Linguistic component: Lemmatizer for the Russian language

Technical description

SemanticAnalyzer Group, 2013-08-29
www.semanticanalyzer.info

This document describes technical details of lemmatizer for the Russian language.

It is assumed, that prior to using this component an input text has been preprocessed with Tokenizer component (see the corresponding Technical Description).

Demo package sent upon request contains the following:

- Java library of tokenizer in a form of a binary
- run_lemmatizer.sh script for swift checking the functionality of the module
- messages_to_lemmatize.txt file containing examples of generic text and tweets for tokenization using the run_lemmatizer.sh script

Algorithm is based on combination of the following:

- dictionary search
- algorithm calculating morphological properties of unknown words
- compound word analyzer
- analyzer of numbers
- rule-based analyzer

Speed of processing

Server: Intel(R) Xeon(R) CPU X3363 @ 2.83GHz
Operating system: ubuntu 10.04, Java 1.7.0_21 64 bit server
5037 characters/ms
880 tokens/ms

Tests were conducted in a single thread.

Format of the messages_to_lemmatize.txt file

This file describes input data for the tokenizer module for demo purposes.Формат:

Format:
Text\tText type

Text contains textual data in Russian for lemmatization
\t – tab symbol
Text type: supported values are GENERAL_TEXT and TWITTER.

Examples of lemmatization

The run_lemmatizer.sh script will generate the following file: messages_to_lemmatize.out.
For the following input file messages_to_tokenize.txt:

Прекрасный вечер))) прогулка по Набережной - самое то;) только маккафе подпортило настроение(
TWITTER
This output gets generated:

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

Examples of using the library from the Java code

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

output:

```
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]
```