ааМіністерство освіти і науки України Національний технічний університет України "Київський політехнічний інститут ім. Ігоря Сікорського" Фізико-технічний інститут

Лабораторна робота № 1

«Експериментальна оцінка ентропії на символ джерела відкритого тексту»

Виконав:

Студент 3 курсу, Групи ФБ-04 Швидкий Максим

Мета 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.

Порядок виконання роботи

- 0. Уважно прочитати методичні вказівки до виконання комп'ютерного практикуму.
- 1. Написати програми для підрахунку частот букв і частот біграм в тексті, а також

підрахунку 1Н та 2Н за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення 1Н та 2Н на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення 1Н та 2Н 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.

Хід роботи

Зчитав текст, відсіяв усе що не належить алфавіту, те що отримав — записав у файл, також записав у файл текст без пробілів. Далі знайшов частоту, ентропію для кожної моно та біграми. Потім порахував загальну ентропію та надлишковість. Результати записав у тхт файли.

Результати

Монограми:

Ентропія з пробілами: 4.411377794286419 Ентропія без пробілів: 4.4765458333728185

Надлишковість з пробілами: 0.10956742165472799 Надлишковість без пробілів: 0.09641331249079155

Біграми без перетину:

Ентропія з пробілами: 4.013315834081912 Ентропія без пробілів: 4.163563029343598

Надлишковість з пробілами: 0.1899158647250877 Надлишковість без пробілів: 0.15958860559999422

Біграми з перетином:

Ентропія з пробілами: 4.013418153632705 Ентропія без пробілів: 4.163289764393376

Надлишковість з пробілами: 0.18989521161722067 Надлишковість без пробілів: 0.15964376388059132

Монограми:

| ≣m | ono_s.txt | ≡ mono.txt |
|----|--------------------------|-------------------------------------|
| 1 | в: 0.0417504397021252 | 1 в: 0.03541366683307026 |
| 2 | л: 0.0492908633810627 | 2 л: 0.04180962466851538 |
| 3 | a : 0.07778730323967763 | 3 a: 0.06598094919303875 |
| 4 | д: 0.029630990409901036 | 4 д: 0.025133675951602524 |
| 5 | и : 0.06656766990192176 | 5 и: 0.05646420254684069 |
| 6 | м : 0.036615195152900674 | 6 м: 0.03105783631080344 |
| 7 | p : 0.04425767735219375 | 7 p: 0.037540362490513554 |
| 8 | н : 0.06650473381436917 | 8 : 0.15177739238833404 |
| 9 | 6 : 0.017256394816787948 | 9 н: 0.05641081873454394 |
| 10 | o: 0.11239194554157354 | 10 6: 0.014637264209472308 |
| 11 | к: 0.03686353755243257 | 11 o: 0.09533338912182186 |
| 12 | т : 0.058232890739550056 | 12 к: 0.03126848594851492 |
| 13 | e : 0.08468816019105355 | 13 т: 0.04939445443186639 |
| 14 | c: 0.05095611823819778 | 14 e: 0.07183441207108993 |
| 15 | я : 0.02265188859291918 | 15 c: 0.043222131485772496 |
| 16 | з: 0.016426318851229294 | 16 я: 0.01921384400961486 |
| 17 | п : 0.02674953818519539 | 17 з: 0.013933175009450377 |
| 18 | ч : 0.015376817066906164 | 18 п: 0.022689563031854265 |
| 19 | г: 0.017776893270601363 | 19 ч: 0.013042963869258715 |
| 20 | x : 0.0090032624707007 | 20 г: 0.015078762765223765 |
| 21 | y: 0.027395908814114012 | 21 x: 0.007636770769909998 |
| 22 | ж : 0.009596902864102277 | 22 y: 0.02323782921219921 |
| 23 | ф: 0.0020088518756655066 | 23 ж: 0.008140309972384699 |
| 24 | ь: 0.01876345896737189 | 24 ф : 0.0017039535762825822 |
| 25 | щ: 0.003915645230975442 | 25 ь: 0.015915590093118683 |
| 26 | э: 0.0026637273812804264 | 26 щ: 0.003321338808300173 |
| 27 | ш: 0.008356891841782078 | 27 э: 0.0022594337853162776 |
| 28 | й : 0.015060435653803892 | 28 ш: 0.007088504589565052 |
| 29 | ю: 0.006943381334857408 | 29 й: 0.012774602002037241 |
| 30 | ы: 0.020532473320201805 | 30 ю: 0.00588953302149492 |
| 31 | ц: 0.003983684244545822 | 31 ы: 0.017416108060378535 |
| 32 | | 32 ц: 0.003379051037810167 |
| | | 33 |

