

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ» ФІЗИКО-ТЕХНІЧНИЙ
ІНСТИТУТ

Лабораторна робота №1

Виконали:
студенти ФІ-04
Кравченко Антон
Давидюк Данил

Київ – 2023

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 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.
2. За допомогою програми CoolPinkProgram оцінити значення $H(10)$, $H(20)$, $H(30)$.
3. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела

Хід роботи:

Для роботи було обрано декілька глав з твору “Мертві душі”

Так як роки написання цього твору припадають на 1842, то алфавіт дещо відрізняється від теперішнього поняття алфавіту (входять більше літер).

[Результат роботи програми](#) за посиланням якщо пробіл **не** входить до алфавіту

[Результат роботи програми](#) за посиланням якщо пробіл **входить** до алфавіту

1.1 Кількість та частота літер, якщо вважати, що пробіл не входить до алфавіту:

'r': 2	'r': 4.1968578125557394e-6
'a': 37185	'a': 0.0780300788799426
'o': 53516	'o': 0.11229952134836647
'щ': 1399	'щ': 0.00293570203988274
'o': 2	'o': 4.1968578125557394e-6

'м': 14218	'м': 0.02983546218945875
'і': 73	'і': 0.0001531853101582845
'п': 13525	'п': 0.028381250957408188
'м': 1	'м': 2.0984289062778697e-6
'д': 14922	'д': 0.03131275613947837
'э': 965	'э': 0.0020249838945581443
'ю': 3144	'ю': 0.006597460481337623
'ы': 9142	'ы': 0.019183837061192286
'л': 21292	'л': 0.0446797482724684
'т': 6	'т': 1.2590573437667218e-5
'u': 1	'u': 2.0984289062778697e-6
'а': 1	'а': 2.0984289062778697e-6
'е': 8	'е': 1.6787431250222958e-5
'ц': 1761	'ц': 0.0036953333039553287
'ѣ': 138	'ѣ': 0.000289583189066346
'v': 5	'v': 1.049214453138935e-5
'в': 21803	'в': 0.045752045443576396
'н': 29307	'н': 0.06149865595628553
'й': 4971	'й': 0.01043129009310729
'y': 14621	'y': 0.030681129038688734
'g': 1	'g': 2.0984289062778697e-6
's': 3	's': 6.295286718833609e-6
'г': 8523	'г': 0.017884909568206285
'з': 7996	'з': 0.016779037534597845
'к': 20107	'к': 0.04219311001852913
'ч': 8888	'ч': 0.018650836118997708
'е': 40316	'е': 0.0846002597854986
'б': 9176	'б': 0.019255183644005733
'ъ': 9568	'ъ' 0.020077767775266656
'и': 32585	'и': 0.06837730591106439
'x': 4798	'x': 8.393715625111479e-6
'р': 20000	'р': 0.0419685781255574
'e': 3	'e': 6.295286718833609e-6
'ê': 2	'ê': 4.1968578125557394e-6
'я': 9210	'я': 0.01932653022681918

'x': 4	'x': 0.010068261892321219
'ϕ': 503	'ϕ': 0.0010555097398577684
'ë': 130	'ë': 0.00027279575781612305
'n': 14	'n': 2.9378004687890178e-5
'c': 1	'c': 2.0984289062778697e-6
'ш': 4840	'ш': 0.01015639590638489
'ж': 5172	'ж': 0.010853074303269143
'i': 1	'i': 2.0984289062778697e-6
'p': 2	'p': 4.1968578125557394e-6
'c': 24014	'c': 0.05039167175535676

