UNSTRUCTURED: Cleaning Test — Page ONE

This document is deliberately MESSY. It contains extra spaces, tabs, and inconsistent casing. Email: TEST_User+spam@Example.com — url: HTTP://EXAMPLE.com/Path?Q=1.

Headings? maybe. ——— Section A ——— Some sentences... have odd punctuation !!! And wrong—dash—usage — like this.

Line-break hyphenation ex- ample should be merged across lines. Also multiple spaces should become one.

Smart quotes: “curly” and ‘single’ vs "straight" quotes. Non-breaking spaces here. Unicode oddities: café, naïve, coöperate, emoji: □□ — ensure normalization (NFC).

Lists, Markup & Fragments (p.2)

Shopping list (messy bullets): * apples - Bananas • oranges 1) milk 2.) bread 3 . eggs

Markdown-ish: ### Title??? > A blockquote that may or may not be real. `inline code` and ```
multiline code block ```

HTML snippet: Hello world!   one   two

Random caps and spacing: ThIs Is a TeST Of NorMALIZation. Odd punctuation spacing :like
this ,and this .And this !?

Numbers, Dates & Entities (03)

Dates (varied): 2025-08-18; 18/08/2025; Aug 18th, 2025; Monday 18th of August, '25. Currencies: $1,234.50; € 999 , 95 ; £0.99 ; JPY 1 000 ; INR₹1000 ; CHF 1'234.56. Phone: +1 (415) 555 - 1234 ; +49-30- 123456 ; (020) 7946   0018.

OCR-ish artifacts: The quick brown f0x jumps 0ver the lazy d0g. rn and m can look similar: mod ern -> modern. Broken words at end of line should be de- hyphenated where appropriate; BUT re- spect true hyphenated-terms like state-of-the-art.

Weird control chars: form feed here; and a tab there; and non-breaking spaces.

Simulated Table & Misc (IV)

CSV-ish table (ragged): id , name , amount , date 1, Alice, 12.00 ,2025/01/02 2 ,Bob, 7.5, 2025-1-3 003, "Carol D." , 1000 , 18-08-2025

URLs & tracking params: https://example.com/search?q=cleaning https://example.com/article?utm_source=Email&utm_medium=CAMPAIGN&utm_campaign=Summer mailto: support@example.org , dev+ops@example.org

Misc: — em dashes — – en dashes – - hyphens - — — doubled — — dashes — — Multiple blank lines follow:

 END OF TEST DOCUMENT.