ПРИКЛАДНА КРИПТОЛОГІЯ 2
КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1

Розрахунок частот символів та біграм відкритих текстів

Порядок виконання роботи

Завдання 1. *Написати програму, яка проводить первинну фільтрацію тексту відповідно до заданого алфавіту.*

Звичайні текстові файли містять багато символів окрім власне літер; для статистичної обробки, роботи із класичними шифрами та їх криптоаналізом такі тексти повинні пройти попередню фільтрацію. Фільтр повинен використовувати два режими: звичайний алфавіт та алфавіт з пробілом. Алфавіт повинен задаватись у програмі довільним зручним чином, наприклад, у вигляді рядка, який містить усі символи алфавіту.

«Пробілом» вважається довільний розділювач слів або довільна послідовність таких розділювачів; окрім власне символу пробілу це також знаки пунктуації, цифри, символи кінця рядку, службові символи тощо. Під час аналізу текстів зазвичай нехтуються усі особливості пунктуації, а пробіл використовується виключно як спеціальний символ, який відділяє одне змістовне слово мови від іншого (та опосередковано надає інформацію, наприклад, про середню довжину слова у мові).

Таким чином, побудовані фільтри повинні робити такі дії:

а) Фільтр у режимі звичайного алфавіту:

– усі символи, окрім символів алфавіту, повинні вилучатись;
– прописні літери – замінюватись на відповідні стрічні.

б) Фільтр у режимі алфавіту з пробілом:

– усі символи, окрім символів алфавіту, повинні замінюватись на пробіл;
– послідовність з декількох пробілів повинна замінюватись на один пробіл;
– пробіли на початку та наприкінці тексту повинні видалятись;
– прописні літери – замінюватись на відповідні стрічні.

Наприклад, текстовий рядок українською мовою

`1. Слава Україні! 2. Героям слава!!`

після застосування фільтру у режимі алфавіту з пробілом перетворюється на текст

`слава_україні_героям_слава`

(тут пробіли показані прочерками для зручності), а після застосування фільтру у режимі звичайного алфавіту – на текст

`славаукраїнігероямслава`

Завдання 2. *Обчислити частоти символів та біграм російської мови.*

Для виконання даного завдання вам необхідно підготувати відфільтрований текст російською мовою об'ємом не менше 1 Мб; художні тексти достатнього розміру можна знайти на сайті lib.ru. Дозволяється склеювати декілька різних текстів у один для досягнення потрібного розміру, але не дозволяється дублювати один й той самий текст декілька разів: така регулярність, непритаманна природній мові, зазвичай суттєво спотворює статистичні дані.

Завдання виконується окремо для текстів, відфільтрованих у звичайному алфавіті та в алфавіті з пробілом.

У відфільтрованому тексті необхідно порахувати кількості усіх символів та кількості усіх біграм, тобто пар символів. Відповідні частоти символів та біграм одержуються шляхом

ділення кількості конкретного символу (біграми) на загальну кількість символів (біграм). Для даних обрахунків необхідно написати окремі функції або окрему програму.

При підрахунку частот біграм треба розглядати як пари букв, що перетинаються, так і пари букв, що не перетинаються (тобто рухатися вздовж тексту з кроком 2). Наприклад, в першому випадку текст «**фільтр**» буде розбитий на біграми **фі, іл, ль, ьт, тр**; в другому – на біграми **фі, ль, тр.** Одержані результати не повинні суттєво відрізнятись, однак в першому випадку використовується більше статистики, а тому чисельні дані більш точні.

Таким чином, після виконання завдання ви повинні одержати шість таблиць із частотами:

1) частоти символів із урахуванням пробілу;
2) частоти символів без пробілу;
3) частоти біграм із урахуванням пробілу, біграми з перетином;
4) частоти біграм із урахуванням пробілу, біграми без перетину;
5) частоти біграм без пробілів, біграми з перетином;
6) частоти біграм без пробілів, біграми без перетину.

Завдання 3. *Обчислити індекс відповідності тексту.*

Індексом відповідності тексту $Y = y_1 y_2 ... y_n$ називається величина

$$I(Y) = \frac{1}{n(n-1)} \sum_{t \in Z_m} N_t(Y)(N_t(Y)-1),$$

де $N_t(Y)$ – кількість появ літери *t* у тексті *Y*. Для текстів російської мови значення індексу відповідності повинно наближатись теоретичним значенням $I_M \approx 0{,}055$, в той час як для зашумлених текстів значення індексу відповідності зазвичай ближче до значення $I_0 = \frac{1}{33} \approx 0{,}0303$.

Для виконання даного завдання вам необхідно написати окрему функцію (або програму), яка обчислює індекс відповідності від заданого *відфільтрованого* тексту. З її допомогою обчисліть значення індексу відповідності від текстів, які ви використовувати у завданні 2 (з пробілами та без пробілів відповідно), та порівняйте одержані значення із теоретичними.

Оформлення звіту

Звіт повинен містити такі ключові моменти:

1) усі написані вами програмні коди; дозволяється надавати посилання на github замість включення текстів програм у звіт;

2) результати застосування написаних вами фільтрів: оригінальний текст файлу, фільтр з пробілами, фільтр без пробілів – **5-6 рядків** на кожен пункт, економимо папір, він вам ще знадобиться;

3) шість таблиць із частотами символів та біграм, вказані у завданні 2; також до кожної таблиці із частотами біграм необхідно окремо навести перелік 10 біграм із найбільшими частотами;

4) значення обчислених індексів відповідності та їх порівняння із еталонними значеннями, наведеними у завданні 3.

Таблицю частот символів потрібно подавати відсортованою за спаданням частот. Таблицю частот біграм зручно подавати у вигляді квадратної матриці, індексованої першою та другою літерами біграм.