КРИПТОГРАФІЯ
КОМП'ЮТЕРНИЙ ПРАКТИКУМ №1

Експериментальна оцінка ентропії на символ джерела відкритого тексту

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.

Необхідні теоретичні відомості

1. Визначення ентропії імовірнісних ансамблів

Розглянемо множину символів $Z = \{z_1, z_2, ..., z_n\}$, в якій кожному символу z_i приписана імовірність $p_i = p(z_i)$. Пару $\langle Z, P \rangle$, де $P = \{p_1, p_2, ..., p_n\}$, ми називаємо *імовірнісним ансамблем*, або просто *ансамблем*.

Ентропія імовірнісного ансамблю $\langle Z, P \rangle$ – це величина

$$H(Z) = -\sum_{i=1}^{n} p_i \log p_i .$$

Значення ентропії різниться в залежності від того, яку основу логарифму використовувати. Тут і надалі вважається, що всюди використовується логарифм за основою 2; в цьому випадку одиницею виміру ентропії є біт.

Ентропія показує нам, скільки інформації міститься в даному ансамблі. Ми розуміємо слова «інформація міститься» таким чином: ентропія показує кількість необхідної інформації для однозначного опису ансамблю (тобто для такого опису, який виключає будь-яку невизначеність), тобто визначає не корисні відомості, що вже містяться в ансамблі, а навпаки, «пусті» місця, своєрідну «ємкість» ансамблю. Звідси випливає інше інтуїтивне визначення ентропії як «міри невизначенності». Відповідно, ентропія приймає значення із проміжку $0 \le H(Z) \le \log n$. Значення $H(Z) = 0$ досягається лише для вироджених розподілів, коли одне значення приймається із імовірністю 1, а всі інші – із імовірністю 0. Значення $H(Z) = \log n$ (максимальне можливе значення для ансамблю розміром n) приймається лише для рівноімовірного розподілу.

Розглянемо два ансамблі $\langle X, P \rangle$ та $\langle Y, Q \rangle$; побудуємо спільний ансамбль на множині $X \times Y$. Для цього необхідно задати сукупний розподіл імовірностей $p_{ij} = p(x_i, y_j)$, який повинен задовольняти вимогам нормування: $\sum_{i,j} p_{ij} = 1$, $\sum_{j} p_{ij} = p_i$, $\sum_{i} p_{ij} = q_j$.

Для побудованого таким чином ансамблю можна також визначити ентропію; вона має назву *сукупної ентропії*:

$$H(X,Y) = -\sum_{i,j} p_{ij} \log p_{ij} .$$

Разом із сукупною ентропією розглядають також *умовну ентропію*:

$$H(X \mid Y) = -\sum_{j} q_j H(X \mid y_j) = -\sum_{i,j} p_{ij} \log \frac{p_{ij}}{q_j} .$$

Сукупна ентропія говорить нам, скільки інформації містять обидва ансамблі (із урахуванням взаємних залежностей). Умовна ентропія говорить нам, скільки інформації залишиться в ансамблі *X*, якщо поведінку ансамблю *Y* буде однозначно визначено.

Ансамблі називаються *незалежними*, якщо виконується умова: $\forall i, j : p_{ij} = p_i q_j$. В цьому випадку $H(X,Y) = H(X) + H(Y)$ і $H(X \mid Y) = H(X)$.

Спільною (або *взаємною*) *інформацією* ансамблів *X* та *Y* називається величина

$$I(X,Y) = H(X) - H(X \mid Y) .$$

Іншими словами, ми вважаємо, що кількість інформації про *X*, яку ми одержуємо зі знання *Y*, визначається зменшенням рівня невизначеності (тобто, нестрого кажучи, зменшенням нашого «незнання» про *X*).

Для взаємної інформації двох ансамблів виконуються також такі співвідношення:

а) $I(X,Y) = H(Y) - H(Y \mid X)$;

б) $I(X,Y) = H(X) + H(Y) - H(X,Y)$;

в) $0 \le I(X,Y) \le \min\{H(X), H(Y)\}$.

Значення $I(X,Y) = 0$ має місце тоді та тільки тоді, коли ансамблі незалежні.

