Національний технічний університет України
«Київський політехнічний інститут»
Фізико-технічний інститут

Криптографія

Комп’ютерний практикум №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

Виконали:
студенти групи ФБ-32
Грабовецький Микита
Драбок Алла

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:

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.

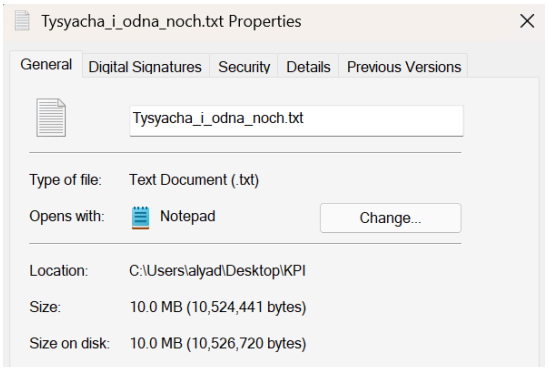
2. За допомогою програми CoolPinkProgram оцінити значення $H^{(10)}$, $H^{(20)}$, $H^{(30)}$.

3. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

Варіант: 13

Хід роботи:

Ми обрали книгу, яку будемо аналізувати в нашій програмі, а саме: «Тисяча і одна ніч». Як бачимо, вона важить 10 мб, але будемо використовувати лише перші 2 мегабайти (прописано в коді). Кількість літер після фільтрації складає 1975786.



Наша програма заміняє всі непотрібні символи на нижні підкреслення (таким чином ми позначити пробіли), а всі подвійні нижні підкреслення (тобто пробіли) видаляє. Замінює «ё» на «е», «ъ» на «ь» і «:» на «_» (знову ж, таким чином позначили пробіл).

Підраховує частоти літер, біграм, що перетинаються та не перетинаються. Топ-30 літер та топ-10 біграм будуть показані нижче, однак у нас також створюється таблиця Excel, де зберігаються абсолютно всі можливі результати (додана у репозиторій). Також обчислюємо ентропії H_1 та H_2 на тексті з та без пробілів.

