Практика №1. Построение словаря. Статистический анализ текста. Энтропия текста.

ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

Определение 1.1. Вероятностной схемой X называется

X	x_1	x_2	...	x_n
P	p_1	p_2	...	p_n

где x_1, x_2, ..., x_n- полная группа попарно несовместных событий, а p_1, p_2, …, p_n – соответствующие вероятности.

Определение 1.2. Количеством информации, содержащимся в сообщении x, называется $h(x) = -\log p(x)$. (Основание логарифма, если не оговорено противное, принимается равным 2.)

Определение 1.3. Энтропией вероятностной схемы X, называется $H(X) = -\sum_{i=1}^{n} p_i \cdot \log p_i$.

Значение функции f(t) = t·log t при t = 0 считаем равным нулю, доопределяя её в этой точке по непрерывности. Таким образом, эта функция определена, по крайней мере, на отрезке [0;1].

Пусть имеются две схемы X и У

X	x_1	x_2	...	x_n
P	p_1	p_2	...	p_n

У	y_1	y_2	...	y_m
P	q_1	q_2	...	q_m

Определение 1.4. Энтропией произведения вероятностных схем X и У, называется

$$H(XY) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i y_j) \cdot \log p(x_i y_j)$$

Если схемы X и У независимы, то энтропия произведения вероятностных схем равна сумме энтропий каждой схемы: $H(XY) = H(X) + H(Y)$.

Определение 1.5. Условной энтропией вероятностной схемы У относительно схемы X называется:

$$H(Y \mid X) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i)\, p(y_j \mid x_i) \log p(y_j \mid x_i),$$

где $p(y_j \mid x_i)$ – условная вероятность события y_j при условии, что получено сообщение x_i.

Энтропия произведения и условная энтропия связаны между собой соотношениями:
$H(XY) = H(X) + H(Y \mid X) = H(Y) + H(X \mid Y)$.

ПРИМЕР

Задание. Событие *A* в каждом из *n* повторных независимых испытаний происходит с вероятностью *p*. Найти энтропию числа появлений события *A*. Составить соответствующую вероятностную схему. Выяснить характер изменения энтропии в зависимости от изменения *p* на промежутке [0;1] при значении *n* = 1, построив график соответствующей функции H(*p*). Определить её наименьшее и наибольшее значение.

Рассмотрим энтропию числа появлений события A в серии из n испытаний.

Если n=1 и X- число появлений события A в серии из n испытаний, то

X	0	1
P	q	p

где q=1-p.

По определению 1.3, функция $H(p) = -p \cdot \log p - (1-p)\log(1-p)$. Построим график H(p) (рис.1):

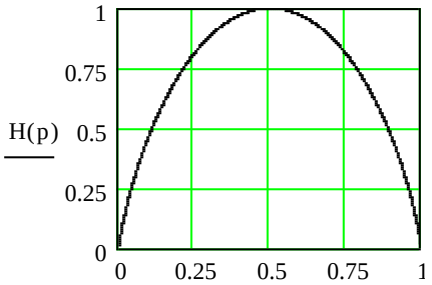

Рис.1 График функции H(p)

При p=0,5 функция H(p) достигает максимума H(0,5)=1, при p=0 или p=1 функция H(p) достигает минимума H(0)=H(1)=0. Функция возрастает на промежутке [0;0,5] и убывает на отрезке [0,5;1].

Таким образом, наименьшее значение, равное нулю, энтропия рассматриваемой вероятностной схемы принимает при p=0 и при p=1, то есть в тех случаях, когда исход опыта с вероятностной схемой X однозначно определён до его проведения. Наибольшее же значение, равное одному биту, энтропия данной схемы принимает только при p=0,5 , то есть в том случае, когда с равными вероятностями можно предполагать, что в результате испытания произойдёт или не произойдёт событие A, что соответствует наибольшей неопределённости исхода опыта с вероятностной схемой X до его проведения. При приближении p к 0,5 , то есть с увеличением неопределённости, энтропия возрастает, а при приближении p к концам отрезка [0;1], то есть с уменьшением неопределённости, энтропия убывает. Следовательно, приведённые выше рассуждения подтверждают тезис о том, что энтропия является мерой неопределённости вероятностной схемы до проведения испытаний с ней.

