

Міністерство освіти і науки України

Національний технічний університет України

"Київський політехнічний інститут імені Ігоря Сікорського"

Фізико-технічний інститут

КРИПТОГРАФІЯ

КОМП’ЮТЕРНИЙ ПРАКТИКУМ No1

Виконала студентка гр. ФБ-24:  
Тішевська Анна

Київ–2025

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*1та *H*2за безпосереднім означенням. Підрахувати частоти букв та біграм, а

також значення*H*1та *H*2на довільно обраному тексті російською мовою достатньої  довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також  одержати значення *H*1та *H*2на 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.

**1. Теоретичні відомості**

**Ентропія символів (монограм):**

**Ентропія біграм:**

**Надлишковість:**

де — кількість символів у алфавіті.

**Умовна ентропія H⁽ⁿ⁾:** показує середню невизначеність наступного символу при врахуванні n попередніх символів.

**2. Вихідні дані**

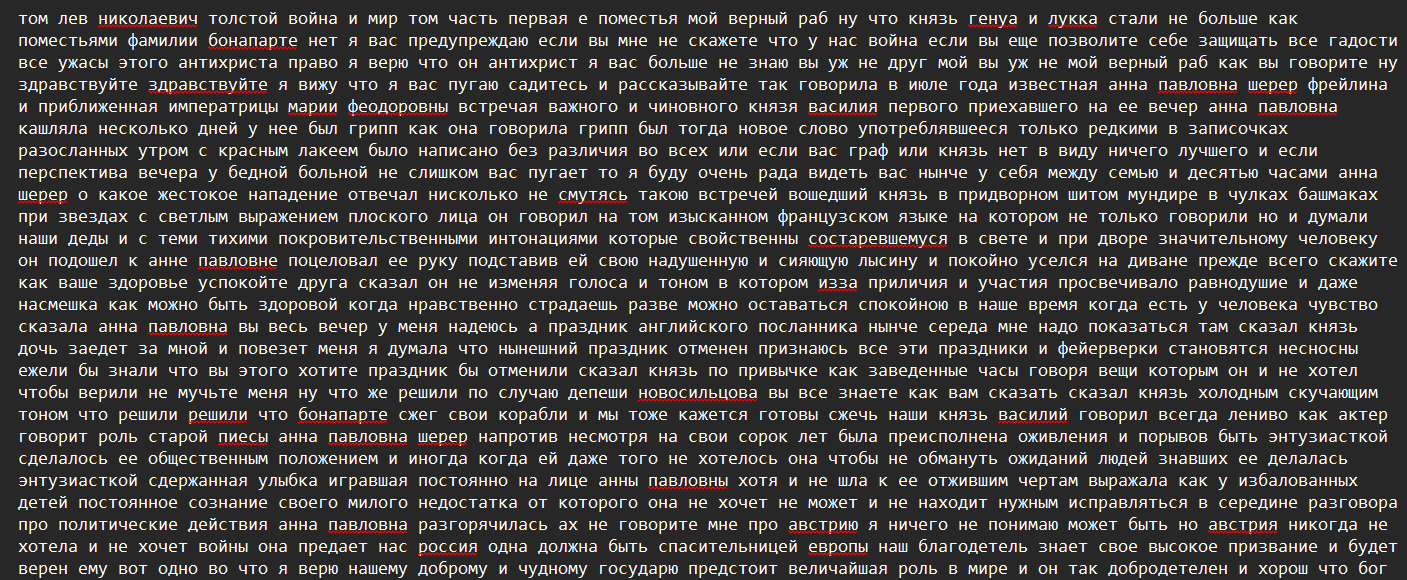
Вхідний текст: фрагмент російського корпусу (≈ 1 МБ після нормалізації).

Алфавіт:

з пробілами: 32 символи (абвгдежзийклмнопрстуфхцчшщыьэюя );

без пробілів: 31 символ.





**3. Підготовка тексту**

Проведено нормалізацію:

ё → е, ъ → ь;

перетворено у нижній регістр;

усі небуквені символи замінені на пробіл;

послідовності пробілів стиснуті до одного;

створено варіант тексту без пробілів.

**4. Використані засоби**

Для основних розрахунків використано Python-скрипт, який:

обчислює частоти символів та біграм;

знаходить , і надлишковість ;

формує CSV-файли:

таблиці частот символів;

матриці частот біграм (для step=1, step=2);

для текстів зі пробілами і без.

