МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМЕНИ Н. Э. БАУМАНА
Факультет информатики и систем управления
Кафедра теоретической информатики и компьютерных технологий

Лабораторная работа №3
по курсу «Автоматическая обработка текстов»

«Исправление ошибок написания»

Выполнил:
студент группы ИУ9-11М
Беляев А. В.

Проверила:
Лукашевич Н. В.

Москва 2018

1 Цель работы

Выяснить, какое исходное слово наиболее вероятно путем расчета вероятности с аддитивным сглаживанием и без него.

Исходные данные:

- Частота буквы «s» – k. Тогда частота буквы «o» – $1.2k$.
- Частота слова «hostel» – c. Тогда частота слова «hotel» – $5c$.
- Количество ошибок типа «замена» для $o \Rightarrow od$: 9. Количество ошибок типа «замена» для $s \Rightarrow d$: 7.

2 Вычисления

Необходимо вычислить значение по следующей формуле и сравнить для обоих слов – hotel и hostel:

$$P = P(x|w)P(w),$$

где x – слово с ошибкой, а w – верное слово.

Без сглаживания формула $P(x|w)$ имеет следующий вид:

$$P(x|w) = \frac{sub(x_i, w_i)}{count(w_i)},$$

где w_i и x_i - буквы слов w и x соответственно

Подставим соответствующие значения для обоих слов:

$$P(hodtel|hotel) = \frac{sub(od, o)}{count(o)} = \frac{9}{1.2k}$$

$$P(hodtel|hostel) = \frac{sub(d, s)}{count(s)} = \frac{7}{k}$$

$P(w)$ - вероятность верного слова в корпусе:

$$P(hostel) = c$$

$$P(hotel) = 5c$$

Вычислим $P(x|w)P(w)$ для слова «hotel»:

$$P(x|w)P(w) = 5c\frac{9}{1.2k}$$

Для слова «hostel»:

$$P(x|w)P(w) = c\frac{7}{k}$$

После сокращений имеем $P(hotel) = 37.5$ и $P(hostel) = 7$

Таким образом, более вероятное исходное слово – hotel.

Аддитивное сглаживание:

Формула для ааддитивного сглаживания следующая:

$$P(x|w) = \frac{sub(x_i, w_i) + 1}{count(w_i) + |A|},$$

где $|A|$ – мощность соответствующего алфавита. В случае английского языка $|A| = 26$

Очевидно, что прибавление 1 к числителю и 26 к знаменателю $P(hodtel|hotel)$ и $P(hostel|hotel)$ ничего не изменит и слово hotel будет все так же наиболее вероятным исходным словом.

3 Выводы

В ходе работы были проведены вычисления наиболее вероятных исходных слов методами с и без аддитивного сглаживания. Вычисления показали, что слово hotel является наиболее вероятным кандидатом в обоих случаях.