НТУУ «Київський політехнічний інститут ім. Ігоря Сікорського»

Навчально-науковий Фізико-технічний інститут

**Криптографія**

Комп’ютерний практикум №1

*Експериментальна оцінка ентропії на символ джерела відкритого тексту*

Варіант №6

Виконали:

Студенти 3 курсу НН ФТІ

групи ФБ-31

Гаврилюк Володимир

Гек Роман

**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 H1 та H2 за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення H1 та H2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення H1 та H2 на тому ж тексті, в якому вилучено всі пробіли.

2. За допомогою програми CoolPinkProgram оцінити значення H10, H20, H30.

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

**Виконання роботи**

Було проаналізовано текст розміром >1MB, текст був попередньо відфільтрований як і було вказано у методичних вказівках

Був написаний скрипт мовою Python, який обчислює:

- частоти появи кожного символу в тексті;

- ентропію H1 на основі частот символів;

- частоти появи біграм двома методами: з перетином (крок 1 символ) та - без перетину (крок 2 символи)

- ентропію H2 для обох методів підрахунку біграм

- усі розрахунки були проведені для двох варіантів тексту: з пробілами та з видаленими пробілами

Результати програмного аналізу:  
Таблиця частот символів (з пробілами)

Зображення, що містить текст, знімок екрана, меню, дизайн

Вміст на основі ШІ може бути неправильним.

Таблиця частот символів (без пробілів)

Зображення, що містить текст, знімок екрана, меню, дизайн

Вміст на основі ШІ може бути неправильним.

Підсумкові значення ентропій:  
Зображення, що містить текст, Шрифт, знімок екрана

Вміст на основі ШІ може бути неправильним.

Надлишковість:  
Зображення, що містить текст, Шрифт, знімок екрана

Вміст на основі ШІ може бути неправильним.

Далі було проведено експерименти з CoolPinkProgram.exe, для кожної n-грами по 50 експериментів. Під час експериментів навмисно підроблювати результати не намагалися, як виходило так і вгадували символи.

Експерименти із CoolPinkProgram

H10

Зображення, що містить текст, електроніка, знімок екрана, програмне забезпечення

Вміст на основі ШІ може бути неправильним.

H20

Зображення, що містить текст, електроніка, знімок екрана, програмне забезпечення

Вміст на основі ШІ може бути неправильним.

H30

Зображення, що містить текст, знімок екрана, Веб-сайт, Веб-сторінка

Вміст на основі ШІ може бути неправильним.

Щодо того, наскільки точні результати експериментів

Було знайдено статті <https://habr.com/ru/articles/791130/>, <https://habr.com/ru/articles/50643/> та дослідження <https://cyberleninka.ru/article/n/analiz-izbytochnosti-russkoyazychnogo-teksta/viewer>

Зображення, що містить текст, знімок екрана, Шрифт

Вміст на основі ШІ може бути неправильним.

Зображення, що містить текст, знімок екрана, Шрифт, число

Вміст на основі ШІ може бути неправильним.

Російська мова має 72.6 відсотків надлишковості, досить великий результат, але ще більше у англійської мови

В загальному надлишковість не є чимось поганим, наприклад як було сказано в одній із статтей – надлишковість допомагає диспетчерам літаків передавати усю потрібну інформацію попри загрозу втрати деякої її частини.

Також щодо ентропій і частот символів

Зображення, що містить текст, чорно-білий, візерунок, шов

Вміст на основі ШІ може бути неправильним.Зображення, що містить текст, Шрифт, ряд, білий

Вміст на основі ШІ може бути неправильним.

Хотілось би прокоментувати, що наші результати більш менш збігаються із результатами іншого дослідження, і це також показує приблизну однозначність ентропії і надлишковості для різних текстів.

**Підсумок та висновки**

В ході виконання роботи було експериментально досліджено статистичні властивості російської мови. Також ми провели оцінку ентропії та надлишковості за допомогою різних моделей. Отримані результати дозволяють зробити декілька висновків

1. Підтверджено теоретичну залежність ентропій від моделі джерела. Чітко видно ієрархію H0 > H1 > H2 > Hn. Врахування зв’язків між сусідніми символами (модель біграм, або n-грам) значно зменшує невизначеність у порівнянні з моделлю, що враховує тільки частоти окремо одного символу.
2. Доведено вплив довжини контексту на точність оцінки ентропій. Оцінка умовної ентропії в експериментах показала, що зі збільшенням кількості символів (n) ентропія значно зменшується. Так середнє значення H10 складає близько ~2.35 біт, тоді як H30 складає ~1.81 біт. Це свідчить, що російська мова має залежності, які дозволяють людині ефективно передбачити наступні символи.
3. Встановлено значну надлишковість російської мови. Найбільш точні оцінки, які були отримані, показують, що надлишковість російської мови може сягати 60-70%. Це означає, що більше половини символів у тексті є передбачуваними і визначаються структурними закономірностями мови. Простіші статистичні моделі (H1, H2) недооцінюють реальну надлишковість, показуючи значення в діапазоні 10-20%.
4. Також проаналізовано вплив алфавіту на ентропію. Видалення пробілів з тексту призвело до незначного зростання ентропії (H1 з та без – 4.43 біт, 4.49 біт). Це пояснюється тим, що пробіл є дуже частим символом, структурно передбачуваним. Його видалення усуває частину передбачуваності з тексту, роблячи його в середньому трішки більш хаотичним для аналізу.