**5. Результати розрахунку**

| **Варіант тексту** |  | **Надлишковість** |
| --- | --- | --- |
| З пробілами | 4.38 | 0.122 |
| Без пробілів | 4.46 | 0.098 |



З пробілами:

|  |  |  |
| --- | --- | --- |
| Символ | Кількість | Частота |
|  | 103318 | 0.160949013132273 |
| о | 61231 | 0.09538578972785194 |
| а | 45176 | 0.07037527456264701 |
| е | 42903 | 0.066834390042528 |
| и | 35789 | 0.05575218481765925 |
| н | 35097 | 0.054674185658872464 |
| т | 30590 | 0.04765317090648513 |
| с | 28102 | 0.04377735890206097 |
| л | 27261 | 0.04246724720763946 |
| в | 24794 | 0.03862414905051953 |
| р | 24545 | 0.0382362562896266 |
| к | 19315 | 0.030088950508622436 |
| д | 16380 | 0.025516800897294096 |
| м | 15921 | 0.02480176966335893 |
| у | 15443 | 0.02405714018662471 |
| п | 13838 | 0.021556867571230508 |
| я | 12469 | 0.019424236287445672 |
| г | 11168 | 0.017397535556836416 |
| ь | 10774 | 0.016783761469319085 |
| ы | 10223 | 0.015925412428146372 |
| з | 9592 | 0.014942439206767093 |
| б | 9302 | 0.014490676553518296 |
| ч | 7340 | 0.0114342685339523 |
| й | 6203 | 0.009663047372766501 |
| ж | 5456 | 0.008499369090087704 |
| ш | 5087 | 0.007924540058884925 |
| х | 4595 | 0.007158101350614553 |
| ю | 3494 | 0.005442961070521708 |
| ц | 2179 | 0.003394451108376303 |
| э | 1628 | 0.002536102067203589 |
| щ | 1511 | 0.0023538392036514884 |
| ф | 1206 | 0.001878709516613961 |

Без пробілів:

|  |  |  |  |
| --- | --- | --- | --- |
| Символ | Кількість | Частота |  |
| о | 61231 | 0.11368294802195272 | |
| а | 45176 | 0.08387484868513884 | |
| е | 42903 | 0.07965474218918256 | |
| и | 35789 | 0.06644671860263046 | |
| н | 35097 | 0.06516193475080392 | |
| т | 30590 | 0.05679413009736137 | |
| с | 28102 | 0.05217484942778847 | |
| л | 27261 | 0.050613428590525276 | |
| в | 24794 | 0.04603313702628237 | |
| р | 24545 | 0.04557083763451241 | |
| к | 19315 | 0.03586069378327999 | |
| д | 16380 | 0.03041150215739716 | |
| м | 15921 | 0.02955931171232724 | |
| у | 15443 | 0.028671845410053992 | |
| п | 13838 | 0.025691963788404267 | |
| я | 12469 | 0.023150245445701174 | |
| г | 11168 | 0.020734777539304732 | |
| ь | 10774 | 0.020003267658351465 | |
| ы | 10223 | 0.018980267799454897 | |
| з | 9592 | 0.01780873801549167 | |
| б | 9302 | 0.017270317037125053 | |
| ч | 7340 | 0.013627620624865396 | |
| й | 6203 | 0.011516639064855592 | |
| ж | 5456 | 0.010129740889545722 | |
| ш | 5087 | 0.009444646610175786 | |
| х | 4595 | 0.00853118757101587 | |
| ю | 3494 | 0.006487044477286061 | |
| ц | 2179 | 0.004045583834002955 | |
| э | 1628 | 0.0030225839751063847 | |
| щ | 1511 | 0.002805358959696405 | |
| ф | 1206 | 0.002239088620379791 | |

**6. Результати для біграм**

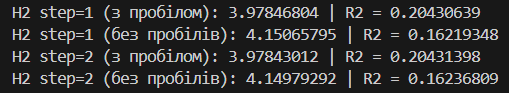
**Методика:**

Біграми формувалися у двох режимах:

