**Фестиваль педагогического мастерства – 2024**

**Мастер-класс «Сжатие информации»**

**Мордовских Е.П.**

**Слайд 1.** Здравствуйте, дорогие коллеги.

* А какие еще слова мы употребляем ежедневно, очень часто? *(отвечают)*

Очень замечательно, запомните, пожалуйста, эти варианты, они нам пригодятся.

**Слайд 2.** «Кто владеет информацией, тот владеет миром». Автор этой фразы Натан Майер Ротшильд **(клик)**, английский банкир, который спонсировал английскую военную компанию против Наполеона. Он прославился тем, что, когда узнал о поражении Наполеона в битве при Ватерлоо **(клик)**, он моментально прибыл на лондонскую биржу и, сокрушаясь о том, что Наполеон битву выиграл, начал продавать английские ценные бумаги. Его примеру последовали множество людей, которые торговали на бирже. А его тайные агенты эти бумаги скупали. И когда через день выяснилось, что Наполеон битву проиграл, оказалось, что Ротшильд успел разбогатеть на 40 млн. фунтов.

Этот день считается первым днем, когда общество перешло в стадию информационного. То есть такое общество, в котором информация является главной мировой ценностью.

* Как вы думаете, может ли быть главная мировая ценность без какой-либо возможности ее измерить или оценить? *(отвечают)*

Конечно, нет. Не будем говорить много умных слов об измерении информации. Лучше всего нам об этом расскажет следующая фотография.

**Слайд 3.** Перед вами один из дата-центов компании Yandex, который располагается на территории индустриального парка «Грабцево» в Калуге. Вот этот комплекс зданий, я думаю можно оценить его масштабы, предназначен исключительно для хранения информации. В нем размещены серверы с информацией **(клик).** Таких гипермасштабируемых дата-центов в мире более 800. Более мелких центов обработки данных уже свыше 7500. И вы понимаете, что затраты на постройку и на обслуживание таких центров очень большие.

* Вопрос! Можно ли каким-нибудь образом минимизировать затраты, чтобы центр с такой же функциональностью хранил больше информации? Можно ли с этой информацией сделать что-то, чтобы туда входило больше информации? *(отвечают)*

Именно об этом я и хочу сегодня с вами поговорить. Тема моего мастер-класса «Сжатие информации». В основе сжатия информации лежит очень простой принцип: чем чаще встречается или передается информация, тем короче она должна быть.

Вспомните, пожалуйста, слова, которые вы произносили пару минут назад – здравствуйте, спасибо, как дела. Отметим, что эти фразы, которые мы произносим часто, являются очень короткими. Эта тенденция общая для всех видов информации, как для таких естественных как текст, так и для тех, которые люди изобретают сами. Думаю, все вы хоть раз в жизни видели игру в футбол.

* Скажите, пожалуйста, когда судья делает так *(свисток)*, что это обозначает? *(ответ – остановка игры, нарушение правил)*
* А когда он делает так *(свисток+желтая карточка)*?
* А что означает желтая карточка? *(ответ – предупреждение)*
* Хорошо, а вот так *(свисток+красная карточка+жест)*? *(ответ – удаление игрока с поля)*
* На какое из этих 3 действий сейчас затратилось больше времени, а на какое меньше? *(ответ – на удаление больше, на предупреждение меньше, на остановку еще меньше)*
* А что в матче встречается чаще, а что реже? *(остановка игры чаще, удаление с поля реже всего)*

То есть тенденция совершенно очевидна. И именно ей мы воспользуемся, чтобы сжать информацию.

**Слайд 4.** Рассмотрим пример: русская скороговорка «Шла Саша по шоссе и сосала сушку».

В основе скороговорки лежит прием аллитерация, повторение звуков с целью создания определенного настроения. А мы этот прием используем для того, чтобы эту скороговорку сжать.

* Какие буквы встречаются наиболее часто? *(отвечают)*

Давайте выпишем все буквы этой скороговорки включая пробелы **(клик1)**, и выпишем сколько раз они встречаются в порядке убывания. Буквы С,А,Ш встречаются наиболее часто, обратите внимание, что пробел является полноценным символом, который входит в скороговорку. Всего у нас **(клик2)** 32 символа. А это значит, что если мы будем хранить ее в памяти компьютера, то **(клик3)** она займет 32 байта или 256 бит информации.

**Слайд 5.** Сейчас мы будем сжимать этот текст по одному из существующих методов, он называется методом Шеннона-Фано по фамилиям ученых, которые его изобрели. Здесь придется считать в уме. Принцип основан на следующем: мы должны провести горизонтальную черту между двумя буквами таким образом, чтобы общие суммы количества букв выше черты и ниже черты были как можно ближе друг к другу. Букв 32, значит половина это 16.

