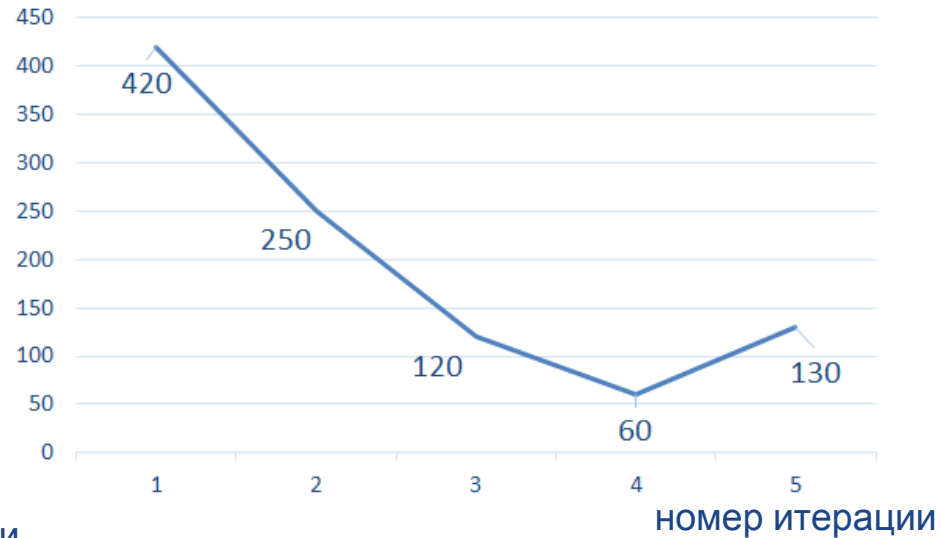
Стало:

