Міністерство освіти і науки України Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського" Фізико-технічний інститут

Лабораторна робота №1ЕКСПЕРИМЕНТАЛЬНА ОЦІНКА ЕНТРОПІЇ НА СИМВОЛ ДЖЕРЕЛА ВІДКРИТОГО ТЕКСТУ

Виконали: студенти гр. ФІ-94, Молдован Д.В., Пелешенко Л.І.

Перевірив: Чорний О.М.

Мета лабораторної роботи:

Засвоєння понять ентропії на символ джерела та його надлишковості, вивчення та порівняння різних моделей джерела відкритого тексту для наближеного визначення ентропії, набуття практичних навичок щодо оцінки ентропії на символ джерела.

Постановка задачі:

- 0. Уважно прочитати методичні вказівки до виконання комп'ютерного практикуму.
- 1. Написати програми для підрахунку частот букв і частот біграм в тексті, а такожпідрахункуН1таН2забезпосереднім означенням. Підрахувати частоти букв та біграм, а також значенняН1таН2на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення Н1 та Н2 на тому ж тексті, в якому вилучено всі пробіли.
- 2. За допомогою програми CoolPinkProgram оцінити значення (10)H ,(20)H ,(30)H.
- 3. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

Опис труднощів:

- Виникали проблеми з формулою для обрахунку ентропії: складно було розібратись та спочатку вона була реалізована з помилками.
- Функції, які обраховують частоти для біграм з перетином і без перетину майже ідентичні (відрізняється лише рядок з циклом). Тому в коді є фрагмент, який двічі повторюється (за винятком рядку з циклом). Існують способи вирішення такої проблеми, адже дана реалізація робить код більш громіздким, але він все рівно є робочим.
- Пошук найкращих біграм робився просто сортуванням, реалізація якого була знайдена в мережі інтернет.

Результат роботи програми:

Ентропія тексту з пробілами;

Ентропія тексту з пробілами та перетином;

Ентропія тексту з пробілами та без перетину;

Ентропія тексту без пробілів;

Ентропія тексту без пробілів та перетину;

Ентропія тексту без пробілів та без перетину відповідно.

```
H1 with gaps = 4.447975437586149

H2 with gaps with intersection = 3.974999153773363

H2 with gaps without intersection = 3.933141321045024

H1 no gaps = 4.421720836108205

H2 no gaps with intersection = 4.020811911404876

H2 no gaps without intersection = 3.974883235762987
```

Біграми, які трапляються найчастіше:

```
Best bigrams [
    ['cT', 0.02063524175601861],
    ['ho', 0.017600647380133522],
    ['pa', 0.013554521545620068],
    ['ha', 0.011936071211814687],
    ['TO', 0.011733764920089014],
    ['oB', 0.011531458628363342]
]
```

Алфавіт без пробіла:

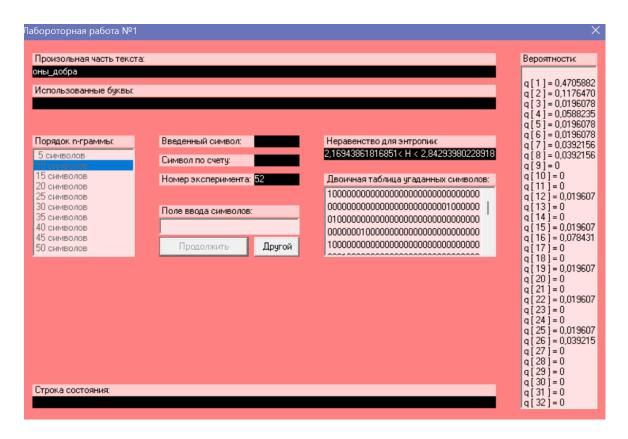
Частота літер:

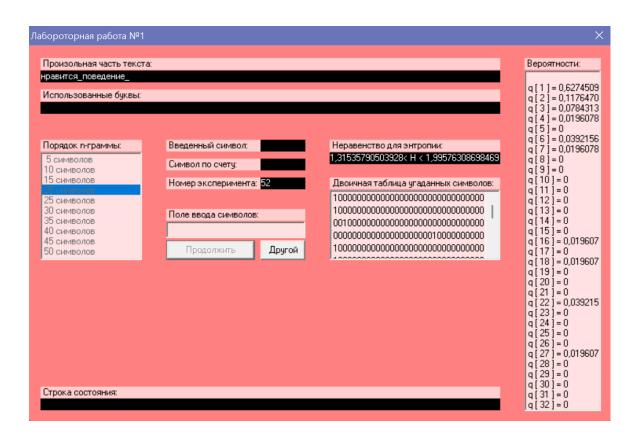
```
Table for literals {
  a': 0.07605177993527508,
  'r': 0.014563106796116505,
  'o': 0.1063915857605178,
  'д': 0.030137540453074433,
  'ы': 0.019417475728155338,
  'm': 0.0056634304207119745.
  'n': 0.040453074433656956,
  'n': 0.08272653721682847,
  '6': 0.011326860841423949,
  'c': 0.061893203883495146.
  'T': 0.06897249190938512
  'p': 0.0505663430420712,
  'H': 0.074231391585760
  'e': 0.08980582524271845,
  'K': 0.028923948220064725
  'n': 0.022451456310679612,
     : 0.007079288025889968
  'B': 0.041868932038834954,
  'M': 0.030946601941747573,
  'ь': 0.0182038834951456
  'a': 0.01719255663430420
  'й': 0.012540453074433657,
  'y': 0.02063106796116505,
  'x': 0.011124595469255663,
  'u': 0.007281553398058253,
  'щ': 0.0034385113268608413,
  'я': 0.01779935275080906,
  '4': 0.013349514563106795,
  'ю': 0.007483818770226537,
  ''': 0.00465210355987055,
  'φ': 0.0028317152103559872
```