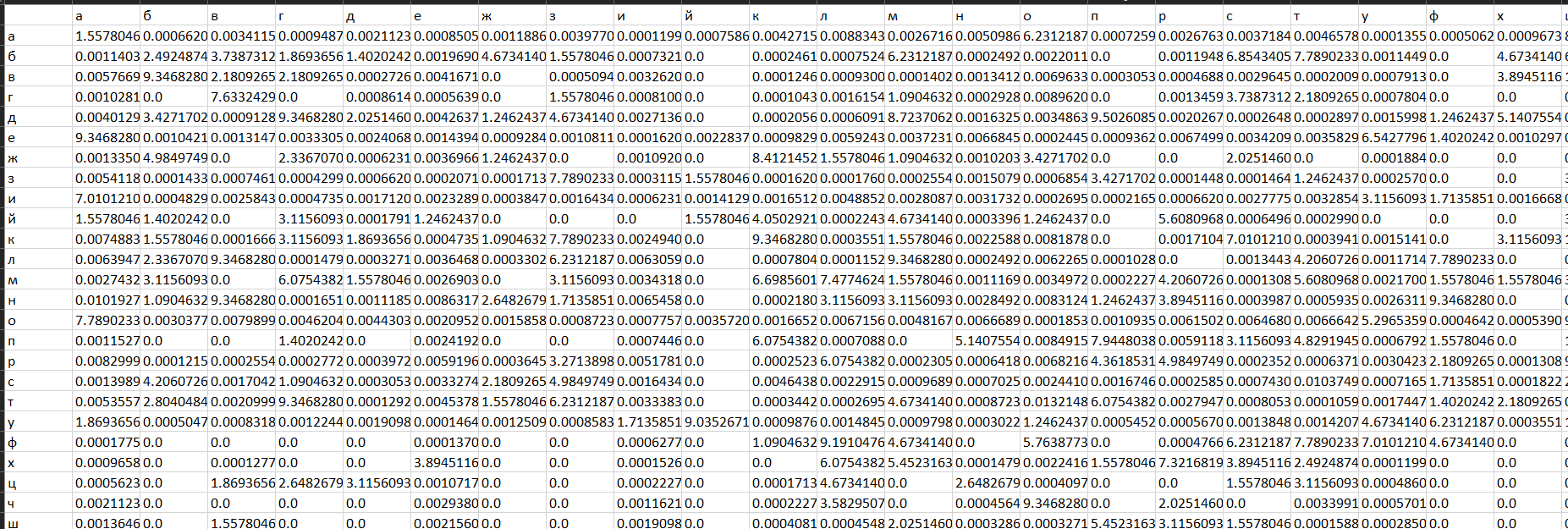
step=1 — *перетинні біграми* (мають спільні символи);

step=2 — *неперетинні біграми* (беруться через два символи).

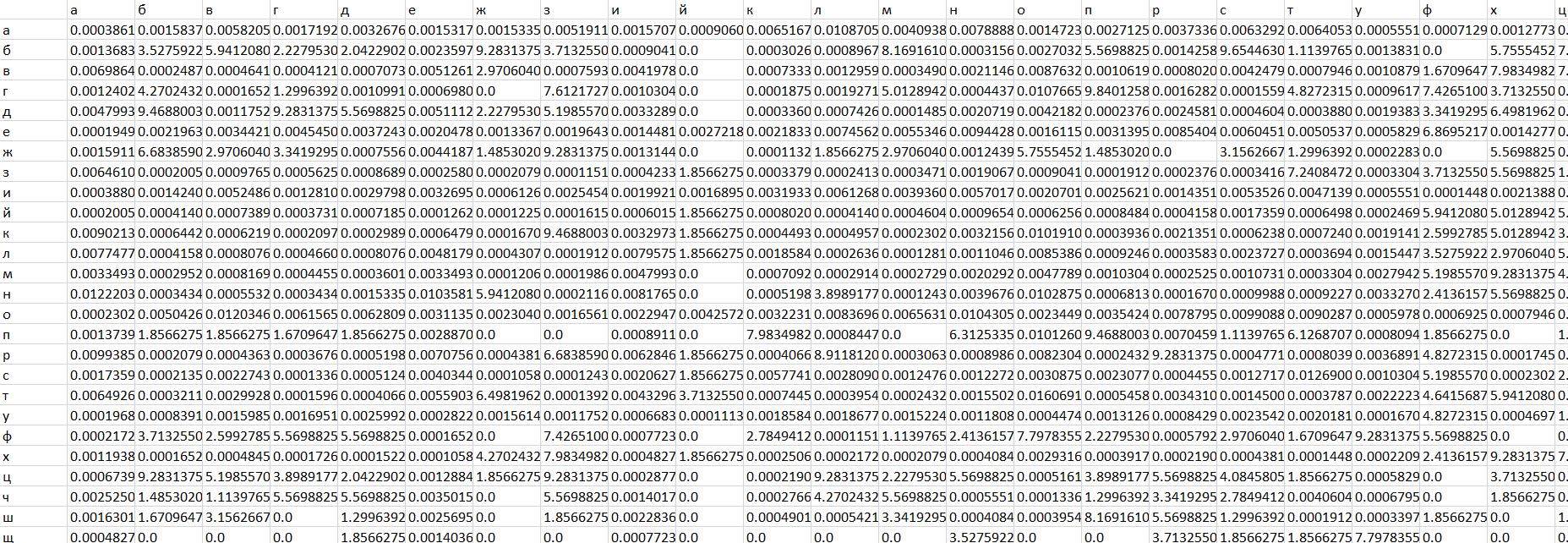
| **Варіант** | **step** |  |  |
| --- | --- | --- | --- |
| З пробілами | 1 | 3.97 | 0.204 |
| Без пробілів | 1 | 4.15 | 0.162 |
| З пробілами | 2 | 3.97 | 0.204 |
| Без пробілів | 2 | 4.14 | 0.162 |