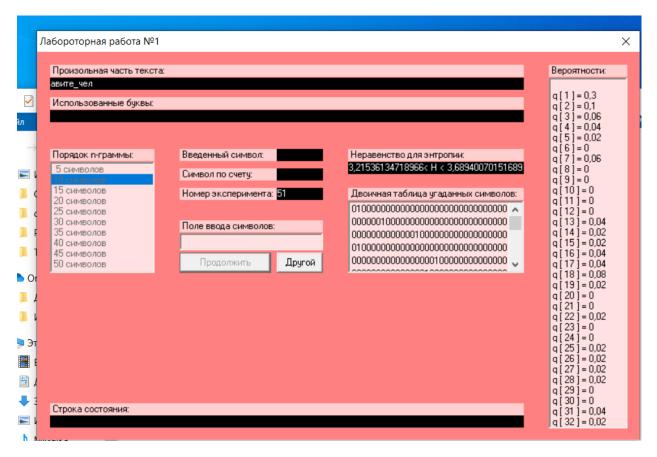
Біграми: Перехресні:

| ≣ big | jr_c_ | s.t | ext | ≣ b | igr_c.t | xt | |
|-------|-------|-----|------------------------|-----|---------|----|-----------------------|
| 1 | вл | : | 0.001110738785875757 | 1 | | | 0.0007820018381371619 |
| 2 | ла | : | 0.007730946067083179 | 2 | ла | : | 0.00651283449695784 |
| 3 | ад | : | 0.0031212950567871584 | 3 | ад | : | 0.0019636614422595526 |
| 4 | ди | : | 0.0025820849570588047 | 4 | ди | : | 0.0021266987258564147 |
| 5 | им | : | 0.004657278400808305 | 5 | ИМ | : | 0.0031352213916458543 |
| 6 | ми | : | 0.004951547635044914 | 6 | МИ | : | 0.003575277776575438 |
| 7 | ир | : | 0.001573404865137941 | 7 | ир | : | 0.0009738952781228493 |
| 8 | рн | : | 0.0014628412800201396 | 8 | р | : | 0.0008469281900120185 |
| 9 | на | : | 0.012061636647235825 | 9 | н | : | 0.012980941951513 |
| 10 | аб | : | 0.0015546941045795437 | 10 | на | : | 0.01020497970690802 |
| 11 | бо | : | 0.0025735800658958966 | 11 | аб | : | 0.000743046027012248 |
| 12 | ок | : | 0.0038203971103781785 | 12 | бо | : | 0.002172868576078535 |
| 13 | ко | : | 0.011263877856155074 | 13 | ок | : | 0.0023979465959113712 |
| 14 | ОВ | : | 0.009625835818179034 | 14 | ко | : | 0.009327752552687734 |
| 15 | | | 0.0076731128071754065 | 15 | ОВ | : | 0.00655900434717996 |
| 16 | ол | : | 0.008358607034905775 | 16 | В | : | 0.006244472242541765 |
| 17 | ЛИ | : | 0.00845216083769776 | 17 | Л | : | 0.0035074658090616987 |
| 18 | | | 0.005895590554127679 | 18 | ло | : | 0.00626034312855562 |
| 19 | | | 0.00684643738614077 | 19 | ол | : | 0.006716270399499057 |
| 20 | | | 0.0001888085838165529 | 20 | ЛИ | : | 0.006922591917679157 |
| 21 | | | 0.0009678566143389063 | 21 | ИТ | : | 0.004324095034865451 |
| 22 | би | : | 0.0011566651981554591 | 22 | та | : | |
| 23 | ик | : | | 23 | а | : | 0.015585210065604472 |
| 24 | вр | | | 24 | а | : | 0.0017775392335516301 |
| 25 | pe | | | 25 | б | | 0.006273328398930591 |
| 26 | | | 0.004616454923226348 | 26 | ба | : | 0.0008137436101648696 |
| 27 | да | | | 27 | би | : | 0.0009623528155673193 |
| 28 | | | 0.005883683706499608 | 28 | ИК | | 0.002139683996231386 |
| 29 | | | 0.0008930135721053178 | 29 | | | 0.003882595842116426 |
| 30 | | | 0.014356256282988346 | 30 | | | 0.006049693186917196 |
| 31 | • | | 0.0073652357470781445 | 31 | 200.0 | | 0.0030962655805209403 |
| 32 | | | 0.00043034749284313407 | 32 | | | 0.00377294244783889 |
| 33 | | | 0.005337669693840928 | 33 | | | 0.004096131399393732 |
| 34 | | | 0.008549116596954908 | 34 | | | 0.0005439385479293543 |
| 35 | ая | : | 0.0035669513537235262 | 35 | то | : | 0.011893064855654292 |