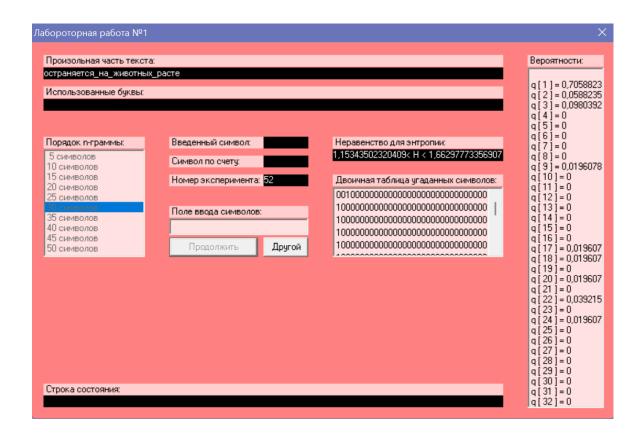
Частота біграм з перетином:

```
0.6
                                                                                               2023062917256726
                                                                                       0.0
'XH': 0.6
                                         ур': 0.0
тд': 0.0
                                                                                       0.0
      0.0
                                                                                       0.0
                                                                                       0.0
                                                                                       0.0
      0.6
                                                                                яр': 0.00
иф': 0.00
                                                                                яр
                                        яю': 0.6
                                                                                       0.0
                                                                                мэ '
                                                                                       0.0
'ьл': 0.6
                                                                                       0.0
                                        MX': 0.6
                                                                                       0.0
                                                                                       0.0
                                                                                       0.0
                                                                                фи':
   : 0.
                                        '∋B': 0.0
                                                                                     : 0.0
                                                                                       0.00
                                        'зя': 0.
                                                                                       0.00
                                                                                xφ': 0.00
тб': 0.6
                                         TT': 0.6
                                        'йч': 0.6
                                                                                φp': 0.0
                                        юв': 0.6
                                                                                3c':
                                                                                       0.0
ую': 0.
                                                                                    : 0.0
юп': 0.6
                                                                                o9': 0.00
                                                                                       0.0
    : 0.
                                                                                нэ': 0.0
                                                                                ЙЯ: 0.00020230629172567267,
Яе': 0.00040461258345134534,
Яяу': 0.00020230629172567267
вт':
      0.
                                              0.6
B9': 0.6
                                        сы': 0.6
                                         кц': 0.6
```

CoolPinkProgram:







Оцінки надлишковості:

Значення ентропії на алфавіті з пробілом $H_0 = 5$; Значення ентропії на алфавіті без пробілу $H_0 = 4,954$;

$$R = 1 - \frac{H_n}{H_0}$$

<u>Для Н₁:</u>

R = 0.11 на алфавіті з пробілами;

R = 0.108 на алфавіті без пробілів;

<u>Для *H*₂:</u>

R = 0.205 на алфавіті з пробілами з перетинами;

R = 0.213 на алфавіті з пробілами без перетинів;

R = 0.188 на алфавіті без пробілів з перетинами;

R = 0.198 на алфавіті без пробілів без перетинів;

Для H⁽¹⁰⁾:

 $0,431 \le R \le 0,566$

Для H⁽²⁰⁾:

 $0,601 \le R \le 0,737$

Для H⁽³⁰⁾:

 $0,667 \le R \le 0,769$

Оцінка надлишковості російської мови:

$$R = 1 - \frac{H_{\infty}}{H_0} = 1 - \frac{4,447975}{5} = 0,1104$$

Висновки

Під час виконання даної лабораторної роботи було обчислено частоти букв та біграм в тексті, а також значення ентропії за безпосереднім означенням. Як і очікувалось, отримані значення ентропій біграм на алфавіті з пробілами та без (а також з перетином та без) відрізнялись мінімально.

Обчислений результат надлишковості російської мови становить більше 11%