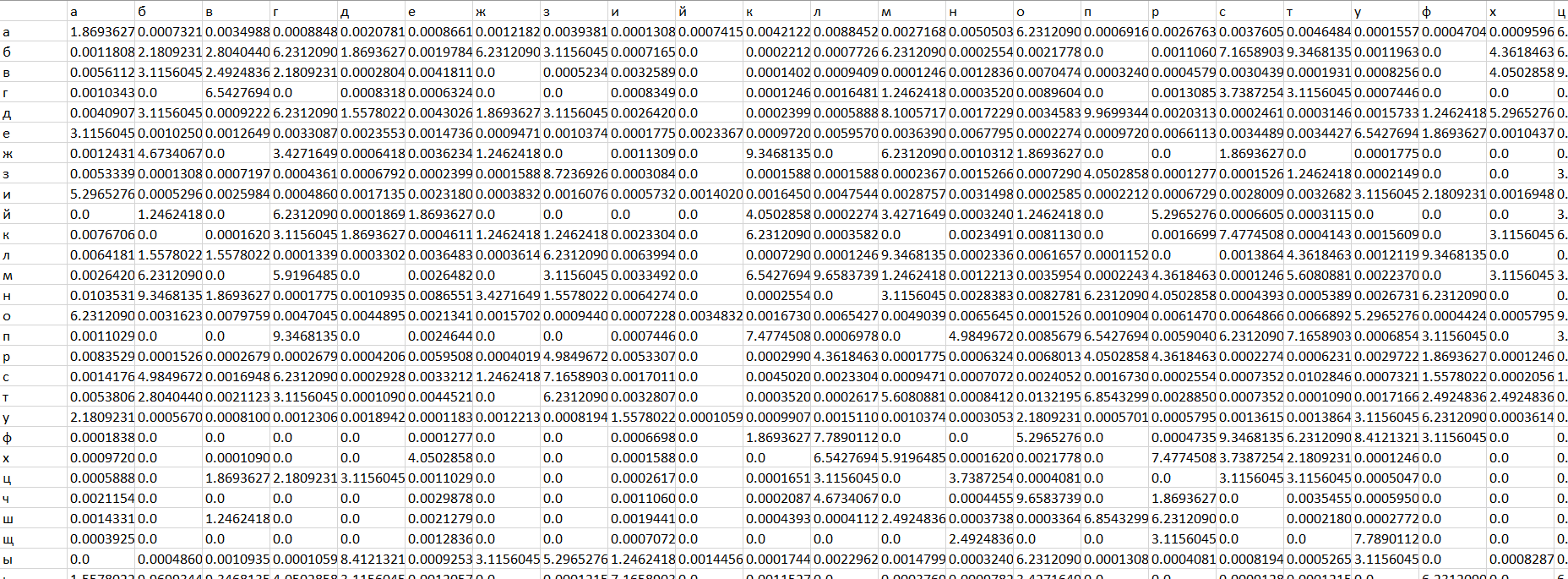
перетинні біграми з пробілами:



