Предсказание следующего слова на основе нескольких предыдущих слов

Вознюк Анастасия, Б05-022

vozniuk.ae@phystech.edu

18 апреля 2022 г.

Цель

Цель: Построить модель, которая бы по датасету могла бы генерировать текст в схожей стилистике и со схожей лексикой с тектос, который подается на вход
Входные данные: Было взято несколько произведений английских классиков - Оскара Уайльда и Чарльза Диккенса.

Основные формулы

Считаем, что появление слов в нашем тексте - независимые события.
Хотим узнать, какова вероятность сгенерировать предложение $w_1 ... w_n$.
Из формулы полной вероятности:

$$P(w_1 ... w_n) = P(w_1) \times P(w_2|w_1) \times P(w_3|w_1 w_2) \times \cdots$$
$$\cdots \times P(w_n|w_1 ... w_{n-1}) = \prod_{i=1}^{n} P(w_i|w_1 ... w_{i-1}) \quad (1)$$

Пример:

P(capital | London is the) = P(London) $\times$ P(is|London) $\times$
$\times$ P(the| London is the)

Основные формулы

Основное утверждение - мы можем пренебречь всем контекстом, оставив лишь информацию о самом ближайшем контексте

$$P(w_n|w_1...w_{n-1}) \approx P(w_n|w_{n-1}) \tag{2}$$

или в общем случае,

$$P(w_n|w_1...w_{n-1}) \approx P(w_n|w_{n-N+1}...w_{n-1}) \tag{3}$$

где N - размер скользящего окна для нашего контекста
Тогда вероятность сгенерировать предложение $w_1...w_n$:

$$P(w_1...w_n) \approx \prod_{i=1}^{n} P(w_i|w_{i-N+1}...w_{i-1}) \tag{4}$$

Основные формулы

Вероятность 2-граммы: $P(w_2|w_1) = \dfrac{C(w_1 w_2) + \alpha}{C(w_1) + \alpha|V|}$

Вероятность 3-граммы: $P(w_3|w_1 w_2) = \dfrac{C(w_1, w_2, w_3) + \alpha}{C(w_1, w_2) + \alpha|V|}$

где

- $C(w_1)$: количество появлений слова w_1 (unigram count)
- $C(w_1, w_2)$: количество появлений 2-граммы w_1, w_2
- $C(w_1, w_2, w_3)$: количество появлений 3-граммы w_1, w_2, w_3
- $|V|$: размер словаря
- $0 \leq \alpha \leq 1$: сглаживающий гиперпараметр

Основные формулы

Вероятность 2-граммы: $P(w_2|w_1) = \dfrac{C(w_1 w_2) + \alpha}{C(w_1) + \alpha|V|}$

Вероятность 3-граммы: $P(w_3|w_1 w_2) = \dfrac{C(w_1, w_2, w_3) + \alpha}{C(w_1, w_2) + \alpha|V|}$

Если $\alpha = 0$, то сглаживания нет. В этом случае если в тренировочной выборке не встречалось (w_1, w_2), то возникает деление на 0.

Формула для сравнения правдоподобности

Чем больше **правдоподобность** у входного текста, тем больше вероятность того, что модель сгенерировала бы сама подобный текст. Сгенерируем предложение, а также возьмем предложение из тестовой выборки, причем предложения должны быть одной длины. Посяитаем вероятности для обоих предложений, а дальше посмотрим на их отношение (предварительно перейдем в логарифмический масштаб):