H1 chars = 4.479795715390369

H2 chars = 2.2398978576951847

Кількість: строка 24-65.

```

1 amount: 816; "x": 2575; "y": 922; "z": 16; "x2": 236; "y2": 321; "z2": 137; "w": 170; "w2": 201; "x3": 555; "y3": 86; "z3": 138; "w3": 170; "w4": 170; "x4": 93; "y4": 28; "z4": 142; "w4": 162; "x5": 2575; "y5": 922; "z5": 16; "x6": 236; "y6": 321; "z6": 137; "w6": 170; "w7": 201; "x7": 555; "y7": 86; "z7": 138; "w7": 170; "w8": 170; "x8": 93; "y8": 28; "z8": 142; "w8": 162; "x9": 2575; "y9": 922; "z9": 16; "x10": 236; "y10": 321; "z10": 137; "w10": 170; "w11": 201; "x11": 555; "y11": 86; "z11": 138; "w11": 170; "w12": 170; "x12": 93; "y12": 28; "z12": 142; "w12": 162; "x13": 2575; "y13": 922; "z13": 16; "x14": 236; "y14": 321; "z14": 137; "w14": 170; "w15": 201; "x15": 555; "y15": 86; "z15": 138; "w15": 170; "w16": 170; "x16": 93; "y16": 28; "z16": 142; "w16": 162; "x17": 2575; "y17": 922; "z17": 16; "x18": 236; "y18": 321; "z18": 137; "w18": 170; "w19": 201; "x19": 555; "y19": 86; "z19": 138; "w19": 170; "w20": 170; "x20": 93; "y20": 28; "z20": 142; "w20": 162; "x21": 2575; "y21": 922; "z21": 16; "x22": 236; "y22": 321; "z22": 137; "w22": 170; "w23": 201; "x23": 555; "y23": 86; "z23": 138; "w23": 170; "w24": 170; "x24": 93; "y24": 28; "z24": 142; "w24": 162; "x25": 2575; "y25": 922; "z25": 16; "x26": 236; "y26": 321; "z26": 137; "w26": 170; "w27": 201; "x27": 555; "y27": 86; "z27": 138; "w27": 170; "w28": 170; "x28": 93; "y28": 28; "z28": 142; "w28": 162; "x29": 2575; "y29": 922; "z29": 16; "x30": 236; "y30": 321; "z30": 137; "w30": 170; "w31": 201; "x31": 555; "y31": 86; "z31": 138; "w31": 170; "w32": 170; "x32": 93; "y32": 28; "z32": 142; "w32": 162; "x33": 2575; "y33": 922; "z33": 16; "x34": 236; "y34": 321; "z34": 137; "w34": 170; "w35": 201; "x35": 555; "y35": 86; "z35": 138; "w35": 170; "w36": 170; "x36": 93; "y36": 28; "z36": 142; "w36": 162; "x37": 2575; "y37": 922; "z37": 16; "x38": 236; "y38": 321; "z38": 137; "w38": 170; "w39": 201; "x39": 555; "y39": 86; "z39": 138; "w39": 170; "w40": 170; "x40": 93; "y40": 28; "z40": 142; "w40": 162; "x41": 2575; "y41": 922; "z41": 16; "x42": 236; "y42": 321; "z42": 137; "w42": 170; "w43": 201; "x43": 555; "y43": 86; "z43": 138; "w43": 170; "w44": 170; "x44": 93; "y44": 28; "z44": 142; "w44": 162; "x45": 2575; "y45": 922; "z45": 16; "x46": 236; "y46": 321; "z46": 137; "w46": 170; "w47": 201; "x47": 555; "y47": 86; "z47": 138; "w47": 170; "w48": 170; "x48": 93; "y48": 28; "z48": 142; "w48": 162; "x49": 2575; "y49": 922; "z49": 16; "x50": 