- Словарь в ПЗУ – 35+95МБ
- Библиотека в ОЗУ 550МБ

Динамика изменения в ОЗУ, ГБ



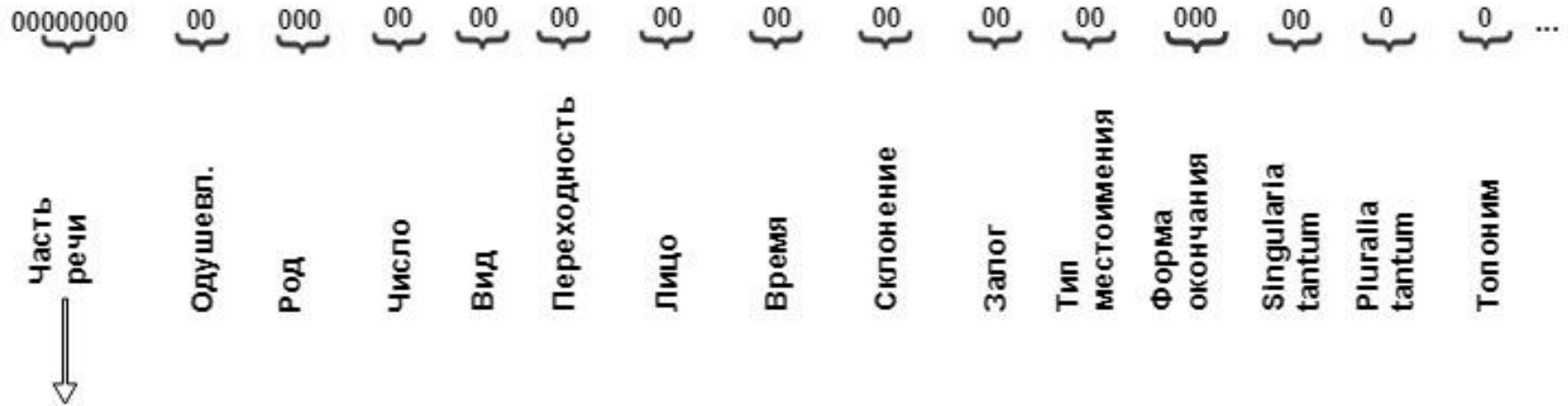
Динамика изменения в ПЗУ, МБ



ИСХОДНАЯ СТРУКТУРА СЛОВАРЯ OpenCorpora

```
rev="1"><l t="ёж"><g v="NOUN"/><g v="anim"/><g v="masc"/><
rev="2"><l t="ёж"><g v="NOUN"/><g v="inan"/><g v="masc"/><
rev="3"><l t="ёжик"><g v="NOUN"/><g v="anim"/><g v="masc"/><
rev="4"><l t="ёжиком"><g v="ADVB"/></l><f t="ёжиком"></f><
rev="5"><l t="ёжистый"><g v="ADJF"/><g v="Qual"/></l><f t="
rev="6"><l t="ёжист"><g v="ADJS"/><g v="Qual"/></l><f t="ё
rev="7"><l t="ёжистее"><g v="COMP"/><g v="Qual"/></l><f t="
rev="8"><l t="ёжу"><g v="VERB"/><g v="impf"/><g v="tran"/>
rev="9"><l t="ёжить"><g v="INFN"/><g v="impf"/><g v="tran"
rev="10"><l t="ёжимый"><g v="PRTF"/><g v="impf"/><g v="tr
rev="11"><l t="ёжим"><g v="PRTS"/><g v="impf"/><g v="pres
rev="12"><l t="ёжа"><g v="GRND"/><g v="impf"/><g v="tran"
rev="13"><l t="ёжусь"><g v="VERB"/><g v="impf"/><g v="int
rev="14"><l t="ёжиться"><g v="INFN"/><g v="impf"/><g v="i
rev="15"><l t="ёжащийся"><g v="PRTF"/><g v="impf"/><g v="
rev="16"><l t="ёжившийся"><g v="PRTF"/><g v="impf"/><g v="
rev="17"><l t="ёжася"><g v="GRND"/><g v="impf"/><g v="int
```


МОРФОЛОГИЧЕСКАЯ ШКАЛА



0001 0001 **Существительное**
 0001 0010 **Полное прилагательное**
 0001 0011 **Краткое прилагательное**
 0001 0100 **Глагол**
 0001 0110 **Полное причастие**
 0001 1011 **Краткое причастие**
 0001 1100 **Деепричастие**
 ...

ПРОМЕЖУТОЧНАЯ СТРУКТУРА СЛОВАРЯ

биогидроакустика 11 100000db"биогидроакустику 10000045
перештемпелёвывающий 16 10258d4"перештемпелёвывающим 10
банк 11 d7"банка 157"банком 2d7"банку 1d7"банке 3d7"бан
обольстить 17 5000
неирландский 12 d4"неирландскому 1d4"неирландском 3dc"н
выщелкал 14 145054"выщелкайте 785060"выщелкало 14505c"в
изыскавшийся 16 10470d4"изыскавшееся 104745c"изыскавшего
туаз 11 d7"туаз 457"туазе 3d7"туазами 2f7"туазы f7"туаз
почтенный 12 2000000000d4"почтенном 2000000003dc"почтен
аяна 11 80000da"аянами 80002fa"аян 800047a"аяны 800015a
банг 11 d7"бангов 177"бангами 2f7"бангах 3f7"банга 157"
экспорт 11 d7"экспорт 457"экспортом 2d7"экспорта 157"эк
выщелкав 19 45000"выщелкавши 4501b

СТРУКТУРЫ ДАННЫХ В СЛОВАРЕ

ID	
№ в словаре	Java hash словоформы (старший байт)
3 байта	1 байта

Часть речи (TypeOfSpeech)	
Существительное	0001_0001
Полное прилагательное	0001_0011
Краткое прилагательное	0001_0101
Глагол	0001_0100
Именное местоимение	0001_1101
Наречие	0000_1001
Союз	0000_1000
...	...

