**Исчисление количества информации**

Методические указания к лабораторным работам по дисциплине

«Теория информации»

**1 Общие положения**

Общее количество неповторяющихся сообщений, которое может быть составлено из алфавита  путем комбинирования по  символов в сообщении, определяется соотношением

. (1.1)

Неопределенность, приходящаяся на символ первичного (кодируемого) алфавита, составленного из равновероятных и вазимонезависимых символов равна

. (1.2)

При этом основание логарифма влияет лишь на удобство вычисления. Так для случая вычисления энтропии имеем:

а) в двоичных единицах

,

б) в десятичных единицах

,

где , ;

в) в натуральных единицах

,

где , .

Так как информация есть неопределенность, снимаемая при получении сообщения, то количество информации может быть представлено как произведение общего числа сообщений  на среднюю энтропию , приходящуюся на одно сообщение

. (1.3)

Для случаев равновероятных и взаимонезависимых символов первичного алфавита количество информации в  сообщениях алфавита  равно

.

Для неравновероятных алфавитов энтропия на символ алфавита

. (1.4)

а количество информации в сообщении, составленном из  неравномерных символов,

. (1.5)

При решении задач, в которых энтропия вычисляется как сумма произведений вероятностей на их логарифм, вероятности всегда должны представлять группу полных событий, независимо от того, являются ли они безусловными , условными  или вероятностями совместных событий .

Количество информации определяется исключительно характеристиками первичного алфавита, объем – характеристиками вторичного алфавита. Объем информации

, (1.6)

где  – средняя длина кодовых слов вторичного алфавита. Для равномерных кодов (все комбинации кода содержат одинаковое количество разрядов)

,

где  – длина кода (число элементарных посылок в коде). Согласно (1.3), объем равен количеству информации, если , т.е. в случае максимальной информационной нагрузки на символ сообщения. Во всех остальных случаях .

Например, если кодировать в коде Бодо некоторый равновероятный алфавит, состоящий из 32 символов, то

;

.

Если кодировать в коде Бодо русский 32-буквенный алфавит, то без учета корреляции между буквами количество информации

; ; ,

т.е. если в коде существует избыточность и , то объем в битах всегда больше количества информации в тех же единицах.

**2 Задание на работу**

Определить количество информации в сообщении и энтропию сообщения на русском языке, содержащем фамилию, имя, отчество, год, месяц и день рождения студента для следующих случаев:

1) для равновероятных символов алфавита;

2) для неравновероятных символов алфавита;

3) для случая двубуквенных сочетаний;

4) для случая трехбуквенных сочетаний;

5) проанализировать полученные значения и сделать выводы.

Вероятность символов появления букв алфавита выбрать в соответствии с Приложением 1.

**3 Ход работы**

При выполнении работы необходимо выделить две группы символов: отдельно буквы и отдельно цифры. Это связано с тем, что появление цифр в тексте носит равновероятный характер.

**3.1 Вычисление количества информации и энтропии сообщения для случая равновероятных символов алфавита**

Для вычисления количества информации при равновероятном появлении символов необходимо воспользоваться формулой Хартли

, (3.1)

где  ‒ общее число букв в алфавите.

В рассматриваемом случае весь алфавит для передачи буквенных сообщений на русском языке представлен 32 символами, а алфавит для передачи численных сообщений ‒ 10 символами.

Тогда, количество информации, которое приходится на одну букву текстового алфавита определяется из соотношения

.

Аналогично количество информации на один символ цифрового сообщения информацию может быть рассчитано как

.

Для вычисления количества информации, которое приходится на все сообщение, необходимо воспользоваться выражением

, (3.2)

где  ‒ число букв в сообщении, ‒ число цифр в сообщении.

Рассмотрим пример сообщения:

**АЛЬЧАКОВ\_ВАСИЛИЙ\_ВИКТОРОВИЧ\_21\_09\_1978**

Количество букв в сообщении с учетом пробелов равно 30, количество цифр ‒ 8, следовательно, количество информации буквенной части сообщения равно

,

Общее количество информации в сообщении, состоящем из равновероятных букв и цифр, определяется по формуле

.

Для расчета энтропии сообщения следует использовать соотношение

.

**3.2 Вычисление количества информации и энтропии сообщения для случая неравновероятных символов алфавита**

Для случая неравновероятных символов алфавита расчеты необходимо начать с составления таблицы вероятностей появления символов алфавита в сообщении. Пример такой таблицы приведен ниже. Следует обратить внимание, что в конец таблицы добавлены три пробела, которые являются частью числового сообщения, однако участвуют в расчете его буквенной части.

|  |  |  |
| --- | --- | --- |
| **Символ сообщения** |  |  |
| А | 0,062 | 0,249 |
| Л | 0,035 | 0,169 |
| Ь | 0,014 | 0,086 |
| Ч | 0,012 | 0,076 |
| А | 0,062 | 0,249 |
| К | 0,028 | 0,144 |
| О | 0,09 | 0,313 |
| В | 0,038 | 0,179 |
| \_ | 0,175 | 0,44 |
| В | 0,038 | 0,179 |
| А | 0,062 | 0,249 |
| С | 0,045 | 0,201 |
| И | 0,062 | 0,249 |
| Л | 0,035 | 0,169 |
| И | 0,062 | 0,249 |
| Й | 0,010 | 0,066 |
| \_ | 0,175 | 0,44 |
| В | 0,038 | 0,179 |
| И | 0,062 | 0,249 |
| К | 0,028 | 0,144 |
| Т | 0,053 | 0,225 |
| О | 0,09 | 0,313 |
| Р | 0,04 | 0,185 |
| О | 0,09 | 0,313 |
| В | 0,038 | 0,179 |
| И | 0,062 | 0,249 |
| Ч | 0,012 | 0,076 |
| \_ | 0,175 | 0,44 |
| \_ | 0,175 | 0,44 |
| \_ | 0,175 | 0,44 |

После составления таблицы необходимо вычислить энтропию сообщения, которая приходится на его буквенную часть. Для этого необходимо воспользоваться формулой

. (3.3)

Выполнив расчеты получим .

Далее, можно вычислить количество информации, содержащейся в буквенной части сообщения по формуле

. (3.4)

Подставляя в (3.4) численные значения параметров, получим

.

Общее количество информации для буквенно-цифрового сообщения рассчитывается из соотношения . Отметим, что количество информации для цифровой части сообщения определяется точно также, как и в подразделе 3.1.

Общая энтропия сообщения рассчитывается аналогично

.

**3.3 Вычисление количества информации и энтропии сообщения для случая дву- и трехбуквенных сочетаний**

Для русского алфавита, с учетом пробела, а также с учетом неравновероятности появления символов ; с учетом двубуквеных сочетаний ; с учетом трехбуквенных сочетаний .

Исходя из этого можно записать выражение для количества информации

,

.

Энтропию сообщений каждого вида вычислим по уже знакомой формуле

,

.

**4 Содержание отчета**

После выполнения расчетной части лабораторной работы каждый студент должен подготовить отчет, в который входит:

1. титульный лист;
2. цель работы;
3. вариант задания;
4. расчетная часть в соответствии с п. 3 данных методических указаний;
5. выводы по работе**;**
6. приложение – исходный код программы.

**Приложение 1**

Распределение вероятностей букв в русских текстах