Так как информацию можно рассматривать как неопределённость, снимаемую при получении сообщения, то можно дать следующее определение.

Определение 1.6. Пусть проводится k независимых испытаний с вероятностной схемой X. Тогда количеством информации, которое несёт в себе сообщение о результатах этой серии опытов, называется $I = k \cdot H(X)$.

В частном, но с практической точки зрения очень важном, случае, когда вероятностная схема X указывает вероятности появления символов алфавита от некоторого стохастического источника сообщений, причём буквы появляются независимо друг от друга, k интерпретируется как длина сообщения, полученного от данного источника, $H(X)$ – среднее количество информации, которое несёт в себе одна буква достаточно длинного сообщения, I - количество информации, которое несёт в себе сообщение из k символов.

Для случая равновероятных и взаимно независимых m символов $I = k \cdot \log m$.

Если схемы X и У статистически зависимы, то возможно измерение количества информации о системе X, которое дает наблюдение за системой У.

Определение 1.7. Информационной избыточностью называется величина

$$D = 1 - \frac{H}{H_{\max}}$$

Частным видом избыточности является избыточность, обусловленная неравномерным распределением символов сообщения: $D_p = 1 - \dfrac{-\sum_i p_i \cdot \log p_i}{\log m}$

ПРИМЕР

Задание. Произвести статистическую обработку данного сообщения, считая, что источник сообщений периодически, достаточно долго выдаёт следующую последовательность символов 12342334551233. Определить энтропию, приходящуюся в среднем на одну букву и на одно двухбуквенное сочетание. Найти длину кода при равномерном кодировании и избыточность.

Пусть имеется сообщение:
1234233455123312342334551233123423345512 33... .
Составим схему появления однобуквенных сочетаний:

X	1	2	3	4	5	∑
n	2	3	5	2	2	14
w	$\frac{2}{14}$	$\frac{3}{14}$	$\frac{5}{14}$	$\frac{2}{14}$	$\frac{2}{14}$	1

Энтропия схемы X равна

$$H(X) = -\left[3 \cdot \frac{2}{14} \cdot \log\frac{2}{14} + \frac{5}{14} \cdot \log\frac{5}{14} + \frac{3}{14} \cdot \log\frac{3}{14}\right] = 2{,}21$$

Составим схему $\overline{XY}$ появления двухбуквенных сочетаний

XУ	12	23	31	34	35	42	45	51	55	∑
n	2	3	1	2	2	1	1	1	1	14
w	$\frac{2}{14}$	$\frac{3}{14}$	$\frac{1}{14}$	$\frac{2}{14}$	$\frac{2}{14}$	$\frac{1}{14}$	$\frac{1}{14}$	$\frac{1}{14}$	$\frac{1}{14}$	1

Энтропия, приходящаяся на одно двухбуквенное сочетание, составляет

$$H(\overline{XY}) = -\left[3 \cdot \frac{2}{14} \cdot \log\frac{2}{14} + 5 \cdot \frac{1}{14} \cdot \log\frac{1}{14} + \frac{3}{14} \cdot \log\frac{3}{14}\right] = 3{,}039$$

Найдем длину кода при равномерном кодировании однобуквенных сочетаний[1]:
m=5, $l = \lceil \log 5 \rceil = 3$

При этом возникает избыточность округления $D_0 = 1 - \frac{\log 5}{3} = 0{,}226$

Подсчитаем информационную избыточность:

$$D_p = 1 - \frac{2{,}21}{\log 5} = 0{,}048$$

[1] $\lceil x \rceil$ – округление в большую сторону.

ЗАДАНИЕ

Выполнить последовательно следующие этапы:

1. Прочитать текст из файла
2. Привести все слова к нижнему регистру, удалить знаки препинания и пробелы (результат сохранить в отдельный файл).
3. Подсчитать частоту появления однобуквенных и двухбуквенных сочетаний.
4. Определить энтропию, приходящуюся в среднем на одну букву и на одно двухбуквенное сочетание.
5. Найти длину кода при равномерном побуквенном кодировании и избыточность.
6. Удалить 20% наиболее часто встречающихся символов, проанализировать как изменится энтропия однобуквенных сочетаний и предположить, почему так происходит.
7. Удалить 20% наиболее редко встречающихся символов, проанализировать как изменится энтропия однобуквенных сочетаний и предположить, почему так происходит.