Environmental Modelling & Software 101 (2018) 249–255

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

A novel search algorithm for quantifying news media coverage as a measure of environmental issue salience

Nicholas A. Roby [a], Patricia Gonzales [b, c], Kimberly J. Quesnel [b, c], Newsha K. Ajami [a, b, *]

[a] Woods Institute for the Environment, 473 Via Ortega, Room 218B, Stanford University, Stanford, CA 94305, USA
[b] ReNUWIt Engineering Research Center 473 Via Ortega, Room 117, Stanford University, Stanford, CA 94305, USA
[c] Department of Civil and Environmental Engineering, 473 Via Ortega, Room 314, Stanford University, Stanford, CA 94305, USA

ARTICLE INFO

Article history:
Received 10 April 2017
Received in revised form
19 December 2017
Accepted 20 December 2017

Keywords:
News media coverage
Python
Google Custom Search Engine API
Search tool

ABSTRACT

News media plays an important role in shaping public opinions and attitudes about the environment. Tracking and analyzing media coverage can provide insight into public exposure to narratives that impact resource consumption, environmental behavior, and emergency response, which can help to inform model development or provide additional model inputs. This paper presents *Articulate*, an open-source, flexible tool for discovering, compiling, and quantifying newspaper coverage on a user-specified topic. *Articulate* is written in Python and interfaces with Google Custom Search Engine API. We demonstrate the tool's application and validate its performance on two case studies of news media coverage in the New York Times about drought in California and flooding in Houston, Texas in recent years. Our results show that *Articulate* can generate data similar to or better than proprietary databases. Thus, *Articulate* can help researchers and environmental managers gain important insights to better understand and quantify changing socio-environmental dynamics.

© 2018 Elsevier Ltd. All rights reserved.

Software availability

Name of software: Articulate
Developers: Nicholas Roby and Newsha Ajami
Contact address: 473 Via Ortega, Room 204, Stanford, CA 94306
Telephone: 650-724-8162 Email address: nickroby12@gmail.com and newsha@stanford.edu
Year first available: 2017
Hardware required: Personal computer, internet access
Software required: Python 2.7
Availability: https://github.com/Stanford-Urban-Water-Policy-Innovation/Articulate
Cost: $0
Program language: Python
Program size: 12 KB

1. Introduction

There is increasing recognition that human behavior plays an integral role in environmental planning decisions (Lund, 2015; Sovacool et al., 2015). Behavior has thus emerged as a central feature of many new environmental modeling methods and applications; for example, in recent years scientists and engineers have attempted to model human actions and reactions within hydrologic systems (Garcia et al., 2016; Gonzales and Ajami, 2017; Noël and Cai, 2017), when predicting energy demand (Ma et al., 2009; Wilson and Dowlatabadi, 2007), as inputs into land-use change projections (Lauf et al., 2012), and for developing emergency response scenarios during extreme events such as floods and wildfires (Giordano et al., 2017; Nara et al., 2017). Assumptions about human behavior can have significant effects on the accuracy and results of models that aim to couple human and natural systems (Noël and Cai, 2017; Sun et al., 2016), and as a result, many different approaches for modeling human-environmental interactions have emerged. One method for understanding human behavior is to measure issue salience—public awareness and education are important drivers of environmental resource consumption and conservation (Stern, 1976), emergency response behavior (Du et al., 2017; Mccaffrey et al., 2017), and attitudes and beliefs (King et al., 2017).

Both social media (Du et al., 2017) and news media (Treuer et al., 2017; Troy et al., 2015) have been recognized as vehicles for measuring community exposure and sensitivity to environmental

* Corresponding author. Woods Institute for the Environment, 473 Via Ortega, Room 218B, Stanford University, Stanford, CA 94305, USA.
E-mail address: newsha@stanford.edu (N.K. Ajami).

https://doi.org/10.1016/j.envsoft.2017.12.012
1364-8152/© 2018 Elsevier Ltd. All rights reserved.

topics and events. In this paper, we focus on news media, one important method of influence (Ball-Rokeach and DeFleur, 1976; Boykoff and Boykoff, 2007; Sampei and Aoyagi-Usui, 2009) that has been relatively unexplored in the modeling of environmental scenarios. Quantifying and analyzing coverage of environmental topics such as droughts, floods, wildfire, clean energy, climate change, or other themes can provide deep insight into awareness levels of the public, which can then impact decision-making. Analyzing news media coverage can also be used to retroactively evaluate the most salient events surrounding a topic (Boykoff and Boykoff, 2007) as environmental news coverage is often driven by specific events (Bolsen, 2011) which are believed to reflect periods of public awareness. This approach of creating news media time series and analyzing coverage has been used in previous research to provide insight into public exposure to environmental topics (Akerlof et al., 2012; Crow et al., 2016; Ruiz Sinoga and León Gross, 2013; Treuer et al., 2017). Quantitatively, researchers have examined news media to develop sustainability indicators (Rivera et al., 2014), validate flooding models (Smith et al., 2012), and as inputs into water demand models (Quesnel and Ajami, 2017). These types of multidisciplinary modeling approaches are likely to be used more frequently in the future.