$$log_2(P_{test}(w_i^n)) - log_2(P_{rand}(w_i^n)) \tag{5}$$

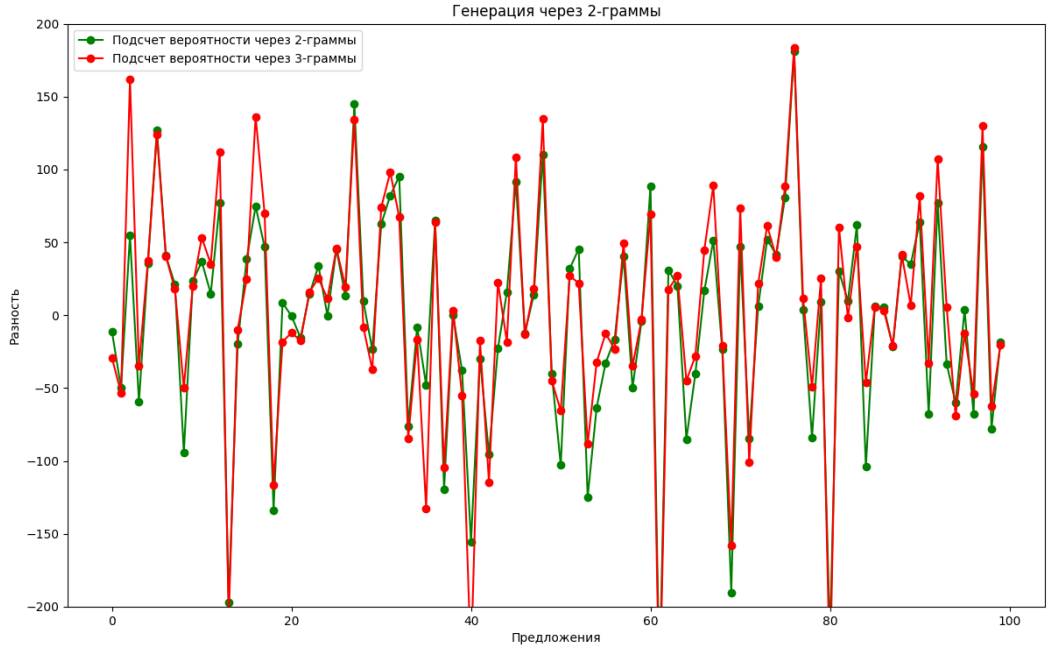
- test - берем из тестовой выборки
- random - генерируем через n-грамму

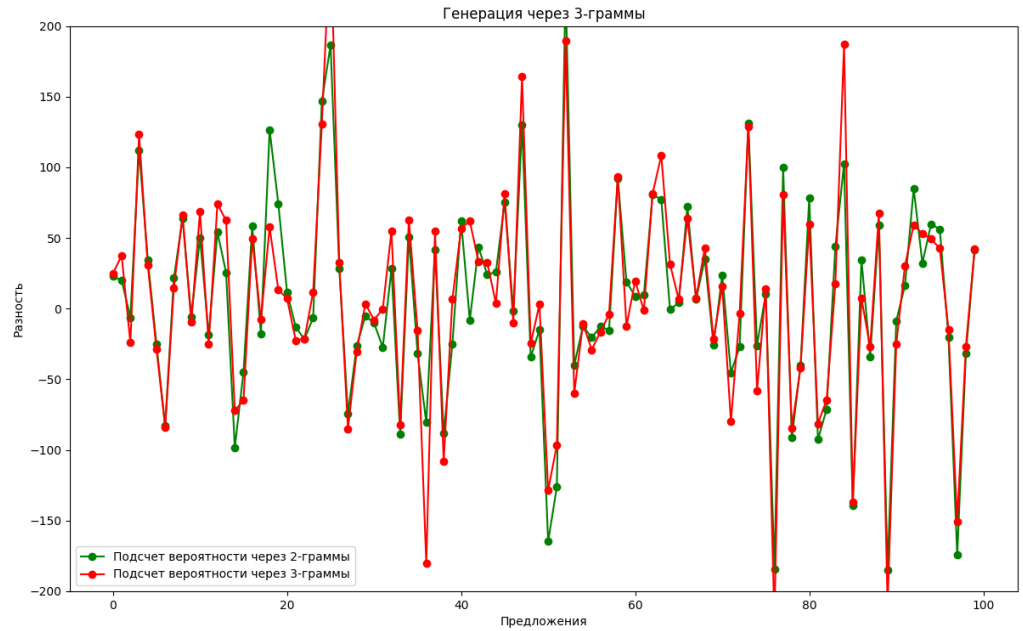
У хорошей модели данная разность будет стремиться к нулю - т.е модель генерирует такие предложения, которые близки к тестовой выборке.