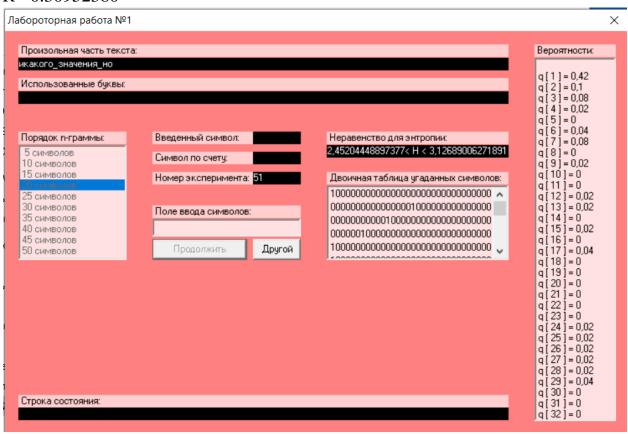
Неперехресні:

| ≣ b | igr.txt | ≣ bigr_s.txt | |
|-----|---------|-----------------------------------|--------------------|
| 1 | ВЛ | | 011804768854461148 |
| 2 | ад | 01991071918094804 2 ад: 0.0 | 032182453418790335 |
| 3 | ИМ | 031510877312456898 3 им: 0.0 | 045824275639651775 |
| 4 | ир | 010243920738023991 4 ир : 0.0 | 0156149536144025 |
| 5 | Н | 132016725004112 5 на : 0.0 | 12168777577062688 |
| 6 | аб | 007560302065809255 6 6o: 0.0 | 025684727622818926 |
| 7 | ок | 023719726328607664 7 ко: 0.0 | 11192417732327717 |
| 8 | ОВ | 063772013608543714 8 ол : 0.0 | 08433435732048756 |
| 9 | Л | 03471390605026158 9 ит: 0.0 | 0599083514487207 |
| 10 | ол | 067061610690613395 10 aa : 0.0 | 00163293632568915 |
| 11 | ИТ | 04250505703411081 11 6a : 0.0 | 010035754501631236 |
| 12 | а | 15590958802124964 12 би: 0.0 | 010648105623764666 |
| 13 | ба | 0008195136590419192 13 вр : 0.0 | 011600651813750004 |
| 14 | би | 0009724510672434042 14 ед: 0.0 | 04565417810572583 |
| 15 | ко | | 058003259068750024 |
| 16 | В | 06342574023148375 16 то: 0.0 | 1412149726653263 |
| 17 | pe | | 004082340814222875 |
| 18 | да | | 0852869035104729 |
| 19 | KT | | 007484291492741938 |
| 20 | ор | | 05031485053529694 |
| 21 | СК | | 06086089763870603 |
| 22 | ая | | 07286978353387832 |
| 23 | 3 | | 019731313935410564 |
| 24 | ам | | 06882146222644064 |
| 25 | ет | | 02265699151893696 |
| 26 | ка | | 015478875587261736 |
| 27 | П | | 07395840775100443 |
| 28 | ри | | 04388516375289591 |
| 29 | ме | | 07099871066069284 |
| 30 | ча | | 007926545080949416 |
| 31 | | | 06613392119041058 |
| 32 | Я | | 04004095948616937 |
| 33 | бо | | 013947997781928157 |
| 34 | | | 06300412656617304 |
| 35 | ар | 002507596372209253 35 ae: 0.0 | 016057207202609977 |

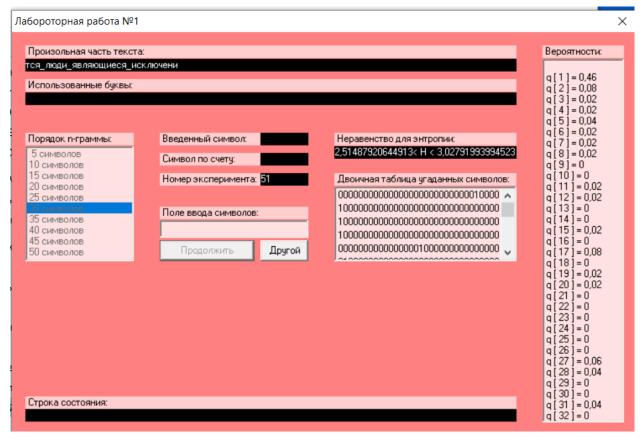
Cool Pink



R = 0.30952380



R = 0,442106544830732



R = 0,445720085360564

Висновки: засвоїв поняття ентропії на символ джерела та його надлишковості, вивчив та порівняв різні моделі джерела відкритого тексту для наближеного визначення ентропії, набув практичних навичок щодо оцінки ентропії на символ джерела. Після аналізу тексту визначив що найчастіші у використанні букви рос. алфавіту «о», «е», «а».