Copyright Clearance Center

TOP 3 CHALLENGES FOR COMMERCIAL TEXT MINERS

How Content Access and Manual Process Hinder Mining Efforts

In biomedical R&D, researchers use text mining tools to extract and interpret facts, assertions and relationships from vast amounts of published information. Mining accelerates the research process, increases discovery and helps companies identify potential safety issues in the drug development pipeline. However, despite the many benefits of text mining, researchers face a number of obstacles before they even get a chance to run queries against the body of biomedical literature. Here are the three primary challenges for researchers as they build a collection of articles (or "corpus") for their text mining projects:

1 Incomplete Information in Article Abstracts

Many researchers build their corpus using scientific article abstracts because they are easily accessible via biomedical databases such as PubMed. While text mining data from abstracts provides some value, there are limitations as to what data can be found within an abstract. The ability to mine the full text of the article — including detailed descriptions of methods and protocols and the complete study results — ensures that researchers don't miss vital data, discoveries and assertions. However, unlike article abstracts, full text is not often readily available from publishers in a format suitable for text mining.

2 Limited Access to Xml-Formatted Content

When researchers have subscriptions the documents are often provided as PDFs, a format not intended for use with text mining software. Researchers must then spend time converting the PDFs to XML (Extensible Markup Language), the preferred format for use in text mining software. XML is a markup language used to encode documents in a format that is easily read by computers.

It is used widely for encoding documents so that computer programs can parse or display the content appropriately. To convert PDFs to XML, researchers must use additional software tools which is not only inefficient but also creates a number of problems with the document itself, including loss of data and tables, conflation of document sections into a "blob of text," and the addition of bad characters and non-words.

3 Inconsistent Licensing Terms and Fees

Because text mining projects depend on access to a broad base of content, businesses must work directly with multiple rightsholders for the use of full-text XML articles, resulting in varying fee structures, inconsistent terms of use and ultimately reduced productivity. Without a common set of terms and conditions for the use of full-text content across publishers, researchers and/or information managers are left with the task of negotiating one-by-one with individual rightsholders to obtain the content and rights they need for text mining.

Copyright Clearance Center

Copyright Clearance Center (CCC) is a global leader in content management, discovery and document delivery solutions. Through its relationships with those who use and create content, CCC drives market-based solutions that accelerate knowledge, power publishing and advance copyright. With its subsidiaries RightsDirect and Ixxus, CCC provides solutions for millions of people from the world's largest companies and academic institutions around the world.

LEARN MORE

For more information about RightFind® XML for Mining, contact CCC today.

@ **1.978.750.8400 (option 3)**

☎ **info@copyright.com**

🌐 **www.copyright.com/xmlformining**