236; "y50": 321; "z50": 137; "w50": 170; "w51": 201; "x51": 555; "y51": 86; "z51": 138; "w51": 170; "w52": 170; "x52": 93; "y52": 28; "z52": 142; "w52": 162; "x53": 2575; "y53": 922; "z53": 16; "x54": 236; "y54": 321; "z54": 137; "w54": 170; "w55": 201; "x55": 555; "y55": 86; "z55": 138; "w55": 170; "w56": 170; "x56": 93; "y56": 28; "z56": 142; "w56": 162; "x57": 2575; "y57": 922; "z57": 16; "x58": 236; "y58": 321; "z58": 137; "w58": 170; "w59": 201; "x59": 555; "y59": 86; "z59": 138; "w59": 170; "w60": 170; "x60": 93; "y60": 28; "z60": 142; "w60": 162; "x61": 2575; "y61": 922; "z61": 16; "x62": 236; "y62": 321; "z62": 137; "w62": 170; "w63": 201; "x63": 555; "y63": 86; "z63": 138; "w63": 170; "w64": 170; "x64": 93; "y64": 28; "z64": 142; "w64": 162; "x65": 2575; "y65": 922; "z65": 16; "x66": 236; "y66": 321; "z66": 137; "w66": 170; "w67": 201; "x67": 555; "y67": 86; "z67": 138; "w67": 170; "w68": 170; "x68": 93; "y68": 28; "z68": 142; "w68": 162; "x69": 2575; "y69": 922; "z69": 16; "x70": 236; "y70": 321; "z70": 137; "w70": 170; "w71": 201; "x71": 555; "y71": 86; "z71": 138; "w71": 170; "w72": 170; "x72": 93; "y72": 28; "z72": 142; "w72": 162; "x73": 2575; "y73": 922; "z73": 16; "x74": 236; "y74": 321; "z74": 137; "w74": 170; "w75": 201; "x75": 555; "y75": 86; "z75": 138; "w75": 170; "w76": 170; "x76": 93; "y76": 28; "z76": 142; "w76": 162; "x77": 2575; "y77": 922; "z77": 16; "x78": 236; "y78": 321; "z78": 137; "w78": 170; "w79": 201; "x79": 555; "y79": 86; "z79": 138; "w79": 170; "w80": 170; "x80": 93; "y80": 28; "z80": 142; "w80": 162; "x81": 2575; "y81": 922; "z81": 16; "x82": 236; "y82": 321; "z82": 137; "w82": 170; "w83": 201; "x83": 555; "y83": 86; "z83": 138; "w83": 170; "w84": 170; "x84": 93; "y84": 28; "z84": 142; "w84": 162; "x85": 2575; "y85": 922; "z85": 16; "x86": 236; "y86": 321; "z86": 137; "w86": 170; "w87": 201; "x87": 555; "y87": 86; "z87": 138; "w87": 170; "w88": 170; "x88": 93; "y88": 28; "z88": 142; "w88": 162; "x89": 2575; "y89": 922; "z89": 16; "x90": 236; "y90": 321; "z90": 137; "w90": 170; "w91": 201; "x91": 555; "y91": 86; "z91": 138; "w91": 170; "w92": 170; "x92": 93; "y92": 28; "z92": 142; "w92": 162; "x93": 2575; "y93": 922; "z93": 16; "x94": 236; "y94": 321; "z94": 137; "w94": 170; "w95": 201; "x95": 555; "y95": 86; "z95": 138; "w95": 170; "w96": 170; "x96": 93; "y96": 28; "z96": 142; "w96": 162; "x97": 2575; "y97": 922; "z97": 16; "x98": 236; "y98": 321; "z98": 137; "w98": 170; "w99": 201; "x99": 555; "y99": 86; "z99": 138; "w99": 170
```