* Сколько букв нужно сгруппировать по порядку, чтобы их сумма была примерно 16? После какой буквы будет черта? *(отвечают – после А)* **(клик1)**

Спасибо большое, вы правы. Сумма верхней группы букв 17, а нижней 15. Дальше мы берем верхнюю группу и приписываем с не 0, а к нижней группе 1 **(клик2).**

Теперь у нас 2 группы букв – верхняя и нижняя. И с каждой их них мы делаем абсолютно то же самое.

* Где в верхней группе нужно провести горизонтальную черту, чтобы поделить верхнюю группу на 2 части, чтобы сумма количества символов были как можно ближе друг к другу? *(отвечают – после С)* **(клик3).**
* Совершенно верно. Верхней половинке приписываем 0, а нижней 1 **(клик4)**. Абсолютно также делаем с нижней группой, их сумма 15. Где нужно провести черту чтобы два числа были максимально близки друг к другу? *(отвечают – после О)* **(клик5).** Верхней половинке приписываем 0, а нижней 1 **(клик6)**.

Обратите снимание, что буква С осталась в своей группе одна, а это значит, что мы можем выписать ее код – 00 **(клик7)**. Берем следующую группу, в которую сейчас входят пробел и А, она делится на 2 половинки единственно возможным образом **(клик8)**, а это значит, что мы можем выписать их коды.

* Какой код будет у пробела? А у буквы А? *(отвечают)* **(клик9).**

Абсолютно таким же способом мы будем продолжать выполнять этот алгоритм **(клик10) (клик11).** Теперь может сказать код букв Ш и О **(клик12)**. Снова разбиваем группы **(клик13)** и определяем кода следующих букв Л и У **(клик14)**.

Снова разбиваем 2 группы **(клик15)** и снова **(клик16)** и получаем коды сразу 4 букв **(клик17)**. Мы получаем достаточно длинный коды у последних букв и это хорошо, они встречаются достаточно редко, а значит у нас все закодировано в соответствии с принципом чем чаще встречается или передается информация, тем короче она должна быть.

Буква С кодируется 2 битами, а всего 6 букв С, значит на их кодировку уйдет 2\*6=12 бит… так можно посчитать суммарное количество для каждой буквы в сообщении **(клик18)**. Получается общий вес сообщения 102 бита **(клик19)**. Напомню, что до сжатия сообщение занимало 256 бит, то есть сжатие в 2,5 раза.

Я попрошу зал поработать, у вас есть листочек с другой скороговоркой. Она звучит как «Карл у Клары украл кораллы, а Клара у Карла украла кларнет». Я прошу вас закодировать текст, т.е. получить коды букв для каждого символа, входящего в эту скороговорку. В то время как зал будет по методу Шеннона-Фано находить коды символов, я попрошу фокус-группу, у которой уже есть листочки с уже посчитанными кодами, выполнить очень важную задачу. Я прошу оценить, насколько сильно нам удастся эту скороговорку сжать. То есть посмотреть сколько 0 и 1 войдет в код полученной скороговорки, если ее закодировать.

Пожалуйста, считайте, метод пока перед вами на экране. На всякий случай напоминаю, что верхней половинке мы присваиваем 0, а нижней 1.

* Может быть, кто-нибудь получил код для буквы А? Код для пробела и буквы Л?

…

Давайте с вами начнем сверяться.

**Слайд 6.** Буквы и частота их использования перед вами. Давайте сравним с вами те коды, которые вы успели получить **(клик)**. Коды перед вами на доске. Действительно, код буквы А – 00, у пробела – 010 и у Л – 011. Остальные коды вы видите на доске.

У меня вопрос к фокус-группе. Скажите, пожалуйста, сколько 0 и 1 займет полученный двоичный код. *(Ответ 180)*

**Слайд 7.** Итак, скороговорка вместе со всеми пробелами, запятой, которые кодируются также как и обычные символы, состоит из 58 символов, а значит она занимала 58 байт = 464 бита. После применения метода сжатия **(клик)** она стала занимать 180 бит. Т.е. за счет того, что какие-то буквы повторялись часто, мы смогли сжать **(клик)** ее в 2,6 раза.

**Слайд 8.** Теперь давайте поговорим о проблемах, связанных со сжатием информации. Информация подобна материи, и принципы сжатия информации напоминают принципы сжатия материи. Сейчас мы проведем эксперимент. В качестве информации у нас будут выступать одинаковые куски материи, а сжатая информация эту будет та же самая ткань, но в пакете.

Я приглашаю 1-ую пару участников. Возьмите пакет и ткань. Ваша задача сжать информацию, поместив ткань в пакет. На выполнение у вас есть ровно 6 секунд, время пошло… Спасибо, ваше время закончилось, пожалуйста, оставьте результат таким каким он получился.

Я приглашаю 2-ую пару участников, чтобы выполнить аналогичную задачу, но ваше время 10 секунд. Время пошло. Это больше чем 6 секунд, но ненамного… Спасибо.