2. Методи оцінювання ентропії джерел символів

Текстом ми називаємо довільну послідовність символів $x_1, x_2, x_3, \ldots$ з деякого скінченного алфавіту *Z*, яку формує деяке джерело (генератор текстів). *n-грамою* називається відрізок тексту з *n* послідовних символів $(x_{i+1}, x_{i+2}, \ldots, x_{i+n})$, $i \ge 0$. Джерело вважається повністю описаним, якщо для кожної *n*-грами ($n \ge 1$) заданий розподіл імовірностей $P(x_{i+1} = z_1, x_{i+2} = z_2, \ldots, x_{i+n} = z_n)$, $i \ge 0$, $z_j \in Z$, $1 \le j \le n$. Джерело називається *стаціонарним*, якщо

$$P(x_{i+1} = z_1, x_{i+2} = z_2, \ldots, x_{i+n} = z_n) = P(x_1 = z_1, x_2 = z_2, \ldots, x_n = z_n)$$

для всіх $i \ge 0$ і довільних $n \ge 1$, $z_j \in Z$, $1 \le j \le n$, тобто якщо розподіл всіх *n*-грам не залежить від зсуву за часом (або, що те ж саме, від місця появи в тексті).

Ентропія на символ стаціонарного джерела визначається як

$$H_\infty = \lim_{n \to \infty} H_n , \text{ де } H_n = \frac{1}{n} H(x_1, x_2, \ldots, x_n),$$

а $H(x_1, x_2, \ldots, x_n)$, в свою чергу, – ентропія *n*-грами відкритого тексту $(x_1, x_2, \ldots, x_n)$:

$$H(x_1, x_2, ..., x_n) = -\sum_{z_1, z_2, ..., z_n} P(x_1 = z_1, ..., x_n = z_n) \cdot \log_2 P(x_1 = z_1, ..., x_n = z_n).$$

Величина H_n називається *питомою ентропією на символ n-грами*, а H_∞ – *ентропією джерела* або *ентропією мови*. Максимальне значення H_∞ приймає в тому випадку, коли всі символи тексту незалежні і рівноімовірні. Тоді $H_\infty = H_0 = \log_2 m$, де m – кількість букв в алфавіті Z.

Для реальних джерел відкритого тексту (таких, як природні мови) значення H_∞ набагато менше за H_0 через нерівноімовірність букв алфавіту в тексті та залежність між ними. Як послідовні наближення до H_∞ можна розглядати значення $H_1, H_2, H_3, ...$, які враховують відповідно імовірності букв алфавіту в мові, зв'язок букв всередині біграм, триграм і т.д. Проте експериментальна оцінка H_n при достатньо великих n нездійсненна з огляду на величезне число можливих значень n-грам. Тому розроблені підходи, що дозволяють непрямо оцінити значення H_∞ за допомогою деяких статистичних дослідів. Один з таких підходів спирається на той факт, що H_∞ може бути також визначена як границя умовних ентропій:

$$H_\infty = \lim_{n\to\infty} H^{(n)}, \text{ де } H^{(n)} = H(x_n \mid x_1, x_2, ..., x_{n-1}).$$

Величина $H^{(n)}$ називається *умовною ентропією джерела*; вона визначає, скільки інформації про наступний символ ми матимемо із значень $(n-1)$ попередніх. Метод оцінки $H^{(n)}$ полягає в тому, що експериментатор за випадково вибраною $(n-1)$-грамою вгадує наступну за нею n-ту букву тексту. Нехай $q_1^{(n)}, q_2^{(n)}, ..., q_m^{(n)}$ – імовірності того, що буква буде правильно вгадана з 1-ої, 2-ої, …, m-тої спроби (природно, число спроб не може бути більше m – числа букв в алфавіті). Тоді має місце нерівність

$$\sum_{i=1}^{m-1} i(q_i^{(n)} - q_{i+1}^{(n)}) \log_2 i + m q_m^{(n)} \log_2 m \le H^{(n)} \le -\sum_{i=1}^{m} q_i^{(n)} \log_2 q_i^{(n)}$$

Ця нерівність дає не вельми точну оцінку H_∞, так як ліва і права його частини не прямують до єдиної границі при $n \to \infty$ і, крім того, через неможливість врахування експериментатором усіх закономірностей мови наведена оцінка буде завищеною. Проте з огляду на простоту реалізації наведений метод є цілком придатним для навчальних потреб.

