

Комп'ютерний практикум 1

Мета

Дослідити і навчитись працювати та визначати ентропію на символ джерела

Аналіз тексту

1. Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку H_1 та H_2 за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення H_1 та H_2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення H_1 та H_2 на тому ж тексті, в якому вилучено всі пробіли.

Для аналізу було взято переклад книги Mein Kampf.

Для зменшення впливу скорочень, спецсимволів і чисельно-символьно-буквених сполук, що не є “чистими” словами, за допомогою регексів з текстів видалено: дати виду 1927, 1934-36 гг., 1920-е, 1936 г.; числа виду 12%, 12-13, 12/13; ініціали виду А., К.С. та інші символи, що не входять в рядок

Також замінено великі літери на малі та ‘ё’ на ‘е’, бо в повсякденному житті ‘ё’ майже не використовується. ‘ь’ та ‘Ъ’ не замінювались, бо правила вживання ‘ь’ можуть допомогти при аналізі.

Обсяг зі space-ом – 1 570 727 символів, без space-а – 1 354 233 символів

Аналіз літер

```
letters - with ' '
' ' 0.1378310808 216495
'o' 0.0967443738 151959
'e' 0.0747545563 117419
'и' 0.0694296335 109055
'a' 0.0625538365 98255
'н' 0.0615453863 96671
'т' 0.0576064459 90484
'с' 0.0504390642 79226
'р' 0.0421626419 66226
'в' 0.0401412849 63051
-- H1 = 4.4111

letters - without ' '
'o' 0.1122104632 151959
'e' 0.0867052322 117419
'и' 0.0805290379 109055
'a' 0.072554038 98255
'н' 0.0713843714 96671
'т' 0.0668157302 90484
'с' 0.0585025313 79226
'р' 0.0489029945 66226
'в' 0.0465584922 63051
'л' 0.0370099067 50120
-- H1 = 4.4453
```

Аналіз біграм з перетином (crs) і без (seq)

```
bgs, crs - with ' '
'и ' 0.017563216 27587
'o ' 0.0174976412 27484
'e ' 0.015983055 25105
'ст' 0.0153260339 24073
'п' 0.0150592783 23654
'н' 0.014145688 22219
'с' 0.0138025346 21680
'в' 0.0135262293 21246
'но' 0.0130156374 20444
'на' 0.0118130088 18555
-- H2 = 3.9868

bgs, crs - without ' '
'ст' 0.0179422861 24298
'но' 0.0151694947 20543
'ен' 0.0142604917 19312
'то' 0.013770915 18649
'на' 0.0137266094 18589
'ни' 0.012039305 16304
'ос' 0.0119270641 16152
'ов' 0.0118480525 16045
'ро' 0.010688723 14475
'пр' 0.0100795211 13650
-- H2 = 4.1008
```

```
bgs, seq - with ' '
'и ' 0.017642797 13856
'o ' 0.0173779513 13648
'e ' 0.0159620456 12536
'ст' 0.0153292172 12039
'п' 0.0148568242 11668
'н' 0.0141539645 11116
'с' 0.0139731564 10974
'в' 0.0135084031 10609
'но' 0.0129838049 10197
'на' 0.0119116892 9355
-- H2 = 3.9868

bgs, seq - without ' '
'ст' 0.0180072543 12193
'но' 0.0152809858 10347
'ен' 0.0142516201 9650
'на' 0.0137613053 9318
'то' 0.0137037081 9279
'ос' 0.0119920368 8120
'ни' 0.0118768424 8042
'ов' 0.0117631248 7965
'ро' 0.0108120322 7321
'пр' 0.0100617915 6813
-- H2 = 4.1001
```

Видно, що зі space-ами H в 1- і 2-граммі менше, бо space символ зустрічається помітно частіше за інші і, відповідно, якщо починати вгадувати у малих n -грамах з нього, то шанс вгадати найвищий.