Частота: строка 66-300

[illegible]

H1 bigram1 = 8.287495344039405

H2 bigram1 = 4.143747672019702

[illegible]

301 350444169171777", ("m", "0.0001269286619086"), ("m", "0.0004289054553512"), ("m", "0.000588868030959"), ("lp", "0.198489085766741e-6"), ("p", "0.0001004947225154161708"), ("m", "0.0001972519037120413"), ("c", "3.357492046131766e-5")
302 351 ("m", "0.000251804649548224"), ("m", "0.0004250482323526"), ("m", "0.000494498018713169"), ("m", "0.01021901591745248"), ("m", "0.005199498919144659"), ("p", "0.012548057552741491"), ("c", "0.0019319359889144683"), ("c", "0.0019319359889144683")
303 352 ("m", "0.0001269286619086"), ("m", "0.0004289054553512"), ("m", "0.000588868030959"), ("lp", "0.198489085766741e-6"), ("p", "0.0001004947225154161708"), ("m", "0.0001972519037120413"), ("c", "3.357492046131766e-5")
304 353 37086e-5", ("m", "1.259054710729421e-5"), ("m", "0.00048263765631441"), ("m", "0.0001856588264267188"), ("m", "0.001248101157407017"), ("m", "0.0027927924949144"), ("c", "0.001792918787077"), ("c", "1.968490805766741e-6"), ("m", "0.01680676210799332"), ("m", "0.01405524723031101"), ("m", "0.0023928393388688"), ("x", "1.4168490805766741e-6"), ("p", "0.00036803218806917653"), ("m", "0.0039368011532426e-6"), ("m", "1.968490805766741e-6"), ("s", "6.29523725804970e-5")
305 354 350444169171777", ("m", "0.0001269286619086"), ("m", "0.0004289054553512"), ("m", "0.000588868030959"), ("lp", "0.198489085766741e-6"), ("p", "0.0001004947225154161708"), ("m", "0.0001972519037120413"), ("c", "3.357492046131766e-5")
306 355 350444169171777", ("m", "0.0001269286619086"), ("m", "0.0004289054553512"), ("m", "0.000588868030959"), ("lp", "0.198489085766741e-6"), ("p", "0.0001004947225154161708"), ("m", "0.0001972519037120413"), ("c", "3.357492046131766e-5")
307 356 350444169171777", ("m", "0.0001269286619086"), ("m", "0.0004289054553512"), ("m", "0.000588868030959"), ("lp", "0.198489085766741e-6"), ("p", "0.0001004947225154161708"), ("m", "0.0001972519037120413"), ("c", "3.357492046131766e-5")
308 357 350444169171777", ("m", "0.0001269286619086"), ("m", "0.0004289054553512"), ("m", "0.000588868030959"), ("lp", "0.198489085766741e-6"), ("p", "0.0001004947225154161708"), ("m", "0.0001972519037120413"), ("c", "3.357492046131766e-5")
309 358 350444169171777", ("m", "0.0001269286619086"), ("m", "0.0004289054553512"), ("m", "0.000588868030959"), ("lp", "0.198489085766741e-6"), ("p", "0.0001004947225154161708"), ("m", "0.0001972519037120413"), ("c", "3.357492046131766e-5")
310 359 350444169171777", ("m", "0.0001269286619086"), ("m", "0.0004289054553512"), ("m", "0.000588868030959"), ("lp", "0.198489085766741e-6"), ("p", "0.0001004947225154161708"), ("m", "0.0001972519037120413"), ("c", "3.357492046131766e-5")
311 360 350444169171777", ("m", "0.0001269286619086"), ("m", "0.0004289054553512"), ("m", "0.000588868030959"), ("lp", "0.198489085766741e-6"), ("p", "0.0001004947225154161708"), ("m", "0.0001972519037120413"), ("c", "3.357492046131766e-5")
312 361 350444169171777", ("m", "0.0001269286619086"), ("m", "0.0004289054553512"), ("m", "0.000588868030959"), ("lp", "0.198489085766741e-6"), ("p", "0.0001004947225154161708"), ("m", "0.0001972519037120413"), ("c", "3.357492046131766e-5")
313 362 350444169171777", ("m", "0.0001269286619086"), ("m", "0.0004289054553512"), ("m", "0.000588868030959"), ("lp", "0.198489085766741e-6"), ("p", "0.0001004947225154161708"), ("m", "0.0001972519037120413"), ("c", "3.357492046131766e-5")
314 363 350444169171777", ("m", "0.0001269286619086"), ("m", "0.0004289054553512"), ("m", "0.000588868030959"), ("lp", "0.198489085766741e-6"), ("p", "0.0001004947225154161708"), ("m", "0.0001972519037120413"), ("c", "3.357492046131766e-5")
315 364 350444169171777", ("m", "0.0001269286619086"), ("m", "0.0004289054553512"), ("m", "0.000588868030959"), ("lp", "0.198489085766741e-6"), ("p", "0.0001004947225154161708"), ("m", "0.0001972519037120413"), ("c", "3.357492046131766e-5")
316 365 350444169171777", ("m", "0.0001269286619086"), ("m", "0.0004289054553512"), ("m", "0.000588868030959"), ("lp", "0.198489085766741e-6"), ("p", "0.0001004947225154161708"), ("m", "0.0001972519037120413"), ("c", "3.357492046131766e-5")
317 366 350444169171777", ("m", "0.0001269286619086"), ("m", "0.0004289054553512"), ("m", "0.000588868030959"), ("lp", "0.198489085766741e-6"), ("p", "0.0001004947225154161708"), ("m", "0.0001972519037120413"), ("c", "3.357492046131766e-5")
318 367 350444169171777", ("m", "0.0001269286619086"), ("m", "0.0004289054553512"), ("m", "0.000588868030959"), ("lp", "0.198489085766741e-6"), ("p", "0.0001004947225154161708"), ("m", "0.0001972519037120413"), ("c", "3.357492046131766e-5")
319 368 350444169171777", ("m", "0.0001269286619086"), ("m", "0.0004289054553512"), ("m", "0.000588868030959"), ("lp", "0.198489085766741e-6"), ("p", "0.0001004947225154161708"), ("m", "0.0001972519037120413"), ("c", "3.357492046131766e-5")
320 369 350444169171777", ("m", "0.0001269286619086"), ("m", "0.0004289054553512"), ("m", "0.000588868030959"), ("lp", "0.198489085766741e-6"),

