

Отчет: морфологический анализатор

Ход выполнения: морфологический анализатор был выполнен на языке программирования Java, с использованием открытого морфологического словаря OpenCorpora.

Результат: получившийся анализатор способен принимать на вход текст довольно большого объема, разбивать его на токены и для каждого токена при помощи словаря находить все возможные леммы, от которых этот токен может формироваться. Кроме того анализатор способен давать информацию о граммемах произвольного слова.

Как результат работы программы, мы имеем отсортированный список записей, каждая запись содержит слово из входного текста, лемму, к которой это слово относится и его часть речи.

На этапе первой версии морфологический анализатор не способен разрешать омонимию, поэтому токены, имеющие несколько возможных лемм, в текущей версии анализатора не рассматриваются. Из-за этого присутствует большое количество слов, которые не попадают в финальный список (таких слов сейчас около 40%).

```
рука 1671 NOUN
человек 1733 NOUN
говорию 1764 VERB
мы 1898 NPRO
за 1934 PREP
мой 2089 ADJF
ибо 2391 CONJ
вы 2406 NPRO
земля 2606 NOUN
народ 2686 NOUN
бог 3264 NOUN
я 3444 NPRO
свой 3984 ADJF
от 4073 PREP
сын 4106 NOUN
сказал 4202 VERB
который 4203 ADJF
твой 4336 ADJF
они 4430 NPRO
господь 4706 NOUN
он 4739 NPRO
весь 4923 ADJF
есть 5744 VERB
ты 6333 NPRO
не 8600 PRCL
```

Особенности лексики:

В качестве корпуса был взят Ветхий завет. Как можно заметить, среди самых встречаемых слов в тексте есть частицы, местоимения, предлоги (что в целом естественно для любого текста), а также слова, присущие конкретно этому тексту: есть (как глагол), господь, бог, сын, сказал, народ, земля, ибо и т.д.