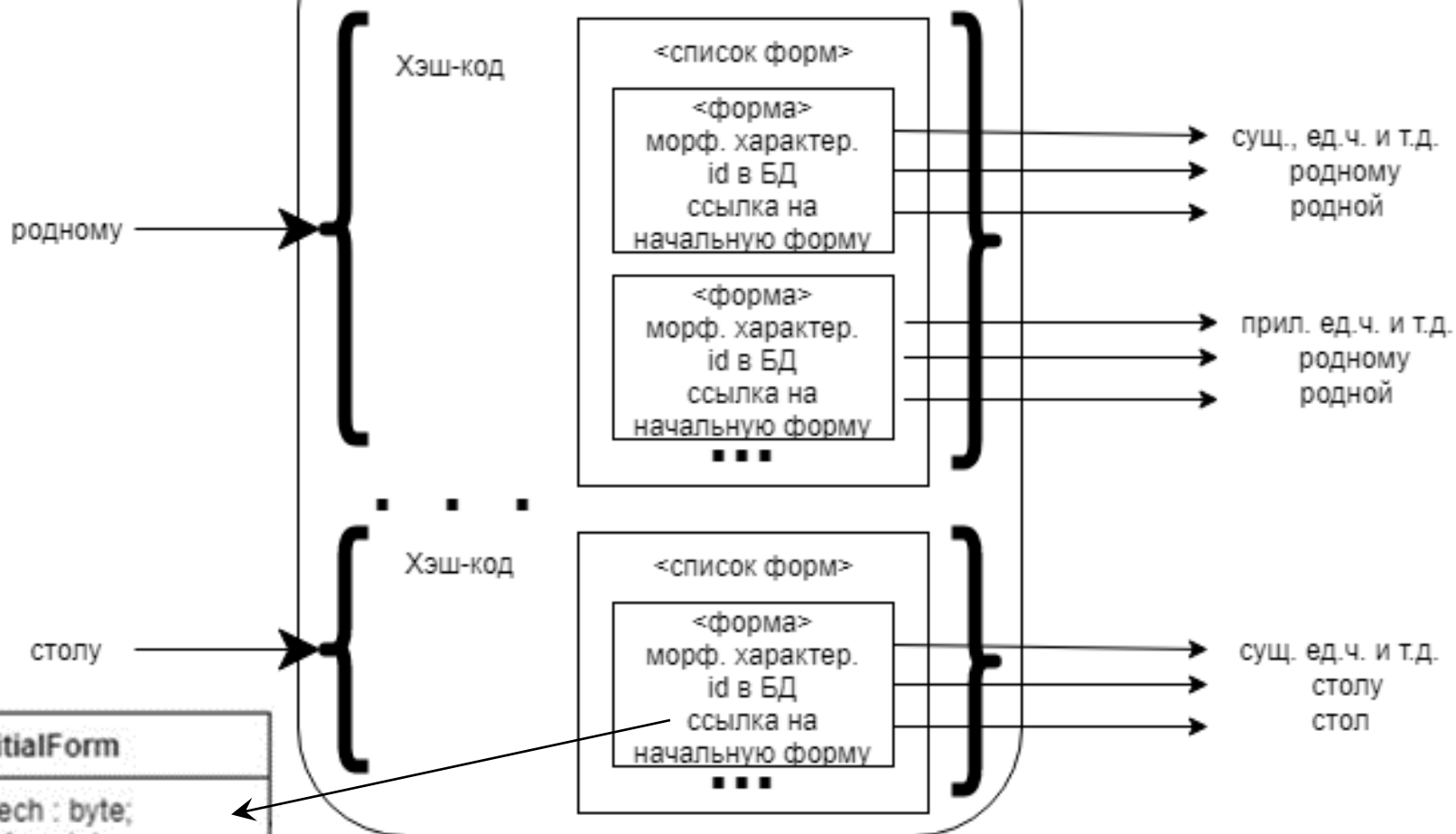
Начальная форма (InitialForm)			
CityHash	ID	Часть речи	Морфологические характеристики
4 байта	4 байта	1 байт	8 байт

Морфологические характеристики (MorphologicalCharacteristics)	
Одушевленное	0000_..._0000_0000_0010
Единственное число	0000_..._0000_0000_0010
Множественное число	0000_..._0000_0000_0011
...	...
Именительный падеж	0000_..._0000_1000_0000
Родительный падеж	0000_..._0010_1000_0000
...	...