Гипотеза

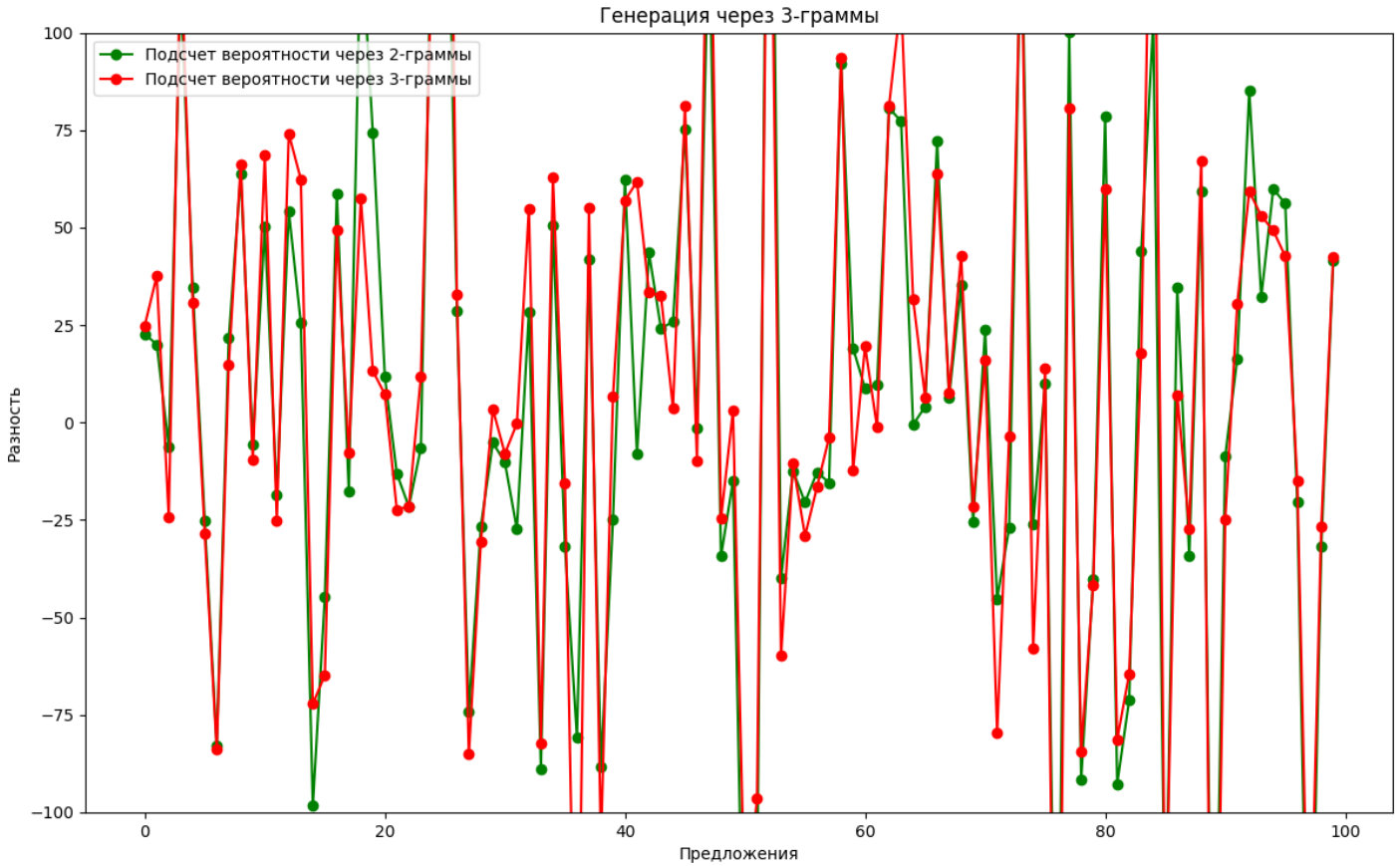
Гипотеза

У модели, построенную на 3-граммах, данная метрика будет лучше, чем у модели, построенной на 2-граммах. Модель, построенная на 1-граммах будет иметь самый низкое значение этой метрики


Генерация через 2-граммы
Подсчет вероятности через 2-граммы
Подсчет вероятности через 3-граммы
Разность
Предложения
200
150
100
50
0
−50
−100
−150
−200
0
20
40
60
80
100

