Simple Crawler Thread

```
procedure CrawlerThread(frontier)
    while not frontier.done() do
        website ← frontier.nextSite()
        url ← website.nextURL()
        if website.permitsCrawl(url) then
            text ← retrieveURL(url)
            storeDocument(url, text)
            for each url in parse(text) do
                frontier.addURL(url)
            end for
        end if
        frontier.releaseSite(website)
    end while
end procedure
```

How to Include the Shared data Structures

- Duplicate Post Map
- Politeness Timer
- File Counter
- Post Counter

Just building directory of Post JSON Files

Protect with a Mutex.

Simple Crawler Thread

How to limit?

```
procedure CrawlerThread(frontier)
    while not frontier.done() do   and not at limit
        website ← frontier.nextSite()
        url ← website.nextURL()
        if website.permitsCrawl(url) then   and check Politeness Timer
            text ← retrieveURL(url)
            storeDocument(url, text)   file counter logic here
            for each url in parse(text) do
                frontier.addURL(url)   If not in Post map
            end for                    Post Counter ++
        end if
        frontier.releaseSite(website)
```

end while
end procedure

Storage Requirements

1 post / Row

10 MB files

at least 500 MB
= 50 Post files

What details to include in JSON for post?

Post, Linked Website Titles, Linked Posts,
Post Comments, Post titles, tags
Post upvotes, Comment upvotes
usernames, "trendy" / "new", Subreddit

All in JSON

How to use these to rank?