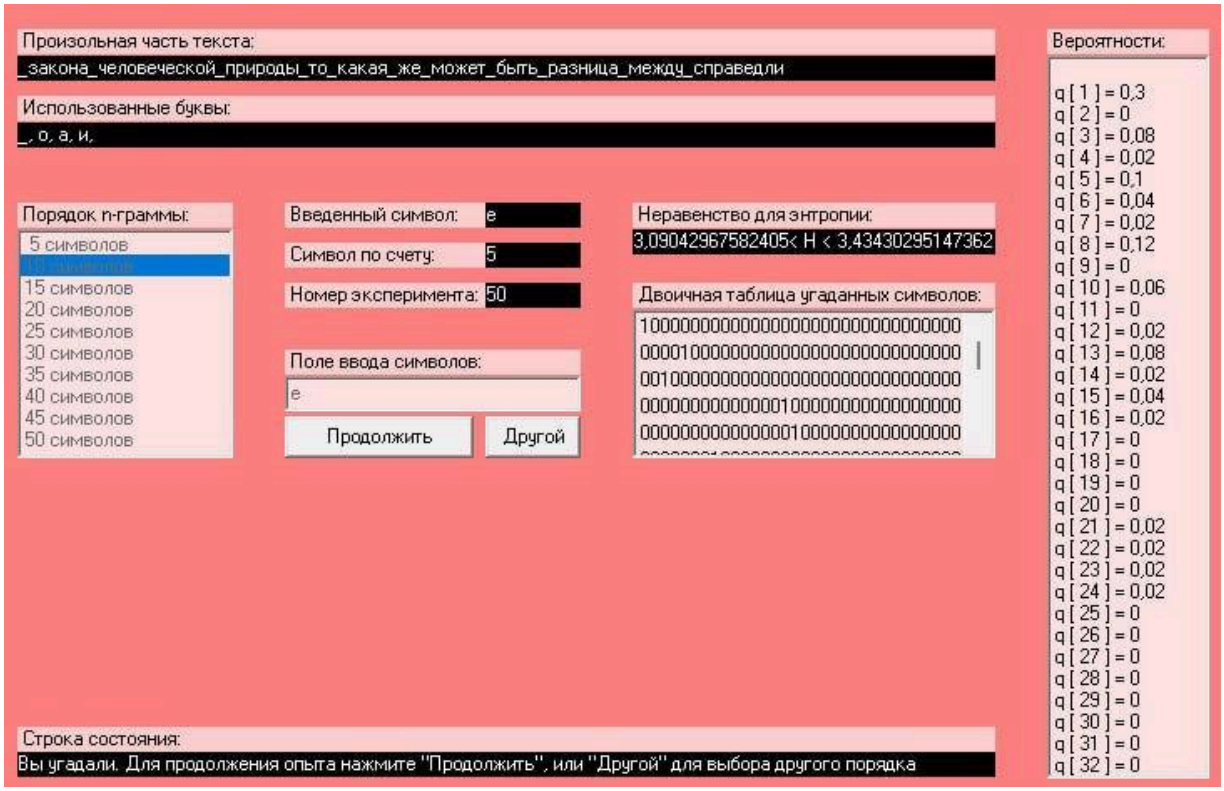
```
Топ-30 літер:
'_': 0.18351
'о': 0.08441
'а': 0.07528
'и': 0.06751
'е': 0.06434
'н': 0.04734
'т': 0.04695
'л': 0.04586
'с': 0.04331
'в': 0.03350
'р': 0.03288
'д': 0.02742
'к': 0.02636
'м': 0.02566
'у': 0.02451
'п': 0.02113
'я': 0.01799
'ь': 0.01661
'з': 0.01538
'б': 0.01459
'г': 0.01421
'ы': 0.01332
'ч': 0.01061
'й': 0.00854
'х': 0.00850
'ш': 0.00823
'ж': 0.00669
'ю': 0.00545
'ц': 0.00474
'э': 0.00241
```

```
Топ-10 біграм (перетин):
'и_': 0.03079
'_и': 0.02214
'а_': 0.02111
'о_': 0.01829
'_с': 0.01704
'_п': 0.01563
'е_': 0.01451
'_н': 0.01433
'_о': 0.01366
'_в': 0.01335

Топ-10 біграм (без перетину):
'и_': 0.03080
'_и': 0.02218
'а_': 0.02102
'о_': 0.01824
'_с': 0.01711
'_п': 0.01560
'е_': 0.01453
'_н': 0.01426
'_о': 0.01379
'_в': 0.01325
```

```
Оцінка ентропії та надлишковості
Максимальна ентропія H_0 = log2(32) = 5.0000 біт/символ
-----------------------------------------------
Модель 1-грам:
  Ентропія H_1 ≈ 4.3155 біт/символ
  Надлишковість R_1 ≈ 13.7%
Модель 2-грам:
  Ентропія H_2 ≈ 3.8284 біт/символ
  Надлишковість R_2 ≈ 23.4%

Аналіз без пробілів ('_'):
Модель 1-грам (без пробілів):
  Ентропія H_1 ≈ 4.4400 біт/символ
  Надлишковість R_1 ≈ 11.2%
Модель 2-грам (без пробілів):
  Ентропія H_2 ≈ 4.0536 біт/символ
  Надлишковість R_2 ≈ 18.9%
```

Тепер переходимо до програми CoolPinkProgram (повний аналіз всіх результатів нижче).

Результати для $H^{(10)}$



Результати для $H^{(20)}$

Результати для $H^{(30)}$

Аналіз результатів:

Спочатку проаналізуємо наші значення **ентропій**:

- H0 = 5 біт/символ - це максимальна ентропія, яка відповідає ситуації, коли всі 32 символи алфавіту з'являються з однаковою ймовірністю. Однак це дуже ідеалізована ситуація, а мова має свої особливості.
- H1 ≈ 4.32 - значення вже менше, оскільки частота появи літер нерівномірна (наприклад, літери 'о', 'а', 'е' зустрічаються набагато частіше, ніж 'ф', 'щ' або 'э'), тож враховуючи це можна трошки краще передбачити наступний символ.
- H2 ≈ 3.83 - це значення ще нижче, тому що модель біграм враховує залежність між сусідніми літерами (наприклад, після приголосної літери ймовірність появи голосної набагато вища, ніж іншої приголосної), а також існують стійкі буквосполучення та майже неможливі. Тому знання попередньої літери суттєво звужує коло можливих наступних літер, що ще більше знижує ентропію.

Маємо: H0>H1>H2.

Тепер проаналізуємо H(10), H(20) та H(30) з програми **CoolPinkProgram**. По ідеї, зі збільшенням кількості символів у послідовності діапазон значень ентропії мав би зміщуватись вниз, тобто мало б бути H(10)>H(20)>H(30). Однак фактично вийшли такі результати:

- 3.09<H(10)<3.43
- 2.98<H(20)<3.5
- 3.05<H(30)<3.69

Це є аномалією, адже фактично виходить, що у нас H(30)>H(10)>H(20). Найімовірніше, що це похибка, спричинена специфікою тексту, а також кількістю наших експериментів. Для більш точних значень краще було б взяти більшу кількість тестів.

