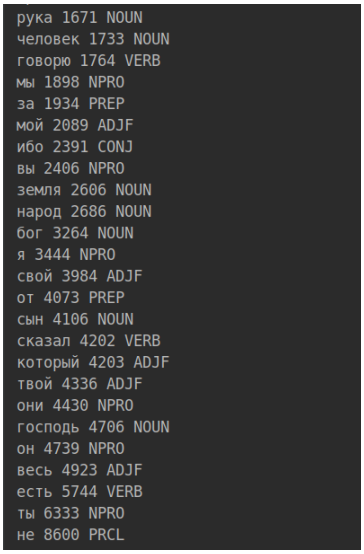
Отчет: морфологический анализатор

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%).



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.