Також, через великий обсяг тексту, рахування з перетином і без мають майже однакові результати

Cool pink program

Лабораторная работа №1

Произвольная часть текста:
существа_во_всех_ч

Использованные буквы:

Порядок n-граммы:
5 символов
10 символов
15 символов
20 символов
25 символов
30 символов
35 символов
40 символов
45 символов
50 символов

Введенный символ:

Символ по счету:

Номер эксперимента: 51

Поле ввода символов:

Продолжить Другой

Неравенство для энтропии:
 $1,27842396434638 < H < 1,97945832095488$

Двоичная таблица угаданных символов:

01000000000000000000000000000000	▲
10000000000000000000000000000000	
10000000000000000000000000000000	
01000000000000000000000000000000	
00000010000000000000000000000000	▼

Вероятности:

$q[1] = 0,66$
$q[2] = 0,08$
$q[3] = 0,04$
$q[4] = 0$
$q[5] = 0,02$
$q[6] = 0$
$q[7] = 0,06$
$q[8] = 0,02$
$q[9] = 0,04$
$q[10] = 0$
$q[11] = 0,02$
$q[12] = 0,02$
$q[13] = 0$
$q[14] = 0$
$q[15] = 0,02$
$q[16] = 0$
$q[17] = 0$
$q[18] = 0$
$q[19] = 0$
$q[20] = 0$
$q[21] = 0$
$q[22] = 0,02$
$q[23] = 0$
$q[24] = 0$
$q[25] = 0$
$q[26] = 0$
$q[27] = 0$
$q[28] = 0$
$q[29] = 0$
$q[30] = 0$
$q[31] = 0$
$q[32] = 0$

Строка состояния:

[illegible]

Произвольная часть текста:
том_роде_относительно_чего_они_оба_согласны_точно_так_же_не_имело_бы_смысла

Использованные буквы:
_

Порядок n-граммы:
 5 символов
 10 символов
 15 символов
 20 символов
 25 символов
 30 символов
 35 символов
 40 символов
 45 символов
 50 символов

Введенный символ: и

Символ по счету: 2

Номер эксперимента: 77

Неравенство для энтропии:
 $1,05357592485804 < H < 1,72910527706972$

Двоичная таблица угаданных символов:

10000000000000000000000000000000
10000000000000000000000000000000
00000000000000000000000000000000
10000000000000000000000000000000
00100000000000000000000000000000

Вероятности:

q[1] = 0,6363636
q[2] = 0,1558441
q[3] = 0,0909090
q[4] = 0,0649350
q[5] = 0
q[6] = 0
q[7] = 0,0129870
q[8] = 0
q[9] = 0
q[10] = 0
q[11] = 0
q[12] = 0
q[13] = 0
q[14] = 0
q[15] = 0,012987
q[16] = 0
q[17] = 0
q[18] = 0
q[19] = 0
q[20] = 0
q[21] = 0,012987
q[22] = 0
q[23] = 0
q[24] = 0
q[25] = 0
q[26] = 0
q[27] = 0
q[28] = 0
q[29] = 0
q[30] = 0
q[31] = 0,012987
q[32] = 0

Строка состояния:
Вы угадали. Для продолжения опыта нажмите "Продолжить", или "Другой" для выбора другого порядка

Продолжить Другой

Помітно, що зі збільшенням довжини n-грамів, збільшується "швидкість" вгадування і зменшується H .

Також, для великих n збільшення довжини n-грамів до $(n+k)$ дає відносно невеликий приріст шансу вгадування не пробільних символів (не '_'). Якщо ж треба вгадати наступний символ після '_', то це зробити приблизно однаково складно для будь-яких великих n .

Надлишковість

Нехай $H_\infty \sim H^{(30)}, H^{(30)} \in (1.055; 1.729), H_0 = \log_2 32 = 5 \Rightarrow$

$$R \in \left(\left(1 - \frac{1.729}{5} \right); \left(1 - \frac{1.055}{5} \right) \right) \Rightarrow R \in (0.65; 0.79)$$

Висновки

Було цікаво :)