ОПТИМИЗАЦИЯ СТРУКТУР ДАННЫХ

Внешнее представление

Внутреннее представление

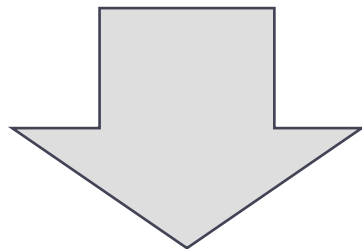


InitialForm

```
- typeOfSpeech : byte;  
- initialFormKey : int;  
- wordFormList List<WordForm>  
- morfCharacteristics : long;  
- formKeyInBD : int;
```

ПОИСК СЛОВОФОРМЫ

1. Получение хэш-кода по алгоритму Google CityHash64 (ключ в HashMap).
2. Получение хэш-кода строки по встроенному алгоритму Java (*.hashCode()*).
3. Пересечение результатов поиска по этим двум хэш-кодам.



4. **Разрешение коллизий.**

 [google / cityhash](#)



СТРУКТУРА ЗАПИСИ СЛОВАРЯ

Лемма 1

Лемма 2

17 байт	16 байт	16 байт	...	16 байт	4 байта	17 байт	16 байт	...
InitialForm	WordForm	WordForm	...	WordForm	ControlValue	InitialForm	WordForm	...

InitialForm			
CityHash	ID	TypeOfSpeech	MorphologicalCharacteristics
4 байта	4 байта	1 байт	8 байт

WordForm		
CityHash	ID	MorphologicalCharacteristics
4 байта	4 байта	8 байт

ControlValue
1111_1111 x 4 байта

ИТОГОВАЯ СТРУКТУРА СЛОВАРЯ

```

?♀♂ @z◀ Q?R?•?? f????± ???? ????± ?&E?
> ????> &Zm/△ δ5 @f ??? ?> @??>4 ????±
> u????> J5 ???g#> *5 ???2?> J5 @6??>Q ?> @u$?>7? <5
?>p??> "? @> Z>fu> #? ? ? ??iδ> $? @> ????>?? ? ?
,$o?>+> +>
<,$??> >
?5??> ->
0Rs??> < ,;:A'> /> <,@????????> >δ > ????H?9b
5- ?>,d3bΔ?> 0? ?>.,??_??> 1? ?>.,?3bΔ?> 0? ?>.&H?9b?> 25 ?>.'#?!!> 35
>? i o?↓> B5 i0_>?I> C? ?> ?NT ?> D? ?>0J6t> E5 ?
<$Z??> H?
<<\t??> I?
<.>?t?>v J?
<0?>4?Δ K?
?><< Δn>?> L? <<0????2<??> J?δ < ????>
*?> ?><dΔ??> M? ?><??c↓> N? ?><?Δ??> M? ?>&??>
?>?'>d?> P? ?>=dδ?2> Q? ?>=?~1??> R? ?><h?<?> S? ?><??<?>
8$3?>R h6
8<??X+> i6
8.[W??> j6
80 ?>w< k> δx >?d??> l> δx0!>?>? m6 δ? ?A?J? n6 δ?0????> o>
?> tΔ ?>8??Wδ?> uΔ ?>8?I?
?> tΔ ?>:&??\>? u? ?>:'?u?>? w? ?>9d?z?b> x? ?>?/?>?*> y? ?>8hX_??>
?> tΔ ?>8??Wδ?> uΔ ?>8????< >? ?>: ,?u?>? w? ?>91?z?b> x? ?>??????>
* ?>P ?>?@J?> ?>P ?>@Q??8?> ?? ?>'G1?> ?? ?>@ly>n> ?? ?>@????x> ?? ?>@????+?>
* ?>P ?>?WΔ?> ?? ?>@,G1?> ?? ?>@ly>n> ?? ?>@????x> ?? ?>@????+?>
?>?>C> ?? ?> ?>????>+?> ↓ L?>↓ ?> d?> l?> ?>r ?> ?>?m?> ?>r ?> ?> l?>
?>?>v>k> ?? ?> ?>I^??> ?? ?> ?> >/?W?> ?? ?> ?>????~r??> ?>?>↓ k*??>h>
i_?> ?>6 i0, l?>?> ?> ?>XI?<?> ?> ?>0s*:>?> ?>6 ?> ?>M?> ?>6 ?>
<$c/5i> ?>6
<<B*Q> ?>6
<.>z?<n> ?>6
<0????>+ ?> ?><< ?>?b1?> ?> <<0????SZ?> "6δ < ?>??U+?> #G_ ?><d??>k1?> ??
?>?>?>?>?>
    
```

