Idio and Content

Idio ingests web pages (and a variety of other web-based resources) for numerous reasons, the goal of which is to facilitate the:

- Generation of abstracts/summaries for those resources
- Analysing resources in aggregate to identify “gaps” in coverage of interesting topics
- Grouping of resources based on metadata

Unfortunately, the web is a messy place, and this content comes in all shapes and sizes. Ingesting the content and extracting accurate information from it in a reliable, deterministic and performant way is not a simple task.

Some of the problems faced when ingesting web-based content are things like how to:

- Extract the title and body of a resource
- Extract metadata from a resource, potentially including but not limited to:
    - link/meta tags
    - JSON-LD
    - Internal metadata (in the case of PDFs, images, videos, etc)
- Extract images from a content item
    - Both body images or meta/link tag image references
- Avoid ingesting the same page with different parameters (that don’t alter the page)
- Reliably canonicalise a URL
- Orchestrate requests to remote servers
    - Authentication, cookie retention and rate limiting all differing between sites
- Save (and index) a content item
    - What are the relevant and important attributes of an online resource?
- Avoid scraping the same resource again and again
- Explicitly scrape the same resource again to update it
- Handle responses of remote servers
    - Following redirects, handling response codes, timeouts, network errors etc
- Cache responses from the server
- Ingest resources other than article-like HTML pages, i.e.:
    - PDFs
    - Videos
    - Images
    - HTML Data Tables

Your task is pick one or more of these problems that you deem fun or interesting and, taking a URL as your input, produce some code as you set about solving them. Solutions are accepted in any language and format, but please make sure you're focusing on the following key points:

- We must be able to execute everything to end-to-end in order to test it
- You're free to use any libraries and tools, but please include explanation as to why you've chosen those specific ones
- Try to keep it 'in-house' as much as possible, i.e. avoid using external APIs during the process
- Installation is provided in format of a Makefile or documentation
- There is at least a short description provided of which problems you've chosen (and maybe why) and what you've accomplished
- If you've also written tests, please add running instructions for them

Some tools and libraries you may find useful:

- http://corpus.tools/wiki/Justext
- https://lxml.de/
- https://www.crummy.com/software/BeautifulSoup/bs4/doc/#
- https://github.com/misja/python-boilerpipe
- https://github.com/grangier/python-goose
- That mozilla thing Vic mentioned [LINK]

Here are a few example URLs to get you started. However if you're feeling adventurous and would like to showcase your code against other content-heavy sites or specific non-HTML resources, then please include the URLs with your submission!

Example 1

URL: http://www.crainscleveland.com/node/688181
Expected Title: *Cleveland Clinic sets opening for new Lakewood Family Health Center*
Expected Body:
Cleveland Clinic is celebrating the opening of its new Lakewood Family […] began transitioning inpatient services out of the hospital in 2016.

Example 2

URL:
https://www.theguardian.com/politics/2018/aug/19/brexit-tory-mps-warn-of-entryism-threat-from-leave-eu-supporters
Expected Title: *Brexit: Tory MPs warn of entryism threat from Leave.EU supporters*
Expected Body:
Pro-Brexit group urges supporters to join to back Boris Johnson or Jacob Rees-Mogg for leader […] can go forward to the final ballot among the party's members, the Telegraph reported on Sunday.
Expected Images:


Example 3

URL:
https://etfs.wisdomtree.eu/institutional/uk/en-gb/products/product/etfs-agriculture-agap-lse
Expected Title: *ETFS Agriculture*
Expected Body:
ETFS Agriculture (AGAP) is designed to enable investors to gain an exposure to a total return investment in a basket of agriculture commodity futures contracts by tracking the Bloomberg Agriculture SubIndex (the "Index") and […] found in the Collateral section of the ETF Securities website (www.etfsecurities.com).

Example 4

URL: https://learning.econsultancy.com/module/2
Authentication: you can register an account for yourself
Expected Title: *Planning Digital Marketing*
Expected Body: *Learn the 5 golden rules of effective planning and useful frameworks to help you day to day. Get your head around SMART objective setting and become familiar with the concept of customer journeys.*

Example 5

URL:
https://www.lazardassetmanagement.com/docs/-sp1-/68744/TheLinkBetweenESGAndFinancial_LazardPerspectives_en.pdf
Expected Title:*The Link between ESG and Financial Productivity*
Expected Body: *While it is widely recognised that environmental, social, and governance (ESG) issues can affect a company's valuation and financial […] puts ESG analysis at its core, offers forward-looking insights that could potentially enhance long-term returns for investors.*