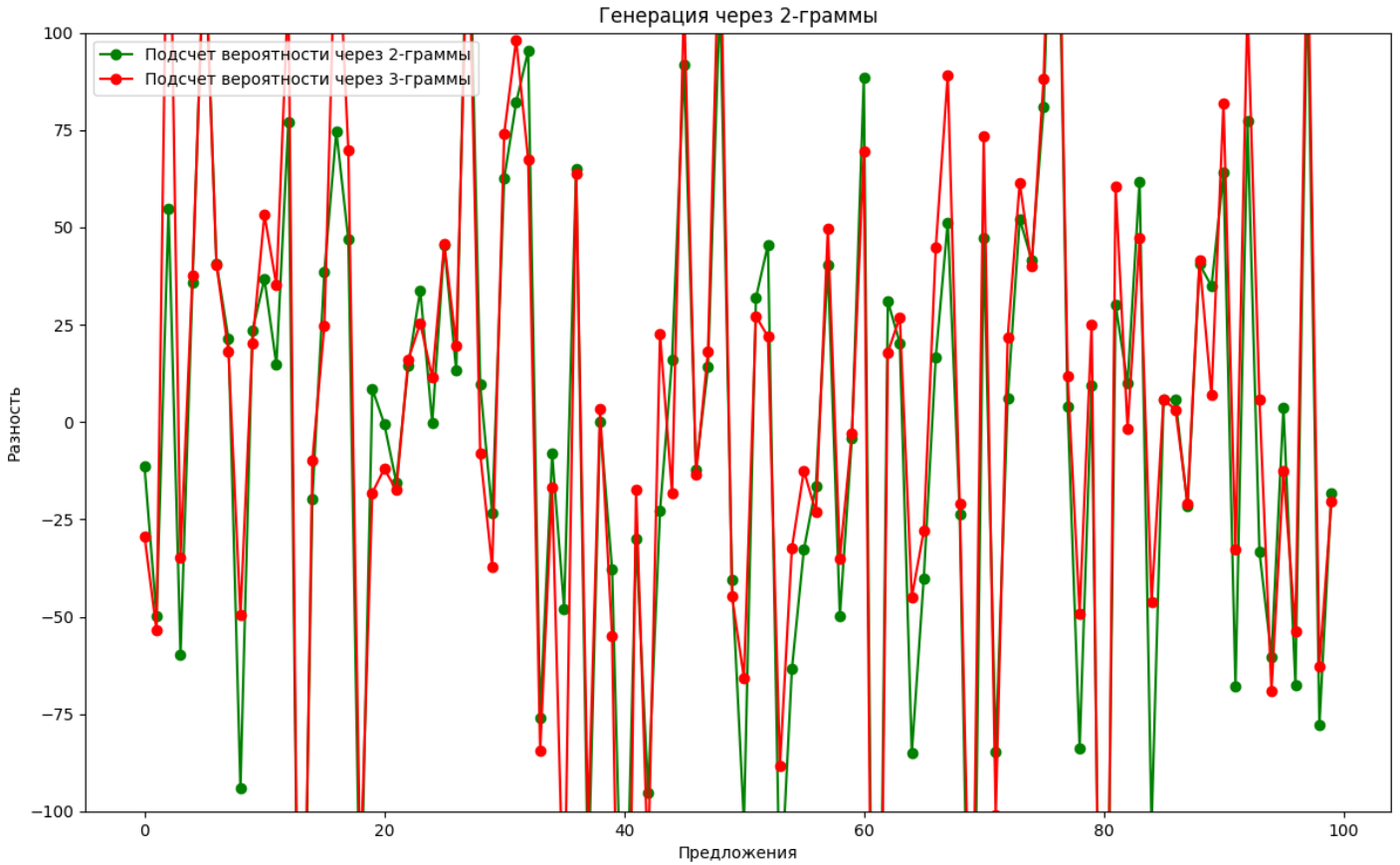

Генерация через 3-граммы
Подсчет вероятности через 2-граммы
Подсчет вероятности через 3-граммы
Разность
Предложения
200
150
100
50
0
−50
−100
−150
−200
0
20
40
60
80
100

