Accelerating Scientific Discovery with Text Mining

SOFTWARE Advanced text mining software provides access to the relevant information researchers need.

In his 2014 TedX Talk, Charles Stryker, CEO of the Venture Development Center points out that scouring journals the usual way and reading them one by one is not very effective. The average oncologist might be able to keep track of six or eight similar cancer cases at a time recalling details that might help him or her go back, re-read one or two of those articles, and determine the best course of care for a patient with an intractable cancer. The data banks of two major cancer institutes, however, hold searchable records of cancer cases that can be reviewed in conjunction with information about three billion DNA base pairs and 20,000 genes. Using that data would mean a vast improvement in the odds of finding clues to help treat a complicated case, or target the best clinical trial for someone with a rare disease. It is difficult, if not impossible, for even the most sophisticated oncologist to locate, read, see patterns emerge, and memorise what is important in that huge amount of information. The ability to finally find the needle in the proverbial haystack is the promise of text and data mining.

KIM ZWOLLO, the article's author is General Manager at RightsDirect.

R&D departments in the biotech sector increasingly rely on text and data mining solutions to improve and accelerate the process of discovery. However, significant obstacles exist. From access to full-text articles, licensing challenges and publisher negotiations to content format inconsistencies, researchers find themselves struggling with manual processes to set-up a text mining project well before they can even begin the actual mining.

Picture: RightsDirect

What Is Text Mining?

Text mining, or text analytics, allows scientists to use technology to sift through vast amounts of unstructured content to gain insight. Text mining is the data analysis of natural language works (articles, books, etc.) using text as input. It is often joined with data mining, the analysis of data works (like filings and reports). Text mining uses advanced software that allows computers to 'read' and digest digital information far more quickly than a human being can possibly do. Text mining software breaks down text-based content into smaller conceptual units, analyses them, permits rapid comparisons and matching across different sources, and comes up with previously invisible connections, such as unexpected patterns in protein interactions that eventually lead to the development of a new drug.

Text Mining Challenges

Consider the difficulties of researchers wishing to use text mining to see if cancer patients taking a certain diabetes drug might have a better outcome than

patients who are not taking the drug. A typical journal subscription doesn't permit the routine reproduction of full text articles that is required for high-level computer-based mining. Researchers may have access to article abstracts and metadata for mining purposes, but the inability to mine the full-text article means that information that may hold the key is excluded from the mining effort. Most publishers offer full-text journal content in PDF format, but advanced text mining software requires machine-readable content feeds in a normalsed XML format.

In these circumstances, the researcher must work with the publisher of each specific journal, negotiate for the commercial text mining rights, work out an arrangement to get access to each publisher's XML content feed, and then independently convert these different feeds into a single normalised XML format. All of these steps need to be taken before any actual text mining can be done. If the top 20 biotech companies did this with the top 20 publishers, it would take 400 agreements, 400 feeds, and maybe as many as 400 XML normalisation conversions. The effort needed is substantial, both for the researchers and for the publishers.

In an ideal world, researchers would be able to access an aggregate of all relevant journals in their field of interest, even if the organisation they work for has no subscriptions to certain journals. Instead of 400 agreements and feeds to navigate, and instead of 400 XML normalisations, one solution could provide researchers with access to XML-formatted content along with the necessary rights for mining. In other words, researchers could get access to the high-value relevant information they need to move research and healthcare forward, in less time, with less effort.

RightFind™ XML for Mining

To enable researchers to find answers and make connections faster, RightsDirect and its parent company Copyright Clearance Center created RightFind™ XML for Mining. Using this solution, researchers can quickly obtain full-text XML normalised content from a number of leading scientific publishers. This integrated solution gives researchers the most complete article collection for mining in one consistent format, which can then be used in their preferred text mining software application, such as Linguamatics I2E. All content available through RightFind XML for Mining is preauthorised for commercial text mining, saving researchers time and money they would have invested in acquiring and licensing articles from individual publishers.

Picture: RightsDirect

Improved Text Mining Results

RightFind XML for Mining enables researchers to make discoveries and connections that can only be found in the full-text version of articles, seeing connections that cannot be found when searching in abstracts only. The new tool allows researchers to even search across full text content of publications from different publishers to which they don't subscribe, and this is a strong feature.

Ensure Copyright Compliance

Because all of the content in RightFind XML for Mining is pre-authorised for commercial text mining, researchers – and the intellectual property teams that advise them – can feel confident that text mining projects comply with an organisation's copyright policy to minimise infringement risk.

Save Time and Money

Because it aggregates article content and normalises the XML feeds from multiple publishers into a central location for fast and easy access, RightFind XML for Mining reduces the time and costs associated with article conversions, content management and negotiations with publishers. That means researchers have more time to focus on high value processes such as analysis and discovery.

Learn more

Learn how RightFind XML for Mining can help improve the results of your text mining efforts, strengthen your compliance programme and save money. Find more information at:
› www.rightsdirect.com/XMLforMining

Contact
Joost Kollöffel
Senior Marketing Manager
RightsDirect
+31(0)-6-46-29-89-24
jkolloffel@rightsdirect.com