Частное профессиональное образовательное учреждение
«АНАПСКИЙ ИНДУСТРИАЛЬНЫЙ ТЕХНИКУМ»

КОДИРОВАНИЕ ТЕКСТОВ

Лекцию подготовил:
Омельченко И. Н.

ЧТО ТАКОЕ КОДИРОВАНИЕ ТЕКСТА?

- Компьютер хранит текст как числа.
- Каждый символ имеет свой код.
- Кодировка — это таблица соответствия символов и чисел.
- Один и тот же текст в разных кодировках может выглядеть по-разному

ASCII

- Первая массовая кодировка (1960-е).
- 7 бит = 128 символов.
- Содержит латиницу, цифры, знаки

Ограничения ASCII

- Содержит только английский алфавит.
- Невозможность представления национальных языков.
- Появление «кракозябр» при неправильном открытии файла.

ОДНОБАЙТНЫЕ КОДИРОВКИ

Однобайтовые кодировки — это системы кодирования символов, в которых для представления каждого символа используется 1 байт (8 бит), что позволяет закодировать до 256 различных символов.

Они являются расширением стандартного кода ASCII, где первые 128 символов (0-127) общие, а оставшиеся 128 (128-255) используются для национальных алфавитов и других символов в зависимости от конкретной кодировки.

ОДНОБАЙТНЫЕ КОДИРОВКИ

Примеры русских кодировок:

Windows-1251

KOI8-R

ISO-8859-5

Недостаток: несовместимы друг с другом.

UNICODE

Поддерживает все письменности мира.

Каждому символу присвоен код: U+XXXX.

Примеры:

A → U+0041 А (русская) → U+0410 😊 → U+1F60A

Эмодзи — это графические символы, используемые для передачи эмоций, действий, предметов, идей и других смыслов в цифровой переписке.

UNICODE ≠ UTF

Unicode — таблица символов.
UTF — способ хранения.

Основные форматы UTF:
UTF-8
UTF-16
UTF-32

UTF-8 — СОВРЕМЕННЫЙ СТАНДАРТ

- Использует 1–4 байта.
- Совместим с ASCII.
- Поддерживает все символы Unicode.
- №1 в интернете, HTML, Linux

СКОЛЬКО БАЙТ ЗАНИМАЕТ СИМВОЛ В UTF-8?

Латиница (A, Z, !) → 1 байт

Кириллица (Я, Б, Л) → 2 байта

Эмодзи 😊 🤓 🤖 → 4 байта

Пример расчёта объёма текста

Строка: АЯ🤓!

A = 1 байт - Я = 2 байта - 🤓 = 4 байта - ! = 1 байт

Итого: 8 байт

ИНФОРМАЦИОННЫЙ ОБЪЁМ

Формула: **I = K × i**

K — количество символов.
i — вес символа (байты/биты).

Объём зависит от кодировки.

ОШИБКИ КОДИРОВОК

Возникают, когда файл открыт «в другой кодировке».

Пример:

Привет → Đ Ñ Đ¸Đ²ĐµÑ

Причина:

байты интерпретируются неверно