3-граммы



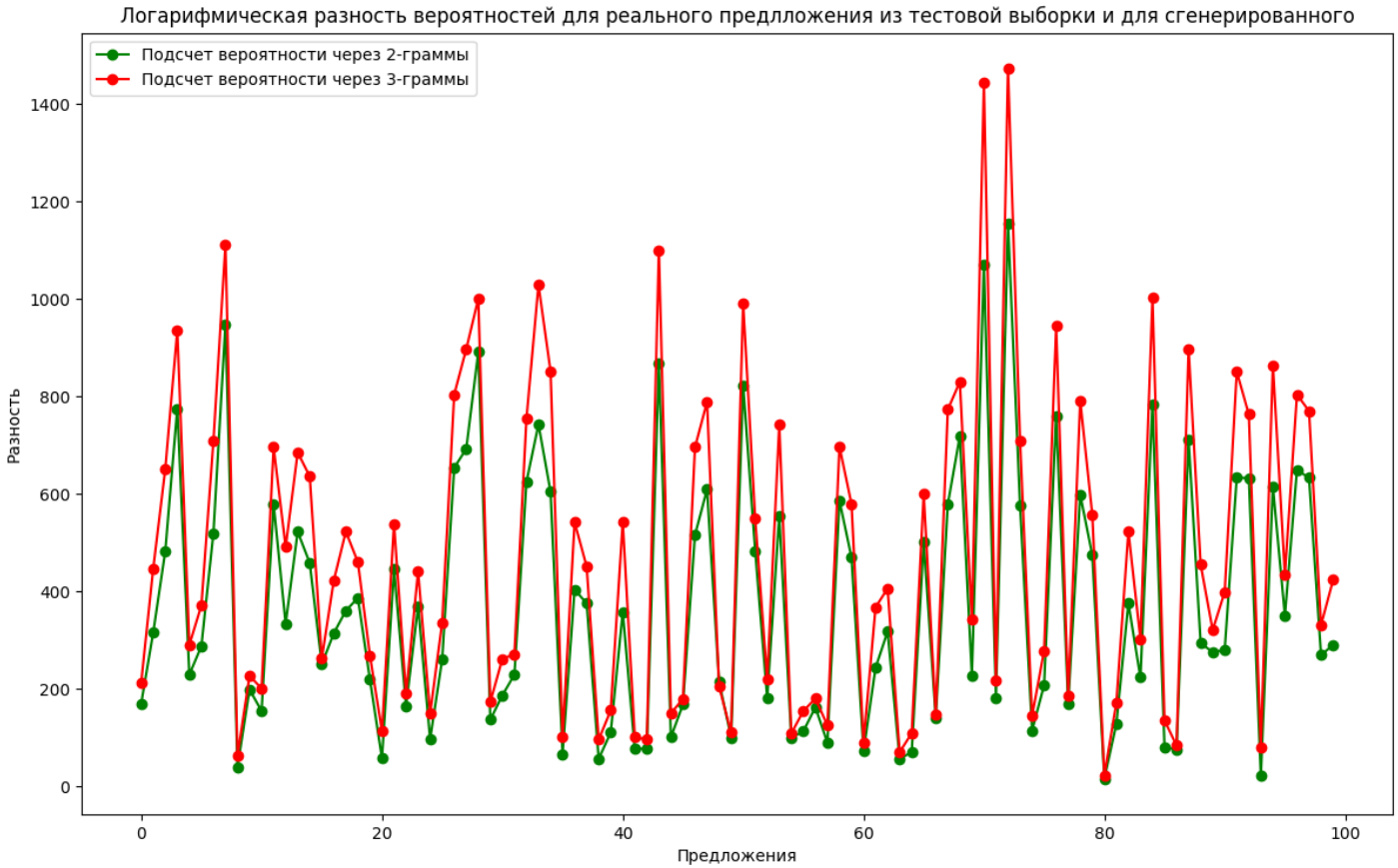
- You must not say these extravagant things to me.
- He is not often wrong about things you don't think, so well he seemed to him how vile and degraded.

2-граммы



- For shame was drawn you to talk to me in this grey sun-bleached pillars loitered upon it
- Every one use your beauty and satisfied.

1-граммы



- That consisted perhaps windows tell hated true he is daughter's good is wore care dominate it his you slowly hubbard grew do he.
- Woman really carefully shall must.

