Raag Patel
Stuart Wurtman
Adrian Telan

Project Proposal: Automated Research Tool for Online Investigations

As information moves online, people across many fields rely on internet research. We are building a core research tool with an extensible interface. The core tool will be a general-purpose bookmarking engine with meta-tagging and an ability to add extension packages and run scripts for niche research needs. We are also building the first extension package, targeted towards professional and citizen journalists gathering evidence for internet-based investigations.

We set out to help journalists working on internet investigations. These journalists face many challenges, which can make their work quite tedious. Their problems include deleted posts, duplicate posts, altered media, and surfacing information. The extension we develop will automate many of the procedures that are done to archive information, find originals of duplicate posts, find altered media, and find new information.

We chose to separate our tool into a core product and an extension for our niche because we believe the core bookmarking engine is valuable as a standalone product to many internet users. Running without extensions, the core product will log a URL and capture metadata associated with it, and save this information into a project file. Metadata will include a description, #tags, and the URL.

We plan to explore development of this tool as a browser extension. Since most internet research is done in a browser, a browser extension allows a tight integration into the research process. The application will store research data, but the location of this storage is to be determined. A cloud service with collaboration is one possibility.

Surfacing information is an important part of the problem for internet journalists. These researchers need to sift through a large amount of content, remove duplicates, find original posts, and they may also want to be informed of related information that they wouldn't have even thought to look for. For example, if there was a post duplicated many times on a social media platform, and commenters frequently referenced a second post, there is an opportunity to surface that second post for the researcher.

Much of the research content we are interested in is found on social media, and each social media service has its own access limitations. We plan to start by solving our problems for Twitter, since it is a very open social media service, then move on to integrate additional social media services.

Research involves iterating through large amounts of information. The problem with research is that people are not suited to iterative tasks in the same way computers are (Enago.com). We see a place for tools to make research easier for people and automate many of the tasks that they are doing manually.

Every internet user could benefit from having a tool to track their internet research and reduce their tab count. The benefit of our research tool is the ability to extend the tool so that every research niche can automate the repetitive parts of their research process.

User Stories for Core Product

As a user,
I want to save research URLs,
so that I can reduce browser clutter.

As a user,
I want to search through my saved URLs,
so that I can find research I have saved.

As a user,
I want to save research URLs by project,
so that I have portable access to my research.

As a user,
I want to add #tags to my saved research,
So that I can make ad hoc classifications of the research.

As a user,
I want to view all of my research for a project in one place,
So that I can understand my research so far.

As a user,
I want to know the date that a piece of research was saved,
So that I can create a timeline of research.

As a user,
I want to be able to group research by domain name,
So that I can understand where research is coming from.

As a user,
I want the option to create a text index of a saved piece of research,
So that I can search through the entire contents of all my saved research from one place.

As a developer,
I want an interface to develop custom extension,
So that I can satisfy the requirements of my research niche.

As a developer,
I want to execute scripts,
So that repetitive actions can be automated while saving research.

User Stories for Internet Investigative Journalism Extension

As a user,
I want to automatically save/archive a copy of the webpage and all media located at each saved URL,
So that I have a copy in case the post is deleted later.

As a user,
I want to search social media and the web for duplicates of a post or page,
So that I can identify the original source.

As a user,
I want to calculate a percent score that estimates the likelihood that a photo has been digitally manipulated by analyzing image noise variance,
So that I can make decisions on whether to further assess the legitimacy of a post.

As a user,
I want to choose a Twitter post I am interested in, and search for secondary posts that are related by on-demand crawling Twitter,
So that I can find areas to investigate that are related to my current research space.

As a user,
I want to aggregate all unique media in one place,
So that I can visualize the “on-the-ground” picture of my investigation.

The steps to executing the project

1. Choose a problem to work on
2. Find a possible solution to the problem
3. Develop user stories
4. Match technologies/solutions to user stories
5. Test technologies and prototype solutions
6. Build core product.
    a. Browser extension
    b. Database to store data research data
    c. Interface to extend product with scripts
        i. Determine how to make scripts run on the product … python?
7. Create an extension for internet investigative journalists
    a. Scripts to push pages/posts to internet archive and link the archived copy to the saved research
    b. Scripts to identify altered media
    c. Scripts to filter duplicate post (finding earliest sources)
    d. Scripts to surface related information
8. Test product, iterate on development.
9. Test product against manual research processes and measure performance improvement.

Detailed Individual Contributions

Raag

1. Established a GitHub repository
2. Edited Project Proposal Document

Stuart

1. Wrote 15 user stories
2. Finalized project introduction and executive summary
3. Finalized rough project plan
4. Researched applications similar to the project

Adrian

1. Wrote a draft of the project introduction and executive summary
2. Wrote a draft of the rough project plan
3. Researched whether the project's research function would save on research time

Works Cited

"Artificial Intelligence in Research and Publishing." Enago.com, Enago Academy, 25 Aug. 2020, www.enago.com/academy/artificial-intelligence-research-publishing/#:~:text=Thus%2C%20AI%20can%20not%20only%20help%20expedite%20scientific,techniques%20from%20generating%20a%20hypothesis%20to%20conducting%20experiments.