Отчёт по практическому заданию "Униграммная и биграммная языковые модели со сглаживанием"

Выполнил: Алексей Грищенко, 209 группа

Постановка задачи

Построить биграммную и униграммную модель и оценить ее качество с помощью перплексии на заданном кусочке текста при различных параметрах сглаживания.

Используемые формулы

Пусть дан тестовый корпус: $w_1 w_2 ... w_n$ Вычисление вероятности для униграммы осуществляется по следующей формуле:

$$P(w_i) = \frac{c(w_i) + 1 * \lambda}{N + \lambda * V}$$

где $c(w_i)$ - частота униграммы в обучающем корпусе, N - количество слов в обучающем корпусе, V - количество слов в словаре, λ - параметр сглаживания.

Вычисление вероятности для биграммы вычисляется по следующей формуле:

$$P(w_i|w_{i-1}) = \frac{c(w_i w_{i-1}) + 1 * \lambda}{c(w_{i-1}) + \lambda * V^2}$$

где $c(w_i w_{i-1})$ - частота подстроки $w_i w_{i-1}$

Как уже было сказано выше, оценка качества языковой модели будет производиться с помощью перплексии. По сути, этот показатель говорит о том, какое количество вариантов в среднем рассматривает наша языковая модель на каждом шаге.

Перплексия для униграммной и биграммной моделей вычисляется по следующем формулам:

$$P_{uni} = \sqrt[n]{\prod_{i=1}^{n} 1/P(w_i)}$$

$$P_{bi} = \sqrt[n]{\prod_{i=1}^{n} 1/P(w_i w_{i-1})}$$

Алгоритм работы программы

Рассмотрим алгоритм работы биграммной языковой модели. В качестве корпуса возьмем стих Самуила Маршака "Дом, который построил Джек". Из этого корпуса выделим два непересекающихся части - одна из этих частей будет использоваться для обучения нашей модели, вторая для оценки качества. Эти части хранятся в фаилах train.txt и test.txt соответственно. Первым делом программа сканирует весь стих, для того, чтобы выделить для себя уникальные слова. Далее программа сканирует обучающий корпус.

В процессе сканирования в соответствующих структурах данных происходит заполнение информации о частотах уникальных слов и биграмм, встреченных в обучающем корпусе. На последнем этапе идет вычисление перплексии за счёт полученных данных при различных параметрах лямбда. Аналогичный алгоритм применяется в униграммной языковой модели.

Результаты

В результате получаем следующую зависимость перплексии от параметра сглаживания (см приложения ниже). Из этих графиков следует, что наиболее точная модель получается при при близком к нулю параметре лямбда, т.к в таком случае распределение вероятности на униграммы/биграммы ни разу не встречавшиеся в обучающем корпусе минимальны.