Perplexity

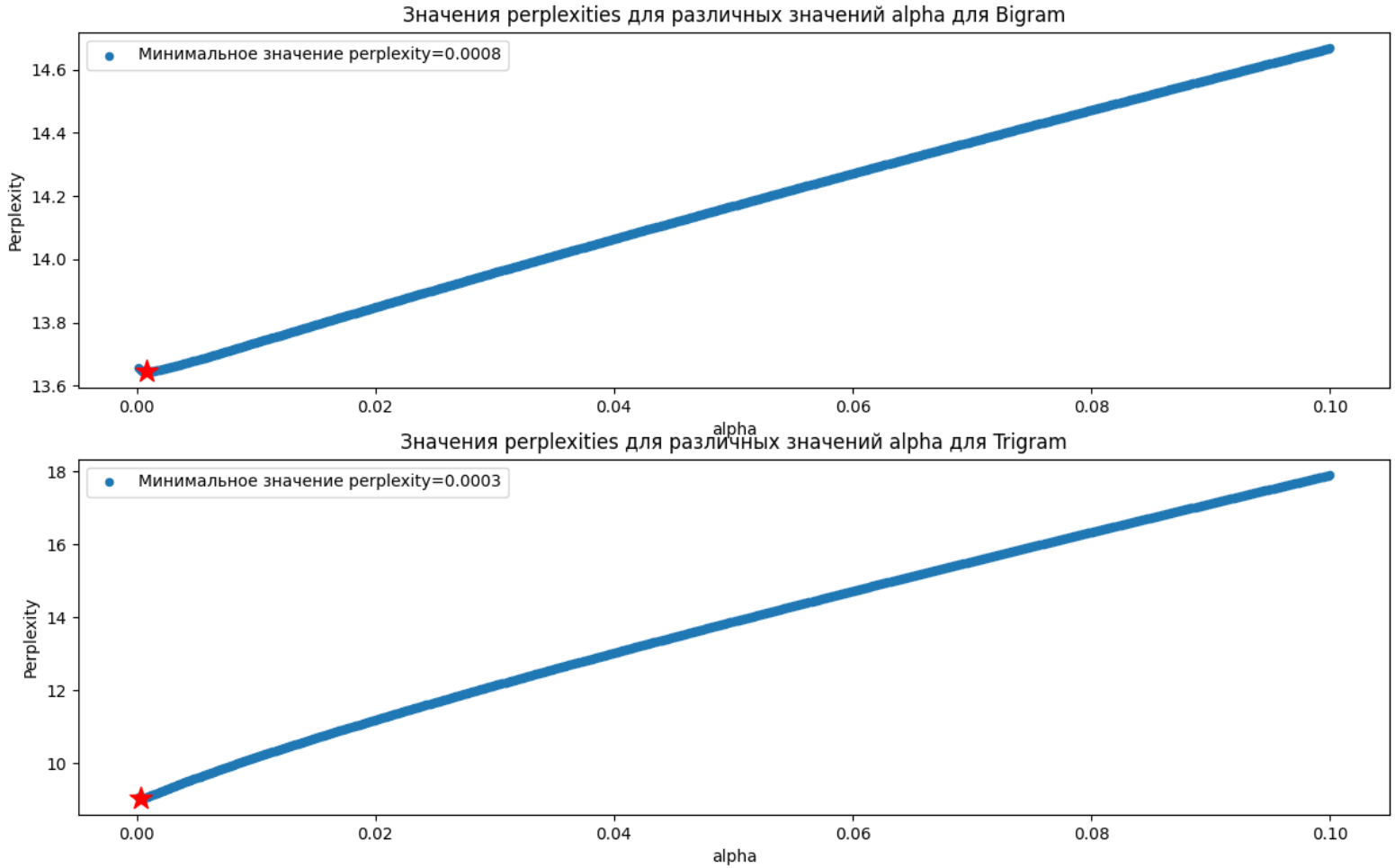
Качество модели будем оценивать по тому, насколько "удивлена" была модель, увидев текст $w_1..w_n$
Введем perplexity - нормализованное обратное значение к вероятности

$$PP(w_1..w_n) = \sqrt[N]{\frac{1}{P(w_1...w_n)}}$$

Это будет критерием качества нашей модели - чем меньше perplexity, тем больше $P(w_1...w_n)$, а значит тем меньше "модель была удивлена увидев наше словосочетание
С учетом наших предположений:

$$PP(w_1..w_n) = \sqrt[N]{\frac{1}{\prod_{i=1}^{n} P(w_i|w_{i-N+1}...w_{i-1})}}$$

Минимизация perplexity



Качество модели

Таким образом, для 2-граммы при a = 0.001 достигается минимальное значение perplexity равное **13.644560245452233**, а для 3-граммы при a = 0.003 достигается минимальное значение равное **9.04139960414555**. При увеличении параметра a значение perplexity монотонно увеличивается

Результаты эксперимента

- Из-за того, что данных немного, 2-граммы и 3-граммы вели себя почти одинаково.
- N-граммы для $N \geq 1$ показывают лучше результаты, чем 1-граммы
- При сглаживании нужно брать параметры сглаживания, сильно меньшие чем 1, однако на данной выборке значение perplexity все равно было достаточно большим.