REGEX REFERENCE SHEET

How to Write Regex from Scratch?

Steps for Creating Regex

1. Observe 3–6 real examples and 1–2 counterexamples.
2. State rules in words (what must be present? what can vary?).
3. Pick blocks (character classes, anchors, quantifiers).
4. Glue blocks into a first draft.
5. Test on examples; note misses/over-matches.
6. Refine (add boundaries \b, alternatives (a|b), ranges {m,n}, escapes \.).

Check Reference Sheet II (PDF) on NILE for Symbols and their meanings

Real Examples

Steps for Creating Email

1. Observe
    - Match - jane.doe@uni.edu, help@service.com, info@mail.co.uk
    - No match - admin@mail, bad@@x.com, a@b.c
2. State Rules in words
    - Username = letters/digits/underscore **with** optional dots or dashes.
    - @ symbol.
    - Domain parts separated by dots.
    - Last part (TLD) is **letters, at least 2 long**.
3. Pick Blocks
    - Username: [\w\.-]+ (\w = letters/digits/underscore; include . and -)
    - @: literal @

- Host: [\w\.-]+
- Dot: \. (escaped)
- TLD: [A-Za-z]{2,} (**"at least 2"** → {2,})

4. Glue Blocks
    - **v1:** \w+@\w+\.\w+ (too strict; misses dots/dashes, short TLDs)
    - **v2:** [\w\.-]+@[\w\.-]+\.[A-Za-z]{2,} - match
5. Final Pattern
    - [\w\.-]+@[\w\.-]+\.[A-Za-z]{2,}

Steps for Year (four digits or modern years)

1. Observe
    - Match - 1998, 2025, 2001
    - No match - 98, 20251
2. State Rules in Words
    - Exactly four digits; optionally restrict to 19xx or 20xx.
3. Pick Blocks & Glue
    - **Any 4 digits**: `\b\d{4}\b` ({4} means exactly 4)
    - **Modern years**: `\b(19|20)\d{2}\b`
4. Final Pattern
    - **`\b\d{4}\b`**
    - **`\b(19|20)\d{2}\b`**

Steps for Date DD/MM/YYYY (strict digits)

1. Observe
    - Match - 03/10/2025, 31/12/2000

- No match - 32/01/2020, 3/1/2020

2. State Rules in Words
    - Day 01–31, slash, Month 01–12, slash, 4 digits.
3. Pick Blocks & Glue
    - **Day**: (0[1-9]|[12]\d|3[01])
    - **Month**: (0[1-9]|1[0-2])
    - **Year**: \d{4}
4. Final Pattern
    - \b(0[1-9]|[12]\d|3[01])/(0[1-9]|1[0-2])/\d{4}\b

Steps for Twitter handle (avoid emails)

1. Observe
    - Match - @nlp_lab, text @User123 end
    - No match - name@domain.com (don't capture the @ in emails)
2. State Rules in Words
    - Start of line **or** whitespace, then @, then 1–15 letters/digits/underscore, then boundary.
3. Pick Blocks & Glue
    - Start/space: (?:(?<=\s)|^) (lookbehind for whitespace **or** start)
    - Handle: @[A-Za-z0-9_]{1,15}\b
4. Final Pattern
    - (?:(?<=\s)|^)@[A-Za-z0-9_]{1,15}\b

TIPS: USE BOUNDARIES

- Use \b to avoid mid-word hits: \bcat\b doesn't match *educate*.

- Use lookbehind to ensure context without consuming it: (?<=@)\w+ gets the part **after** @ only.
- Use (^|\s) (or (?:(?<=\s)|^)) to force “start or whitespace” when needed.