H2 bigram2 = 7.289178349575229

H1 chars = 4.385256604466399
H2 chars = 2.1926283022331994


```
288 amount: 885, {"сн": 485, "кк": 5, "эд": 567, "юо": 18, "хп": 2, "мц": 7, "пи": 1,  
289 "хр": 76, "с": 8681, "жу": 116, "ял": 26, "тк": 351, "йк": 155, "рз": 8, "фл":  
290 ": 28, "ту": 1, "фм": 2, "ра": 3544, "от": 3673, "жн": 467, "ыц": 5, "я": 747,
```

[illegible]

H2 bigram2 = 6.949868202106845

2. H(10)

[illegible]

R = 0,40328551961599

H(20)

Лабораторная работа №1

Произвольная часть текста:
мы_обычно_подразуме

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: 50

Поле ввода символов:

Продолжить

Другой

Неравенство для энтропии:
3.71836713379766 < H < 3.98633331469029

Двоичная таблица угаданных символов:
00000010000000000000000000000000
00000100000000000000000000000000
0000000000000000000001000000000000
00000000000000000000000000000000
00000000000001000000000000000000
00000000000001000000000000000000

Вероятности:
q[1] = 0,1428571
q[2] = 0,0816326
q[3] = 0,0408163
q[4] = 0,1020408
q[5] = 0,0204081
q[6] = 0,1224489
q[7] = 0,0612244
q[8] = 0
q[9] = 0,0204081
q[10] = 0,020408
q[11] = 0,020408
q[12] = 0,061224
q[13] = 0,020408
q[14] = 0,040816
q[15] = 0,061224
q[16] = 0
q[17] = 0,020408
q[18] = 0,020408
q[19] = 0
q[20] = 0,061224
q[21] = 0
q[22] = 0
q[23] = 0
q[24] = 0
q[25] = 0,040816
q[26] = 0,020408
q[27] = 0
q[28] = 0,020408
q[29] = 0
q[30] = 0
q[31] = 0
q[32] = 0

Строка состояния:

R = 0,26796618089263

H(30)

The screenshot shows a software application titled "Лабораторная работа №1" (Laboratory Work No. 1). The interface is divided into several sections:

- Top Section:** Contains a text input field labeled "Произвольная часть текста:" (Arbitrary part of the text) with the value "ество_по_имени_человек_имеет_". Below it is a label "Использованные буквы:" (Used letters).
- Left Panel:** A list titled "Порядок n-граммы:" (Order of n-grams) with options from 5 to 50 symbols. The option "30 символов" (30 symbols) is selected.
- Center Section:** Includes a "Введенный символ:" (Entered symbol) field, a "Символ по счету:" (Symbol by count) field, and a "Номер эксперимента:" (Experiment number) field with the value "50". Below these is a "Поле ввода символов:" (Symbol input field) and two buttons: "Продолжить" (Continue) and "Другой" (Other).
- Right Panel:** Displays "Вероятности:" (Probabilities) as a list of values $q[1]$ through $q[32]$. Below this is a section titled "Неравенство для энтропии:" (Entropy inequality) showing the calculated value $4,08206365667313 < H < 4,12504212044804$. At the bottom of the right panel is a "Двоичная таблица угаданных символов:" (Binary table of guessed symbols) with a grid of 0s and 1s and a vertical scrollbar.
- Bottom Section:** A "Строка состояния:" (Status bar) at the very bottom.

R = 0,04297846377491

Висновок:

Виявилось, що у творі “Мертві душі” : найбільш популярний символ це розділовий знак, воно в принципі зрозуміло чому. Другий за популярністю йде символ “о”, вражає що за 181 рік популярність символу в мові не змінилася.