ID	Строковое представление слова
231512	биогидроакустика



ИТОГОВАЯ АРХИТЕКТУРА

MorphologicalStructures

- Конвертация морфологического словаря в структуру, применяемую в JMorfSdk
- Работа с БД, в которой хранятся текстовые формы словоформ

JMorfSdk

- Получение морфологических характеристик слова
- Получение начальной формы слова
- Получение слова по его начальной форме и морфологическим характеристикам

Бинарный файл с морфологическими характеристиками

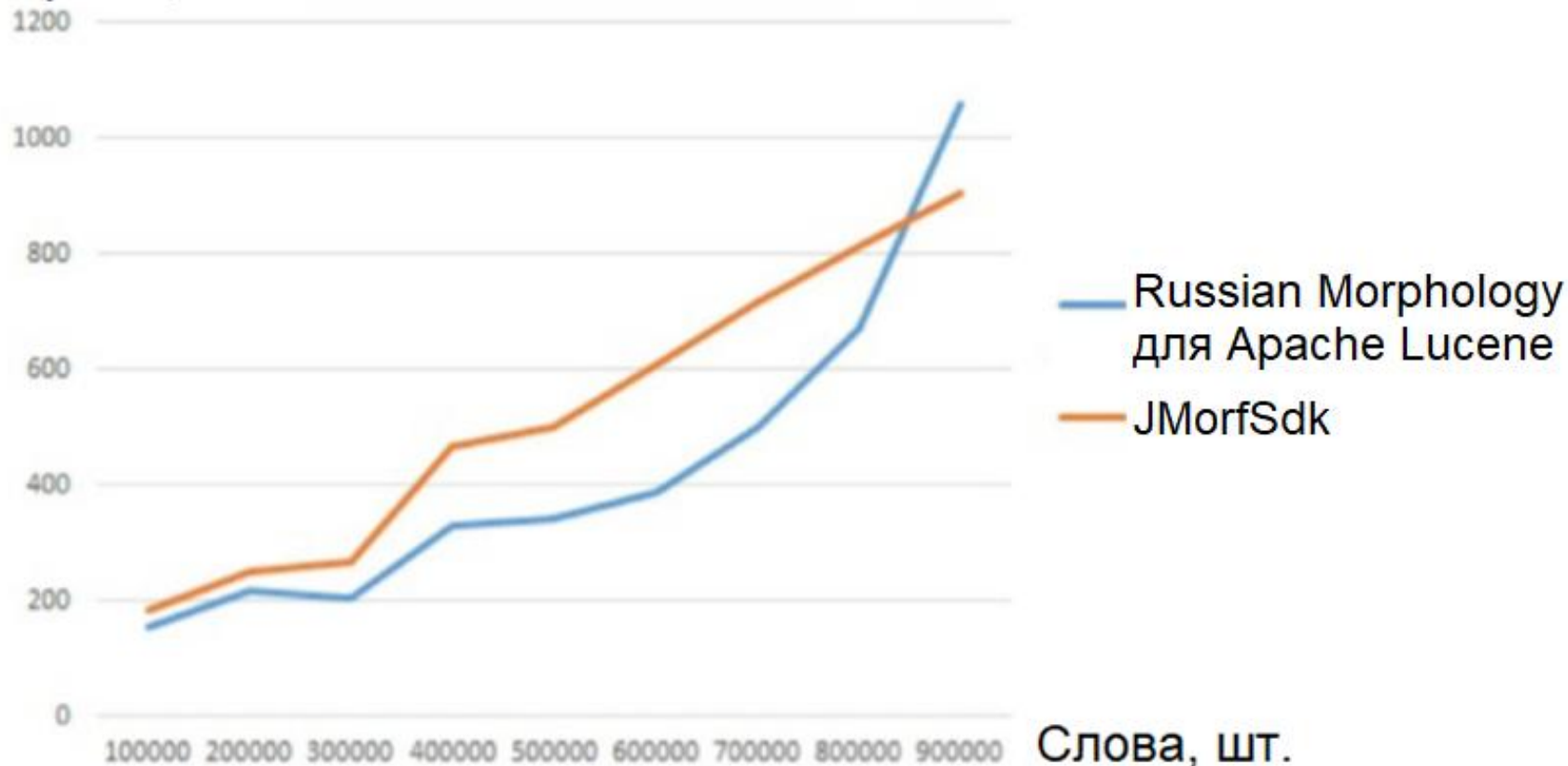
- Связь словоформ с начальной формой
- Морфологические характеристики словоформ в бинарном виде
- Содержит идентификаторы словоформ для БД