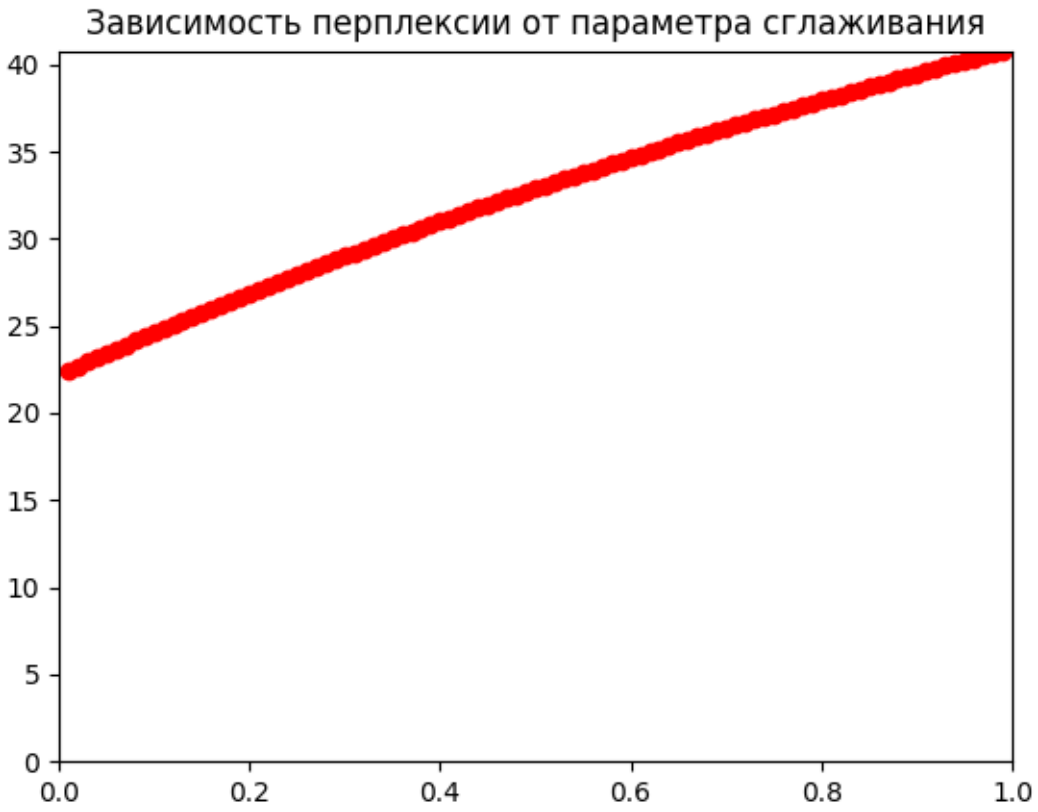

Рис. 1: Зависимость перплексии униграммной модели от параметра сглаживания

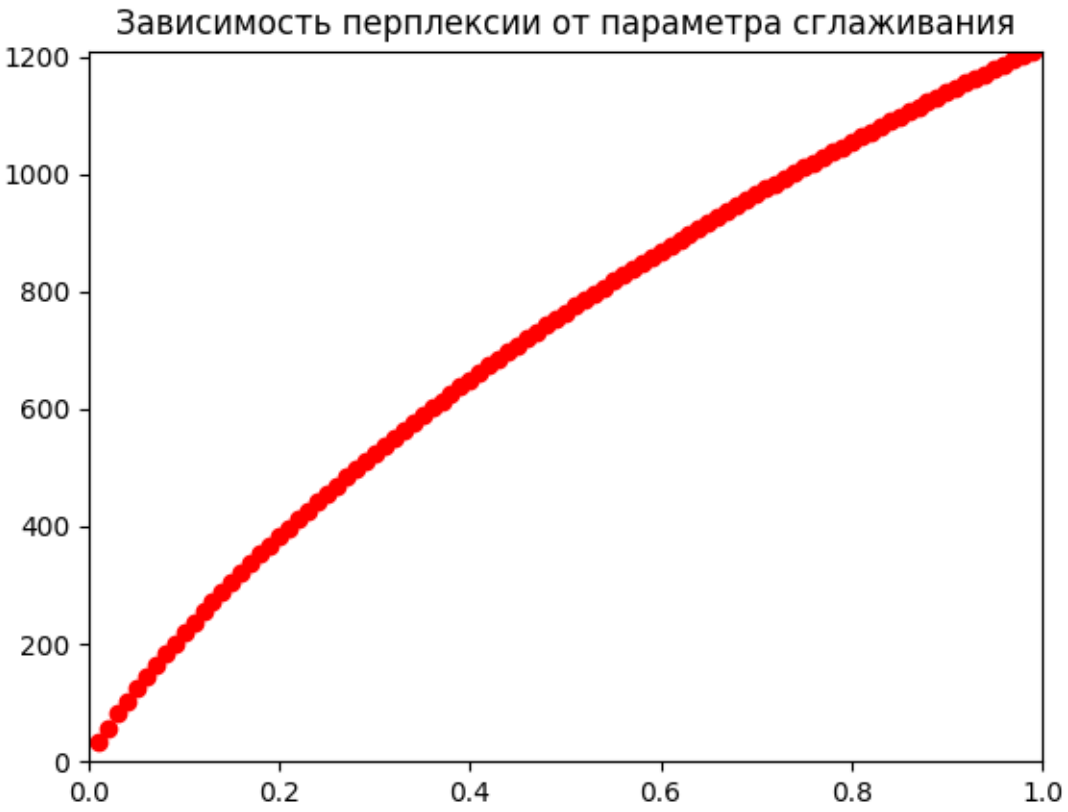

Рис. 2: Зависимость перплексии биграммной модели от параметра сглаживания