# Отчет: извлечение  n-грамм

Ход выполнения: Метод извлечения  n-грамм был реализован на языке программирования Java, с использованием открытого морфологического словаря OpenCorpora, который используется для лемматизации токенов.

Описание программы: реализованный алгоритм извлечения n-грамм принимает на вход текст большого размера, разбивает его на токены, каждый токен лемматизируется. Кроме того программа принимает на вход число N - размер n-грамм

Второй этап работы программы: создаем хэш-таблицу для того, чтобы быстро сравнивать строки. Далее проходимся по всему тексту, и все подстроки длины N добавляем в хэш-таблицу. Строки (n-граммы) являются ключом таблицы, значением являются частота n-граммы и левые и правые контексты.

Третий этап работы программы: фильтруем n-граммы (их частота встречаемости должна превышать 1). Далее фильтруем на основе устойчивости n-граммы (здесь используются левые и правые контексты, хранимые в хэш-таблице)

После фильтрации имеем отфильтрованный набор n-грамм и информацию об их частотах.

Результат:

при N = 4,  П = 0.5:

```
<Ngram: в жертва за грех | Freq: 63.0>
<Ngram: . И сказал я | Freq: 66.0>
<Ngram: , и никто не | Freq: 66.0>
<Ngram: . И сказал царь | Freq: 68.0>
<Ngram: , сын его , | Freq: 82.0>
<Ngram: в тот день , | Freq: 83.0>
<Ngram: , говорю Господь Бог | Freq: 88.0>
<Ngram: Господь , Бог твой | Freq: 90.0>
<Ngram: , и не есть | Freq: 91.0>
<Ngram: : вот , Я | Freq: 92.0>
<Ngram: и сказал они : | Freq: 106.0>
<Ngram: . Так говорю Господь | Freq: 109.0>
<Ngram: , говорю Господь . | Freq: 115.0>
<Ngram: , и вот , | Freq: 117.0>
<Ngram: , говорю Господь , | Freq: 118.0>
<Ngram: и сказал он : | Freq: 129.0>
<Ngram: . И сказал он | Freq: 144.0>
<Ngram: : так говорю Господь | Freq: 167.0>
<Ngram: за тот , что | Freq: 185.0>
<Ngram: , и сказал : | Freq: 193.0>
<Ngram: . И сказал Господь | Freq: 207.0>
```

при при N = 10,  П = 0.5:

```
<Ngram: ; и так истребил зол иза среда себя . Если | Freq: 4.0>
<Ngram: , сын Карея , и весь бывший с они военный | Freq: 4.0>
<Ngram: они , по семейство они , по число имен , | Freq: 4.0>
<Ngram: , когда воцарился , и шестнадцать год царствую в Иерусалиме | Freq: 5.0>
<Ngram: . Посему так говорю Господь Бог : вот , Я | Freq: 5.0>
<Ngram: : отпустил народ Мой , чтобы он совершил Мне служение | Freq: 5.0>
<Ngram: . Ибо так говорю Господь Саваоф , Бог Израилев : | Freq: 5.0>
<Ngram: . И увидел Бог , что это хорош . И | Freq: 6.0>
<Ngram: земля , который Господь , Бог твой , дает ты | Freq: 6.0>
<Ngram: , как повелел Господь Моисею . И сказал Господь Моисею | Freq: 6.0>
<Ngram: так говорю Господь Саваоф , Бог Израилев : вот , | Freq: 8.0>
<Ngram: . И сказал Господь Моисею и Аарону , говоря : | Freq: 11.0>
<Ngram: , и воинство его , вошедший в исчисление его , | Freq: 12.0>
<Ngram: с сын его и брат его ; они -- двенадцать | Freq: 22.0>
<Ngram: . И было к я слово Господне : сын человеческий | Freq: 29.0>
```