Description

Each year, millions of books and journals are published. When researchers need to answer a question, how exactly do they find what they need? In the past, researchers used to gather data through screen scraping, which is the process of capturing data from a website using a computer. Today, sophisticated tools allow text and data mining (TDM), facilitated by organizations such as Crossref. Scholarly publishing has grown by leaps and bounds since 2000 when Crossref was founded. The organization now has over 5,000 members, represented by publishers from all disciplines, academic societies, presses, and open access publishers. Crossref makes it easier to mine journals and books for information using natural language processing (NLP). What is text mining? Text and data mining uses data mining tools to help researchers analyze and filter data resources, at the same time detecting patterns and connections using machines. So how does TDM work in the world of open access content?

Crossref's Starting Point

First, a researcher identifies the journals that he/she wants to work with. This is a tedious process because there are thousands of journals that are available. It is not practical for a researcher to contact each journal on his list for access to publications. It would also require some form of infrastructure to facilitate the process of delivering content to 100 researchers who request to access journals from one publisher. The aim of this process is ultimate to gather journals from which to mine data. To facilitate this, publishers such as Elsevier have created article programming interfaces (APIs), which is a convenient way to download content in bulk. Since Crossref is affiliated with thousands of journals, it solves logistical and technical issues related to TDM by allowing scholarly researchers to access subscription and open access content.

The Role of Crossref Metadata

Second, TDM tools need to be applied to a body of data that you wish to mine. In order to gather large amounts of data, you need to bulk download your content from publishers and across a number of platforms. Digital object identifiers (DOIs) and metadata are useful here because these stabilize the content of journals online and ensure that they remain available where they were published. Crossref is the biggest DOI registration agency. Members can update Crossref metadata if the web address where

a piece of content is hosted changes. The Crossref Metadata API was launched in 2013 and can be leveraged to provide cross-publisher support for TDM purposes. It is free to use and lets anyone search and filter Crossref metadata. It is also easier to integrate into communities, which increases discoverability.

TDM Application and Analysis

Third, you can now analyze your results after applying your TDM tool. Text and data mining are often coupled with visualization techniques to facilitate the discovery of patterns within the data. These techniques include tag clouds, stream graphs, tree maps, heat maps, scatter plots, and time series and can all be used to show the relationships between entities. In addition to detecting patterns, these techniques can be used to automatically assign documents to groups without human intervention, through either categorization or clustering. Lastly, the original research problem can then be answered from the themes that emerged from TDM.

Text and Data Mining and Open Access

What is text mining? Text mining is the act of ‘mining’ data from cross-publisher platforms and is the evolution of screen scraping. Text and data mining is a constantly evolving field with applications that are becoming more and more valuable. The increasing computational capabilities, along with the fast-growing availability of digital content, hold great promise for the future. In the world of open access content, data mining tools are critical for finding information amidst the wealth of journals and online content. Crossref seeks to overcome the logistical and technical barriers of bringing together journals across publishing platforms for scholarly publishing.

Category

1. Career Corner
2. Product & Service Reviews

Date Created
2017/09/27
Author
daveishan