Лингвистический компонент Лемматайзер для русского языка

Техническое описание

SemanticAnalyzer Group, 2013-08-18
www.semanticanalyzer.info

Данный документ описывает технические характеристики лемматайзера для русского языка. Предполагается, что перед применением данного компонента входной текст был предобработан модулем Токенайзера (см. соответствующее техническое описание).

Демо-пакет, предоставляемый по запросу, содержит следующие компоненты:

- Java библиотека лемматайзера в виде бинарного файла
- run_lemmatizer.sh для быстрой проверки работы модуля
- файл messages_to_lemmatize.txt, содержащий примеры общего текста и текстов Твиттера для лемматизации скриптом run_lemmatizer.sh

Алгоритм основан на комбинации следующих компонент:

- словарный поиск
- алгоритм вычисляющий морф. характеристики неизвестных слов
- анализатор сложных слов
- анализатор чисел
- анализатор на правилах

Скорость работы модуля

Сервер: Intel(R) Xeon(R) CPU X3363 @ 2.83GHz
Операционная система: ubuntu 10.04, Java 1.7.0_21 64 bit server
5037 символов/миллисекунду
880 слов/миллисекунду

Тесты проводились в один поток.

Формат messages_to_lemmatize.txt

Данный файл описывает входные данные для демонстрации работы модуля лемматайзера.
Формат:

Текст\tТип текста

Текст – содержимое текста на русском языке для токенизации
\t – символ табуляции
Тип текста: поддерживаются два значения: GENERAL_TEXT и TWITTER.

Примеры токенизации

Скрипт run_lemmatizer.sh генерирует файл: messages_to_lemmatize.out.
Для входного текста в файле messages_to_tokenize.txt:

Прекрасный вечер))) прогулка по Набережной - самое то;) только маккафе подпортило настроение(
	TWITTER

Порождается следующий вывод:

Прекрасный, type: ALPHANUM
MorphDesc[removeNum=0,lemmaEnding=ый,endings=[а, ая, ее, ей, о, ого, ое, ой, ом, ому, ою, ую, ы, ые, ым, ыми, ых],lemma=прекрасный,pos=ADJECTIVE,weight=14317,stem=прекрасн]

вечер, type: ALPHANUM
MorphDesc[removeNum=0,lemmaEnding=,endings=[а, ам, ами, ах, е, ов, ом, у],lemma=вечер,pos=NOUN,weight=39101,stem=вечер]

emopostkn, type: ALPHANUM

emopostkn, type: ALPHANUM

emopostkn, type: ALPHANUM

прогулка, type: ALPHANUM
MorphDesc[removeNum=0,lemmaEnding=а,endings=[ам, ами, ах, е, и, ой, ою, у],lemma=прогулка,pos=NOUN,weight=3054,stem=прогулк]

по, type: ALPHANUM

MorphDesc[removeNum=0,lemmaEnding=,endings=[],lemma=по,pos=PREPOSITION,weight=573564,stem=по]

Набережной, type: ALPHANUM
MorphDesc[removeNum=0,lemmaEnding=ая,endings=[ой, ою, ую, ые, ым, ыми, ых],lemma=набережная,pos=NOUN,weight=2908,stem=набережн]

-, type: PUNCT
MorphDesc[removeNum=0,lemmaEnding=,endings=[],lemma=-,pos=NUMERAL,weight=0,stem=-]

самое, type: ALPHANUM
MorphDesc[removeNum=0,lemmaEnding=ый,endings=[ая, ого, ое, ой, ом, ому, ою, ую, ые, ым, ыми, ых],lemma=самый,pos=ADJECTIVE,weight=0,stem=сам]

то, type: ALPHANUM

MorphDesc[removeNum=0,lemmaEnding=,endings=[],lemma=то,pos=CONJUNCTION,weight=0,stem=то]

MorphDesc[removeNum=0,lemmaEnding=,endings=[],lemma=то,pos=ADVERB,weight=0,stem=то]
MorphDesc[removeNum=0,lemmaEnding=от,endings=[а, е, ем, еми, ех, о, ого, ой, ом, ому, ою, у],lemma=тот,pos=PRONOUN_ADJECTIVE,weight=1139844,stem=т]
MorphDesc[removeNum=0,lemmaEnding=о,endings=[е, ем, еми, ех, ого, ом, ому],lemma=то,pos=NOUN,weight=0,stem=т]

emopostkn, type: ALPHANUM

только, type: ALPHANUM

MorphDesc[removeNum=0,lemmaEnding=,endings=[],lemma=только,pos=PARTICLE,weight=0,stem=только]

MorphDesc[removeNum=0,lemmaEnding=,endings=[],lemma=только,pos=ADVERB,weight=0,stem=только]

маккафе, type: ALPHANUM

подпортило, type: ALPHANUM
MorphDesc[removeNum=0,lemmaEnding=ть,endings=[в, вший, вшего, вшему, вшим, вшем, вшая, вшей, вшую, вшею, вшее, вшие, вших, вшими, вши, л, ла, ли, ло],lemma=подпортить,pos=VERB,weight=190,stem=подпорти]

настроение, type: ALPHANUM
MorphDesc[removeNum=0,lemmaEnding=е,endings=[ем, и, ю, я],lemma=настроение,pos=NOUN,weight=8416,stem=настроени]
MorphDesc[removeNum=0,lemmaEnding=е,endings=[ем, и, й, ю, я, ям, ями, ях],lemma=настроение,pos=NOUN,weight=8416,stem=настроени]

emonegtkn, type: ALPHANUM

Примеры использования из Java

```
MorphAnalyzer morphAnalyzer = MorphAnalyzerLoader.loadDefault();
System.out.println(morphAnalyzer.analyzeBest("русского"));
```

вывод:

MorphDesc[removeNum=0,lemmaEnding=ий,endings=[ая, ие, им, ими, их, ого, ое, ой, ом, ому, ою, ую],lemma=русский,pos=ADJECTIVE,weight=36739,stem=русск]