Test Design – Preprocessing

Estado	Lista

Goals

- Detect `@mentions`, `#hashtags`, `urls`, `emojis`.
- Normalize (lowercase + collapse spaces). Integrative-Task1-2025-2_CyED3

Functions Under Test

`extract_all`, `extract_mentions`, `extract_hashtags`, `extract_urls`, `extract_emojis`, `normalize_text`, `tokenize`. (As defined in your `regexRules.py`).
regexRules

Test Matrix

ID	Scenario	Input	Expected (key checks)	Notes
P1	Basic detection	`Hey @User see https://uni.edu #Icesi 😄`	mentions = `["@User"]`; urls contains `https://uni.edu`; hashtags = `["#Icesi"]`; emojis contains `😄`	Case preserved in raw lists, but normalized lowercases
P2	Normalization	`" HELLO \n WORLD\t "`	`normalized == "hello world"`	Collapse whitespace
P3	Tokenization keeps special tokens	`"go http://a.com #wow @me"`	`tokens` must include the URL, hashtag, and mention as separate tokens	Uses `TOKEN_RE` groups regexRules
P4	Mixed-language + emoji	`"oye bro, you ROCK 😄 #Clase"`	tokens present; emojis detected; normalized lowercases	

ID	Scenario	Input	Expected (key checks)	Notes
P5	Edge: empty	`""`	all lists empty; `normalized == ""`	