Я приглашаю 3-ю пару участников, у вас будет также 10 секунд на выполнение сжатия, но перед этим у вас будет время, чтобы договориться, как вы будете это делать. Договаривайтесь… Будем считать, что вы договорились. Приступайте. Спасибо, ваше время закончилось.

* У меня вопрос к 1ой паре: какие проблемы возникли? *(ответ)*
* Фактически вам не удалось сжать информацию. Правильно? *(ответ)* А раз ее не удалось сжать, значит ее потом невозможно восстановить. **(клик)** Самое главное требование к сжатию информации – это возможность ее восстановить, т.е. обратимость в исходную информацию после сжатия.
* Вопрос ко 2ой паре: удалось ли вам выполнить задачу? *(ответ)* Мы сталкиваемся с таким моментом как скорость сжатия. **(клик)** У вас было больше времени и нам нужно было сжимать с определенной скоростью, чтобы успеть. Медленные методы сжатия на практике не применимы.
* Вопрос к 3ей паре, о чем вы договаривались? *(ответ)* **(клик)** Метод – это самое главное, от него зависит получится ли у вас результат.

Вот 3 момента, на которые нужно обращать внимание, независимо от того, каким видом представлена информация.

Где и как вы можете использовать сжатие информации в вашей предметной области, однозначного ответа у нас нет, потому что существует огромное количество видов информации. Однако постоянно возникают новые способы сжатия и хранения информации, а это значит, что до сих пор не найден самый лучший, самый оптимальный способ сжатия. Возможно, вы или ваши ученики изобретут его. И помните, что тот, кто владеет информацией, владеет миром! **Слайд 9.**

**Скороговорка** «**КАРЛ У КЛАРЫ УКРАЛ КОРАЛЛЫ,**

**А КЛАРА У КАРЛА УКРАЛА КЛАРНЕТ**»

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Символ | Количество символов в сообщении | Код символа | Длина кода символа | Вес символа в битах |
| А | 12 | 00 | 2 |  |
| \_ | 10 | 010 | 3 |  |
| Л | 9 | 011 | 3 |  |
| К | 8 | 100 | 3 |  |
| Р | 8 | 101 | 3 |  |
| У | 4 | 1100 | 4 |  |
| Ы | 2 | 1101 | 4 |  |
| О | 1 | 11100 | 5 |  |
| , | 1 | 11101 | 5 |  |
| Н | 1 | 11110 | 5 |  |
| Е | 1 | 111110 | 6 |  |
| Т | 1 | 111111 | 6 |  |
| Все сообщение весит | | | | бит |

**Скороговорка** «**КАРЛ У КЛАРЫ УКРАЛ КОРАЛЛЫ,**

**А КЛАРА У КАРЛА УКРАЛА КЛАРНЕТ**»

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Символ | Количество символов в сообщении | Код символа | Длина кода символа | Вес символа в битах |
| А | 12 | 00 | 2 |  |
| \_ | 10 | 010 | 3 |  |
| Л | 9 | 011 | 3 |  |
| К | 8 | 100 | 3 |  |
| Р | 8 | 101 | 3 |  |
| У | 4 | 1100 | 4 |  |
| Ы | 2 | 1101 | 4 |  |
| О | 1 | 11100 | 5 |  |
| , | 1 | 11101 | 5 |  |
| Н | 1 | 11110 | 5 |  |
| Е | 1 | 111110 | 6 |  |
| Т | 1 | 111111 | 6 |  |
| Все сообщение весит | | | | бит |

**Скороговорка** «**КАРЛ У КЛАРЫ УКРАЛ КОРАЛЛЫ,**

**А КЛАРА У КАРЛА УКРАЛА КЛАРНЕТ**»

|  |  |  |  |
| --- | --- | --- | --- |
| Символ | Количество символов в сообщении |  | Код символа |
| А | 12 |  |  |
| \_ | 10 |  |  |
| Л | 9 |  |  |
| К | 8 |  |  |
| Р | 8 |  |  |
| У | 4 |  |  |
| Ы | 2 |  |  |
| О | 1 |  |  |
| , | 1 |  |  |
| Н | 1 |  |  |
| Е | 1 |  |  |
| Т | 1 |  |  |

Всего 58 символов

**Скороговорка** «**КАРЛ У КЛАРЫ УКРАЛ КОРАЛЛЫ,**

**А КЛАРА У КАРЛА УКРАЛА КЛАРНЕТ**»

|  |  |  |  |
| --- | --- | --- | --- |
| Символ | Количество символов в сообщении |  | Код символа |
| А | 12 |  |  |
| \_ | 10 |  |  |
| Л | 9 |  |  |
| К | 8 |  |  |
| Р | 8 |  |  |
| У | 4 |  |  |
| Ы | 2 |  |  |
| О | 1 |  |  |
| , | 1 |  |  |
| Н | 1 |  |  |
| Е | 1 |  |  |
| Т | 1 |  |  |

Всего 58 символов