БД (SQLite)

- Две БД по содержанию: только начальные формы и производные от них словоформы
- Текстовое представление словоформ, доступное по идентификатору

СРАВНЕНИЕ С JAVA RUSSIAN MORPHOLOGY ДЛЯ APACHE LUCENE

Время, мс



Слова, шт.

ФИЛЬТРАЦИЯ СЛОВ ПО МОРФОЛОГИЧЕСКИМ ХАРАКТЕРИСТИКАМ

```
List<String> words = Arrays.asList("красного", "красный", "морозное", "солнечная",  
    "бежать", "доска", "топор", "шла");  
  
for (String word : words) {  
    jMorfSdk.getAllCharacteristicsOfForm(word).forEach(form -> {  
        if (form.getTheMorfCharacteristics(MorfologyParameters.Gender.class)  
            == MorfologyParameters.Gender.FEMININ) {  
            System.out.println(form + " - " + word);  
        }  
    });  
}
```

```
initialFormString = 'солнечный', typeOfSpeech = 18, morfCharacteristics = 4200 - солнечная  
initialFormString = доска, typeOfSpeech = 17, morfCharacteristics = 107 - доска  
initialFormString = иду, typeOfSpeech = 20, morfCharacteristics = 670760 - шла
```

ПОЛУЧЕНИЕ ЗАДАННОЙ ФОРМЫ СЛОВА

```
sdk.getAllCharacteristicsOfForm("дорогой").forEach((form) -> {  
    if (form.getTheMorfCharacteristics(MorfologyParameters.Case.IDENTIFIER) ==  
        MorfologyParameters.Case.GENITIVE) {  
        System.out.println(form);  
    }  
});
```

```
initialFormString = дорогой, typeOfSpeech = 18, morfCharacteristics = 4264
```

```
sdk.getAllCharacteristicsOfForm("дорогой").forEach((form) -> {  
    if (form.getTypeOfSpeech() == MorfologyParameters.TypeOfSpeech.NOUN) {  
        System.out.println(form);  
    }  
});
```

```
initialFormString = дорога, typeOfSpeech = 17, morfCharacteristics = 363
```



ПОЛУЧЕНИЕ СЛОВА С ЗАДАНЫМИ ХАРАКТЕРИСТИКАМИ

```
jMorfSdk . getDerivativeForm ( " МЫЛО " ,  
    TypeOfSpeech . NOUN ,  
    Numbers . SINGULAR )  
    . forEach ( ( wordString ) -> {  
        System . out . println ( wordString ) ;  
    } ) ;
```

МЫЛОМ

МЫЛО

МЫЛЕ

МЫЛА

МЫЛУ



ГДЕ ИСПОЛЬЗОВАЛИ?

- В системе классификации текстов на основе ключевых слов и словосочетаний.
- При решении задачи выделения именованных сущностей.
- Для восстановления полных форм слов при работе с сокращениями.
- В исследовательских задачах компьютерной лингвистики: морфологический этап анализа - один из основных, необходим для большинства задач.



СОЗДАННАЯ БИБЛИОТЕКА

Реализованная кроссплатформенная библиотека JMorfSdk имеет:

- 1) Режим анализа: определение морфологических характеристик слова и получение начальной формы со сложностью $O(1)$.
- 2) Режим генерации: получение слова по строковому представлению слова и набору морфологических характеристик.



<https://github.com/jalexpr/JMorfSdk>

```
<dependency>  
  <groupId>ru.textanalysis.tfwat</groupId>  
  <artifactId>jmorfsdk</artifactId>  
  <version>2.10.5</version>  
</dependency>
```





СПАСИБО ЗА ВНИМАНИЕ!

Ekaterina Politsyna
Sergey Politsyn
Alexander Porechny

<http://textanalysis.ru/>
tasystem@yandex.ru

