LOCALIZATION ENGINEERING • CLASS 1 • MODULE 2

Respecting Global Alphabets & Accents

Understanding character encodings, UTF-8 standards, and techniques to prevent character corruption on screen.

ENCODING: UTF-8 (UNICODE)

LESSON 2.1 • FOUNDATIONS

How Computers Store Language

Binary Streams

Computers store raw text solely as numeric sequences of zeroes and ones.

Character Maps

Look-up tables mapping numeric code points directly to human glyphs.

Code Points

Unique numeric values assigned to every letter, diacritic, and symbol.

LESSON 2.1 • ENCODING HISTORY

Evolution of Digital Character Standards

1. **ASCII Standard (7-Bit):** Restricted to 128 basic English characters and control symbols.
2. **Regional Encodings (8-Bit):** ANSI, ISO-8859, and Windows-1252 introduced regional code pages.
3. **Unicode Integration:** A single universal standard mapping every global character set.

Global Web Standard

98%

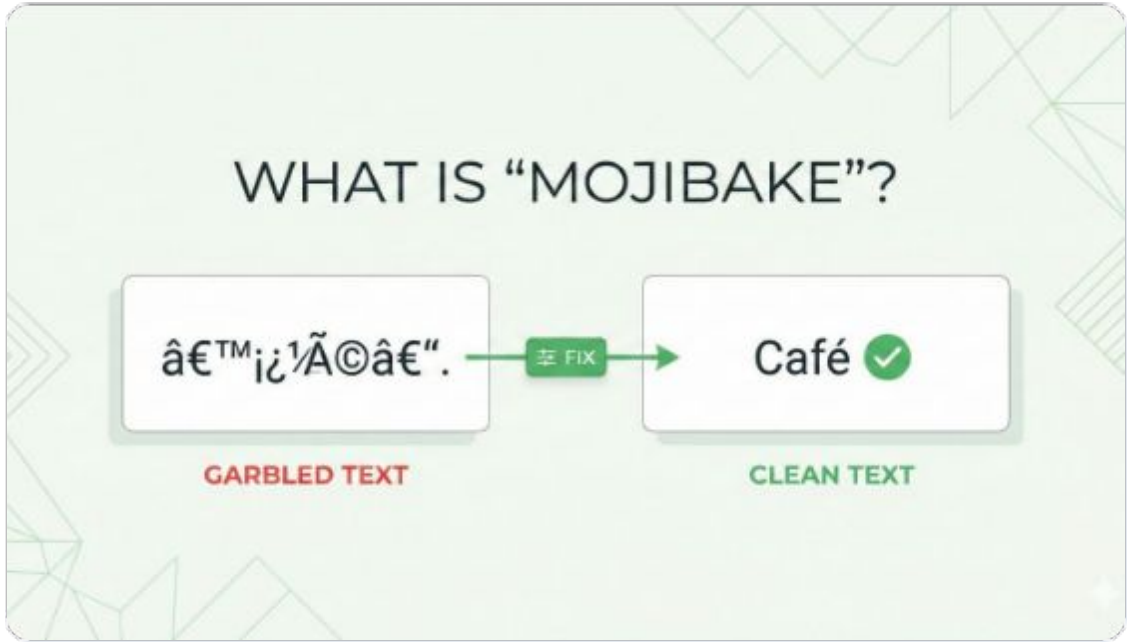
Share of modern web pages and software applications built on UTF-8 encoding

LESSON 2.2 • CHARACTER CORRUPTION

Anatomy of Mojibake

Mojibake occurs when a file written in one encoding profile is interpreted using a different character map.

- **Corrupted Display: Zamoœæ instead of Zamość.**
- **User Impact: Erodes trust and creates unreadable screens.**



LESSON 2.2 • RESTORATION MECHANICS

The 3-Step Encoding Rescue

1. Inspect Encoding

Check VS Code status bar for current active profile.

2. Reopen with Encoding

Select original source encoding to restore characters.

3. Save with UTF-8

Convert and save file profile into UTF-8 standard.

LESSON 2.2 • BATCH OPERATIONS

Folder-Level UTF-8 Remediation

iconv

Command-Line Batch Conversion

Workspace

VS Code Multi-File Extensions

0 Hits

Target Regex Mojibake Audit

HANDS-ON LAB • EXECUTION

Lab 2: The Corrupted Text Rescue

1. **Identify Corruption:** Isolate damaged Polish diacritics in UI files.
2. **Re-Open Profile:** Switch profile to Windows-1252 to reveal correct characters.
3. **Batch Convert:** Convert entire dropdown folder assets to UTF-8 standard.

QUALITY ASSURANCE • AUDIT

Encoding Verification Checklist

- ✅ **Status Bar Check: Confirm UTF-8 explicitly displayed for all target files.**
- ✅ **Diacritic Verification: Check non-English accents (ä, ö, ü, ñ, ść, č) on screen.**
- ✅ **Regex Clean Sweep: Execute [œæ‰âã‡] pattern search across project.**

MODULE 2 COMPLETE

Next: Software & Human Logic

With character encodings protected, Module 3 explores key-value structures, variable placeholders, and non-programmer regex.

NEXT: MODULE 3 — KEY-VALUE LOGIC & SYNTAX

Image Sources



https://texttooldb.com/wp-content/uploads/2025/12/what-is-mojibake_texttooldb.jpg

Source: texttooldb.com