Module Design - Post Preprocessing (Regular Expressions)

Materia	Discretas 3
Estado	Lista

File/Module

`src/moderation/` *`regexRules.py`*

Public API

`extract_all(text: str) → dict`

Input

raw post string

Output (dict):

```
{
  "mentions": ["@alice", "..."],
  "hashtags": ["#icesi", "..."],
  "urls": ["https://...","www..."],
  "emojis": ["😄","..."],
  "normalized": "lowercased single-spaced string",
  "tokens": ["lowercased","tokens","..."]   // space-split tokens
}
```

Notes

- Normalize: lowercase + collapse whitespace.
- Keep URLs/hashtags/mentions in tokens (they are needed later).
- No side effects; pure function.

Optional helpers

- `detect_mentions(text: str) → list[str]`
- `detect_hashtags(text: str) → list[str]`
- `detect_links(text: str) → list[str]`
- `detect_emojis(text: str) → list[str]`
- `normalize_text(text: str) → str`

Example

```
extract_all("Hey @User visit https://uni.edu #Icesi 😄")
# → {"mentions":["@user"],"hashtags":["#icesi"],"urls":["https://uni.edu"],
#     "emojis":["😄"],"normalized":"hey @user visit https://uni.edu #icesi 😄",
#     "tokens":["hey","@user","visit","https://uni.edu","#icesi","😄"]}
```