Тож підсумуючи, маємо: H0>H1>H2>H(30)>H(10)>H(20).

Отже, чим більший порядок моделі (чим довший ланцюжок попередніх символів ми враховуємо), тим точніше ми можемо передбачити наступний символ. Це відбувається тому, що моделі вищого порядку враховують не тільки частоти літер і біграм, але й складніші мовні патерни, синтаксис та семантику. Тому з ростом n ентропія Hn зменшується.

Тепер порівняємо наші результати в розрізі **тексту з та без проблів**.

- H1 з пробілами ≈ 4.32
- H1 без пробілів ≈ 4.44

- H2 з пробілами ≈ 3.83
- H2 без пробілів ≈ 4.05

Як бачимо, значення ентропій у тексті без пробілів більше. Але цьому є логічне пояснення. У таблиці із частотами букв чітко видно, що “_” (тобто пробіл) є найчастішим. І коли ми його видаляємо, то виходить, що розподіл частот літер також змінюється, а від нього залежить ентропія. Тобто видалення одного дуже частого символу зробило розподіл ймовірностей решти літер трохи більш рівномірним, що і призвело до збільшення ентропії. Тож якщо казати простими словами, то коли зникає найбільш передбачуваний елемент, система в цілому стає менш передбачуваною.

Тепер перейдемо до аналізу **надлишковості**.

Надлишковість (R) — це показник, який демонструє, наскільки мова є "стискаємою" або передбачуваною. Шукаємо за формулою: 1 - (Hn/H0). З наших результатів бачимо, що чим менша ентропія, тим більша надлишковість. Що є логічним, адже ми ми виявляємо все більше закономірностей у мові. Тож конкретні цифри:

- При $R_1 \approx 13.7\%$ ми враховуємо лише частоту літер.
- При $R_2 \approx 23.4\%$ ми вже враховуємо зв'язки між парами літер, що розкриває більшу частину структури мови.

При більших n надлишковість буде ще більшою, оскільки модель враховуватиме слова, фрази та граматичні конструкції, роблячи текст ще більш передбачуваним.

За **даними з інтернету** надлишковість української мови оцінюється у приблизно 70%. Результати нашої програми показали 13.7% та 23.4% для літер та біграм відповідно для тексту із пробілами й 11.2% та 18.9% для тексту без пробілів. CoolPinkProgram безпосередньо не показує значення надлишковості, однак ми можемо його порахувати самостійно за нашою формулою, взявши H(10), яке є середнім для нашого випадку. Тож:

H(10) середнє = (3.09+3.43)/2 = 3.26 біт/секунду

R = 1 - (3.26/5) = 1 - 0.652 = 0.348 = 34.8%

Як бачимо, це значення вже ближче, але все ще далеко від реального.

Значення з нашої програми на пайтоні найнижчі, адже наші моделі дуже прості. Вони не враховують існування слів, граматики, синтаксису, сенсу речень, що є основними

джерелами надлишковості в мові. Значення з програми CoolPinkProgram вже краще, адже програма використовує моделі значно вищого порядку, які здатні вловити більш складні статистичні закономірності тексту. Але цього все ще недостатньо для розкриття повної надлишковості мови. Отже, чим складніша модель (чим більше контексту вона враховує), тим повніше вона розкриває справжню надлишковість мови.

Висновки:

У межах лабораторної роботи було проведено аналіз тексту шляхом розрахунку частот символів та біграм, ентропій та надлишковості. Ми виявили, що зі збільшенням порядку моделі (від H0 до H2 і вище), ентропія джерела послідовно зменшується. А значне зниження ентропії від H1≈4.32 до H2≈3.83 біт/символ довело наявність зв'язків між сусідніми літерами. Надлишковість мови, що є показником її передбачуваності, відповідно зросла (від R1≈13.7% до R2≈23.4%), підтверджуючи виявлення мовних патернів. Видалення найбільш частого символу (пробілу) збільшило H1 та H2, що свідчить про те, що його присутність робила текст більш передбачуваним. Експериментальні результати з високопорядковими n-грамами (до n=30) дали оцінку надлишковості близько 35%, що є ближчим до справжніх лінгвістичних показників.