Current tools for evaluating and quantifying news media coverage include proprietary databases such as ProQuest and LexisNexis (Bolsen, 2011; Wei et al., 2015). These database tools, fit with a graphical user interface (GUI), allow the user to query from a selection of news media sources, during specific time-periods, and for specific keyword(s) using query language. These tools then extract, classify, represent, and store various pieces of information from each source. However, database tools can be inflexible and in some cases may lack the temporal and source coverage required for a full and proper analysis. Most limiting is the propriety nature of these databases that require expensive subscriptions. While free alternatives are available by performing manual searches in a specific newspaper web repository or by using existing newspaper-specific API tools, such data collection methods can be labor intensive, time consuming, and limited to specific news sources that may not be representative enough for a comprehensive analysis. These tools, which were built for different needs and objectives, can also produce results in a format that is not exportable and therefore incompatible with quantitative analyses and modeling applications.

Thus, we created a new instrument, *Articulate*, as a compliment to current news media discovery tools. *Articulate* is written in Python and interfaces with the Google Custom Search Engine (CSE) API to provide a free, flexible, customizable tool with functional outputs. Additionally, *Articulate* can offer equal and sometimes greater data coverage than produced by comparable proprietary databases. This tool can be used to gain insight into public exposure to a topic of interest such as environmental issues as well as to evaluate the most important events related to that topic. In this paper, we first describe the functionality and methodological details of the software package. Then, we validate the tool using two case studies of news media coverage of drought in California and flooding in Houston, Texas to demonstrate the algorithm's performance and applicability.

2. Articulate software package

2.1. Basic functionality

The *Articulate* software package uses existing open-source application program interface (API) tools to search through online content and news article databases, collect content of interest, tally search results, and generate a database of articles with pertinent information for further qualitative and quantitative analysis. The general process of the algorithm is shown in Fig. 1. The *Articulate* algorithm is written in Python and uses the Google CSE API to submit queries (web searches) programmatically into the Google search bar, step through multiple pages of results, and scrape only relevant news articles from defined news sources for specific keywords over a user-defined time window. In the process, *Articulate* filters out ads, sidebars, commentaries, and images typically found in news websites, thus eliminating sources of confounding information and returning only relevant results. The tool allows users more flexibility and control over the queries and resulting outputs than existing proprietary applications. While some familiarity with the Python programming language and the Google CSE API is required for initial setup, the *Articulate* package includes a GUI to facilitate its use. In the following sections we present *Articulate* as an accessible algorithm that lowers many of the barriers a user could face when trying to implement Google CSE API applications at socially- and environmentally-relevant scales. *Articulate* is open-source and is available as a GitHub repository at: https://github.com/Stanford-Urban-Water-Policy-Innovation/Articulate.

2.2. Software environment

Articulate uses the Python programing language, a freely available language with extensive capabilities that runs on both Windows and Unix-like platforms (Python Software Foundation, 2016). *Articulate* is written in Python version 2.7 with the following modules: csv.py, time.py, datetime.py, googleapiclient.discovery.py, ast.py, numpy.py, pandas.py, sys.py, dateutil.py, CookieJar.py, urllib.py, urllib2.py, and Tkinter.py. Google CSE API client requires internet access for use, and the user must first create an account with Google CSE API client to obtain an API developer key while also setting up their custom search engine with the parameters relevant to their needs (see https://developers.google.com/custom-search/docs/overview for more information).

2.3. Google CSE API client

The backbone of *Articulate* is the Google CSE API client, a web searching tool developed by Google that allows the user to submit any Google search programmatically (Google, 2017). It performs like Google's search bar, returning the same page of results, ten at a time, that would be achieved had the search been executed within Google's web search bar. The information within each result contains items which can include title of the result, media type (i.e. video, article, etc.), author, short excerpt, content keywords, date of publication, and various other attributes. The information is extracted and stored within the database specified by the user. In the same way that the user steps through various “pages” of results when searching within Google's search bar, the user must also step through various “pages” when using the Google CSE API tool. The flexible and dynamic nature of the CSE tool allows for an array of input parameters; the user has the capability to search various websites by date, by quoted content, by an exclusive query, and other functions.

The functionality of this tool allows the user to step through ten “pages” of the same search, up to the 100th result. Each query submission can retrieve up to 10 results, and each developer key gets 100 queries a day for free. After that, the Google Developer APIs are subject to a fee for each 1,000 queries used in a day, and the user can submit up to 10,000 queries a day with this method. We address these search retrieval limitations in our algorithm by developing a time-step method described in Section 2.6 below to cycle through the queries and results.

Download English Version:

https://daneshyari.com/en/article/6962205

Download Persian Version:

https://daneshyari.com/article/6962205

Daneshyari.com