Надлишковість джерела відкритого тексту (мови) дорівнює $R = 1 - \frac{H_\infty}{H_0}$ і характеризує величину можливого ущільнення тексту деякою схемою кодування символів без втрати його змісту.

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.

Методичні вказівки

Звичайні текстові файли містять багато символів окрім власне літер; для обчислення значень ентропій вони повинні пройти попередню фільтрацію: всі символи, окрім текстових, повинні вилучатись або замінюватись на пробіли; прописні літери – замінюватись на відповідні стрічні; послідовність пробілів (або інших розділових знаків, наприклад, символів кінця рядку) повинна трактуватись як один пробіл або вилучатись, якщо пробіл не входить до алфавіту.

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). Одержані результати не повинні суттєво відрізнятись, однак в першому випадку використовується більше статистики, а тому чисельні дані більш точні. Таблицю частот символів потрібно подавати відсортованою за спаданням частот. Таблицю частот біграм зручно подавати у вигляді квадратної матриці, індексованої першою та другою літерами біграм.

Програма CoolPinkProgram використовує текст, що лежить у допоміжному файлі text. Цей текст написаний російською мовою без знаків пунктуації та великих літер; буква «ё» замінена буквою «е», а «ъ» – буквою «ь». Пробіл також вважається буквою. Таким чином, кількість букв алфавіту $m = 32$. При підрахунку $H^{(10)}$, $H^{(20)}$, $H^{(30)}$ необхідно виконати не менш ніж 50 експериментів.

Оформлення звіту

Звіт до комп’ютерного практикуму оформлюється згідно зі стандартними правилами оформлення наукових робіт, за такими винятками:

- дозволяється використовувати шрифт Times New Roman 12pt та одинарний інтервал між рядками;
- для оформлення фрагментів текстів програм дозволяється використовувати шрифт Courier New 10pt та друкувати тексти в дві колонки;
- дозволяється не починати нові розділи з окремої сторінки.

До звіту можна не включати анотацію, перелік термінів та позначень та перелік використаних джерел. Також не обов’язково оформлювати зміст.

Звіт має містити:

- мету комп’ютерного практикуму;
- постановку задачі та варіант завдання;
- хід роботи, опис труднощів, що виникали, та шляхів їх розв’язання;
- таблиці частот букв і біграм тексту, одержані значення H_1 та H_2, оцінки для $H^{(10)}$, $H^{(20)}$, $H^{(30)}$ (включно із відповідними скріншотами);
- оцінку надлишковості R російської мови у різних моделях відкритого тексту;
- висновки.

Тексти всіх програм здаються викладачеві в електронному вигляді для перевірки на плагіат. До захисту теоретичної частини комп'ютерного практикуму допускаються тільки ті студенти, які оформили звіт та пройшли перевірку програмного коду.

Контрольні запитання

1) Дайте визначення ентропії, сукупної ентропії, умовної ентропії. Який зміст несуть ці поняття?

2) Які визначення ентропії на символ джерела ви знаєте?

3) Порівняйте одержані значення H_1, H_2, $H^{(10)}$, $H^{(20)}$, $H^{(30)}$. Зробіть висновки.

4) Що таке надлишковість джерела? Яка надлишковість російського письмового тексту згідно ваших даних?

5) Які моделі відкритих текстів розглядаються у криптографії?

Оцінювання практикуму

За виконання комп'ютерного практикуму студент може одержати до 7 рейтингових балів; зокрема, оцінюються такі позиції:

- реалізація програм – до трьох балів (в залежності від правильності та швидкодії);
- теоретичний захист роботи – до трьох балів;
- своєчасне виконання практикуму – 1 бал;
- несвоєчасне виконання роботи – (-1) бал за кожен тиждень пропуску.

Програмний код, створений під час виконання комп'ютерного практикуму, перевіряється на наявність неправомірних запозичень (плагіату) за допомогою сервісу *Stanford MOSS Antiplagiarism*. У разі виявлення в програмному коді неправомірних запозичень реалізація програм оцінюється у 0 балів, а за виконання практикуму студент одержує штраф (–10) балів.

Студенти допускаються до теоретичного захисту тільки за умови оформленого звіту з виконання практикуму та проходження перевірки програмного коду.