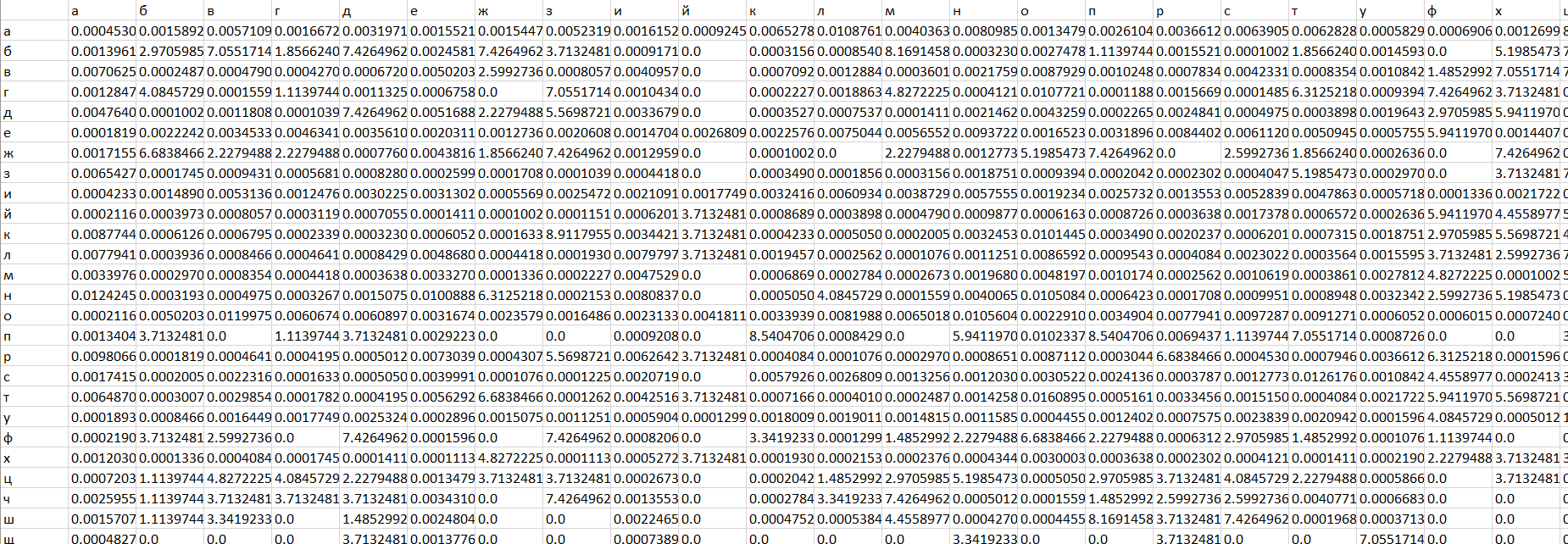
перетинні біграми без пробілів:



неперетинні біграми з пробілами:



неперетинні біграми без пробілів:



**7. Розрахунок у CoolPinkProgram**

| **Варіант** | **n** |  |  |  |
| --- | --- | --- | --- | --- |
| з пробілами | 10 | 3.156 | 3.691 | 3.424 |
| з пробілами | 20 | 1.499 | 2.099 | 1.799 |
| з пробілами | 30 | 1.735 | 2.526 | 2.131 |

**8. Порівняння результатів**

**Спостереження:**

, бо при обліку зв’язків між символами невизначеність зменшується.

При переході до step=2 (неперетинні біграми) значення H₂ дещо збільшується.

При збільшенні порядку n у CoolPinkProgram зменшується (програма “вчиться” краще передбачати наступний символ).

Вилучення пробілів робить розподіл рівномірнішим → H трохи зростає, надлишковість зменшується.

**9. Висновки**

1. Проведено розрахунок ентропій , та на тексті російською мовою.
2. Отримані результати узгоджуються з теорією Шеннона: зі збільшенням порядку n умовна ентропія прямує до деякої межі (~1–2 біт/символ).
3. Надлишковість природної мови становить близько 40–60 %, що свідчить про значну структурованість текстів.
4. Алгоритм у Python та експеримент у CoolPinkProgram підтвердили взаємну залежність символів у природних мовах.