SciPop

Delaney Glass[a], Qiyao Liu[b], Ishika Ray[c]

[a]Department of Anthropology, [b]Department of Applied Mathematics, [c]Department of Psychology, University of Washington

Project Overview

SciPop is a **software tool** which can **identify news articles** which discuss academic research that the user is interested in delving into. Users simply **upload a CSV file** of research article names, DOIs, or author names, and **SciPop creates a CSV output** which compiles the **most relevant Google News search results** related to the users' topic of interest.

Who can use SciPop?

SciPop's primary objective is to make web scraping easy and efficient for **researchers who do not code**.

Our prospective users may include:

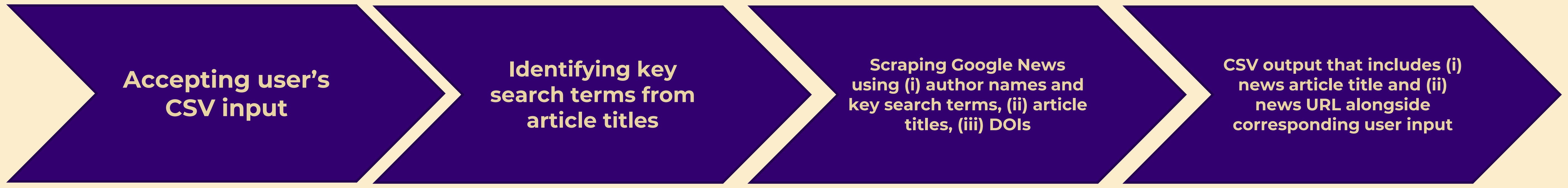
- **Researchers** interested in **reviewing media coverage** of broad research topics
- **Students** who are studying **how specific topics or academic articles have been discussed in news media and popular culture**
- **Non-academic users** who are curious to read **about scientific literature that is indexed on Google News**

Data Preparation

Data Source: We ask users to upload a CSV file which contains 2-3 columns, labeled as follows:
"Author_Name": Contains the names of authors of academic articles. Users may input the full name of the first author, or the names of all authors as one string.
"Article_Title": Contains a string of the specific academic article titles which the user wants to narrow their search to.
"Article_DOI": Contains the strings of Digital Object Identifiers (or DOIs) associated with the article titles or author names.

Files that are not uploaded in this format are considered invalid input by the tool.

Tool Design



Accepting user's CSV input

Identifying key search terms from article titles

Scraping Google News using (i) author names and key search terms, (ii) article titles, (iii) DOIs

CSV output that includes (i) news article title and (ii) news URL alongside corresponding user input

Scraping process: We created 3 scraping functions to identify news articles which: (i) contain the academic article title, (ii) contain the DOIs in the article body, (iii) contain the author name(s) AND at least one of our detected key search terms.

Python Packages Used

Package Name	Purpose
nltk	Identifying keywords from input article titles
pandas	Data manipulation
pygooglenews	Web scraping [selected over beautifulsoup or SerpAPI because it is tailored to Google News, and is relatively easy to learn]
streamlit	Constructing user interface

Milestones hit thus far: Created module containing all relevant scraping functions, created a user-friendly interface. Our code is PEP8-compliant and has considerable unit-test coverage.

What can you expect in Version 2.0?

The current version of SciPop depends heavily on (i) the flexibility and efficiency of our scraping tool, and (ii) users uploading their CSV files with headers that SciPop dictates. In future versions of SciPop, we plan to:

- ***Use a more up-to-date scraping tool***: PyGoogleNews limits our choice of package dependencies, which can lead to environment errors while running SciPop. We plan to resolve this in our next version by using an alternative web scraper.
- ***Autodetect column names from user input***: Using built-in Python modules (such as DiffLib) can help us detect relevant headers from an input CSV file without requiring users to input their data in a specific format. We plan to implement this change so that users may simply download a CSV file from research databases (e.g. Web of Knowledge, PsycInfo) and drag-and-drop it into SciPop.

To leave us any feedback., please visit our GitHub repository: