Crawler(F={seed URLs}) F = frontier queue
queue

① URLs = dequeue(F, batch) → F does not rearrange ...

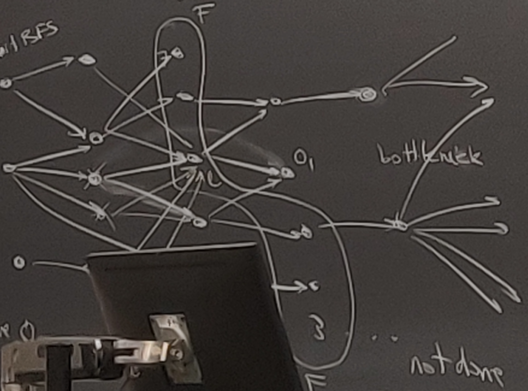
② for each e ∈ URL

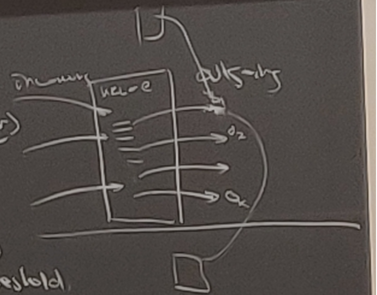
Ⓐ GET(e) – http head call ⇒ meta info
– wget call ⇒ html source
– polite: DELAY (40s), Robots.txt (server)

? { CLEAN(e)
EXTRACT outgoing links(e) ⇒ $\{o_1^e, o_2^e, \dots o_k^e\}$
ANALYZE CONTENT(e) – relevance (topic) ≥ threshold

– diversity/novelty
– near-duplicate (source ≅ source' ∈ INDEX)
– pagerank (links) ≅ importance from links
– domain ≅ importance due to site/author.

Ⓑ ?? maybe
ignore (e, source)
continue FOR (next e)





(2C) store(p_i, source) ⇒ ES (ID)

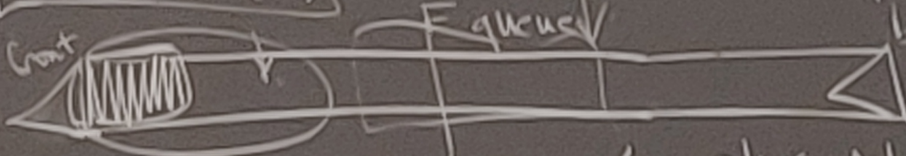
foreach $o_i \in$ outgoinglinks(p)

o_i = cannonical(o_i)

check(o_i) = new?

	BFS wave	incoming → o_i	URL(o) text vs keywords	Security	domain
score	K	count when q=F	BM25	filter	set filter / swap

ENQUEUE(o_i)



default FIFO

- mechanisms (operations, data structures)
- local resorting/rearrangement criteria