NewsLight

The Technology Requirements

The Workflow

Identifying and Categorizing Info Sources	Identifying Foundational Sites	Identifying Additional Info Sources	Develop Information Acquisition Capabilities	Apply Categorization
Collecting Information	Crawl Sites	Create Crawl Record	Compare	Create Change Record
Generating Content	Apply Rules	Send to LLM	Create Content Record	Send to Primus Engine
Organizing Content for Publication	Organizing Content	Notifying Editor	Editing Content	Publishing Decision
Publishing and Sharing Content	Publish to Community	User Signals	User Interests	User Content

Identifying and Categorizing Information Sources

Identifying Foundational Sites

Using tools, develop a list of websites with information that relate to members of a community

Store that list in a database

Identify Additional Info Sources

Identify other potential sources of information (Reddit, Facebook Groups, X/Twitter)

Store sources in database

Develop Acquisition Capabilities

As necessary, develop capabilities to acquire information

Capabilities include live transcription based on videos, reading PDFs, image recognition

Apply Categorization

Each information source is categorized based on geography and interest

Collecting Information

Crawl Site

Crawl site 1x per week

If a page has new content, scrape page every day

If a page has new content at least once every two days, scrape it every hour

Challenge: Develop a crawler that is not blocked

Create Crawl Record

Any time a website is crawled or web page is scraped, a database record is created.

Database record should include text, images, videos, and pdf files

Assuming 500 sites per community X 1GB per site = 500GB per week, or $2 per week, or $100 per year

Compare

New webpage is compared to the old webpage to identify differences

This could include new text, images, videos, files. It could also include new pages.

Create Change Record

Any changes in content are stored in a database

Generating Content

Apply Rules

For new content, specific rules will be applied in order to determine if it is worthy of publishing.

Rule # 1: New text has to have at least 25 characters

Rule # 2: New images have to be associated with text (in order to provide context)

Send to LLM

The new content that survives the rules will be sent to an LLM

The LLM will be provided prompt instructions to go along with the content. Ideally, the instructions would be permanent and used every time.

The LLM will return content

Challenge: Getting the LLM to complete the full task, consistently

Create Content Record

The content returned will be stored in a database.

This database will have all LLM generated content

Send to Primus Engine

Content will be sent to the Primus Engine for processing

Organizing Content for Publication

Organizing Content

The Primus Engine organizes content based on origin source.

It classifies content into 7 geographic layers and numerous different types of interest.

It shows each queue for each community and geographic layer.

Notifying Editor

Each queue has a list of content ready for publishing.

As new content is ready for publishing, an alert is sent to the Editor

Editing Content (optional)

The editor can edit any content and/or the classifications assigned in the Primus Engine.

The content is automatically saved when changes are made.

Publishing Decision

Once the editor has finalized the content, they have two choices, scheduled post or publish now.

The default is scheduled post. For a scheduled post, a time is assigned for when it will be published, ideally different than any other content, in order to ensure a consistent flow of content

Publish now is self-explanatory

Publishing and Sharing Content

Publish to Community

Once a decision is made to publish, the content is posted to the appropriate community group

User Signals

The Perceptor Engine determines whether content is considered, Hot.

Hot content automatically is shown to users first, unless they have already been exposed to it.

If a local story is clicked on by a certain number of users, it is considered hot.

User Interests

Over time, as users interact with more content, they will also be shown content that interests them

This is the second layer of classification.

User Content

As a fast follow, users will be given the opportunity to submit content for consideration

That content will go through the same process as that which is automatically identified through the crawler.