RightsDirect

A Copyright Clearance Center Subsidiary

WHITEPAPER

Turning Text into Insight: Text Mining in the Life Sciences

According to *The STM Report (2015)*, 2.5 million peer-reviewed articles are published in scholarly journals each year.[1] PubMed contains more than 25 million citations for biomedical journal articles from MEDLINE. This enormous and continuously growing body of research holds the potential to help shed light on new discoveries and guide R&D in life sciences industries. Given the volume of scientific literature and the pace at which it is published, it's neither feasible nor cost-effective for researchers to read and analyze this material one article at a time. Only an automated process like text mining can adequately analyze massive amounts of information quickly to extract data, assertions, and facts from text sources.

In this white paper we discuss what text mining is and three approaches to maximize its efficiency and potential for discovery.

WHAT IS TEXT MINING?

Text mining is a process that derives high-quality information from text materials using software. It is often used to extract assertions, facts, and relationships from unstructured text (e.g., scholarly articles), for purposes of identifying patterns or relations between items that would otherwise be difficult to discern. Text mining tools employ sophisticated software which uses natural language processing (NLP) algorithms to read and analyze text. Text mining enables R&D teams to systematically and efficiently examine the biomedical literature to answer questions that ultimately guide business decisions and resource investments.

Broadly, text mining involves two phases. First, identifying the entities that an organization is interested in, such as genes, cell lines, proteins, small molecules, cellular processes, drugs, or diseases; then, analyzing the sentences in which those key entities appear to determine how they are related. A relationship is a connection between at least two named entities; for example, that gene BCL-2 is an independent predictor of breast cancer.

Mining can uncover relationships that might not have been found otherwise, unlocking previously hidden information to help researchers:

- Identify and develop new hypotheses
- Attain knowledge and improve understanding
- Discover links between diseases and existing drugs to find new therapeutic uses
- Detect potential safety issues early

For example, the results of mining projects provide a greater understanding of the underlying biology behind specific diseases, show how they respond to certain drugs and support the target discovery process.

HOW DOES TEXT MINING DIFFER FROM A WEB SEARCH?

While typical Web searches may seem similar to the process of text mining, there are stark differences. Search is the retrieval of documents based on certain search terms. Search engines such as Google, Yahoo or Bing are the common tools used to conduct these types of searches. The output is typically a hyperlink to text/information residing elsewhere, along with a small amount of text that describes what is found at the other end of the link. The purpose is to find the entire existing work so that its content can be used.

In text mining, the researcher is looking to analyze text. The goal is to extract information that is useful for particular purposes, not solely to find, link to, and retrieve documents that contain specific facts. Unlike with search, the output of text mining will vary depending on the use to which the researcher wishes to apply the results.

Where search helps users find the specific document(s) they are looking for, text mining goes well beyond search, to find particular facts and assertions in the literature in order to derive new value.

THREE TIPS TO MAXIMIZE YOUR TEXT MINING EFFORTS

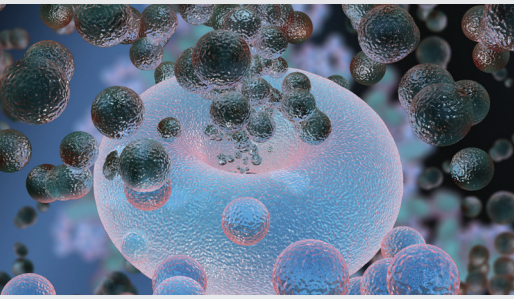
To obtain the full benefits of text mining, commercial researchers should consider using full-text articles versus abstracts; work with content in a format optimized for mining, such as XML (Extensible Markup Language); and take steps to ensure that content is licensed by the publisher for commercial use:

① Mine full-text articles over abstracts

Many researchers elect to use the summary information in article abstracts, which are easily accessible via biomedical databases such as PubMed, to compile a collection of articles (or "corpus") for use in text mining rather than full-text articles. In addition to their general accessibility, abstracts are typically provided in a format that is suitable for text mining (e.g., XML).

However, abstracts often don't include essential facts and relationships, access to secondary study findings, and adverse event data. While abstracts do provide some valuable information, researchers need access to full-text articles to get the best results from text mining projects.

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. XML is the preferred format for use in text mining software



Access More Facts

Full-text articles provide more information than abstracts — including detailed descriptions of methods and protocols and the complete study results. While authors often include their most important findings in the abstract, secondary study findings, discoveries, and observations include critical insights but are found only in the full-text article. Given the size limitations of abstracts and their concise nature, they often exclude, or underrepresent, data or results that are considered to be less relevant or out of scope with the main idea of the publication. Further, in some cases, critical information may reside in a footnote. But, by mining all of a given text, including bibliographic information, researchers can gain richer results that reveal vital patterns and information in the documents.

In addition, new discoveries are more likely to be mentioned in the full text of articles before appearing in abstracts. Following initial publication of a new discovery in a particular journal, the research is often repeated and included in other publications. But there is a substantial delay between when that discovery appears in full articles and when that information appears in abstracts. In fact, it can take one to two years for discoveries to appear in the abstract of a subsequent article.[3]

Lastly, full-text articles are more likely than abstracts to contain information on adverse events. According to a study published in BMC Medical Research Methodology, "Abstracts published in high impact factor medical journals underreport harm even when the articles provide information in the main body of the article."[4] This missing information can reduce the value of abstracts as the "raw material" to mine, especially in pharmacovigilance use cases, or when researchers want to make novel connections that haven't yet been a major focus of the literature.

Uncover More Relationships

Full-text articles also contain more relationships between named entities than abstracts. According to a study published in the Journal of Biomedical Informatics, only 8% of the scientific claims made in full-text articles were found in their abstracts.[5]

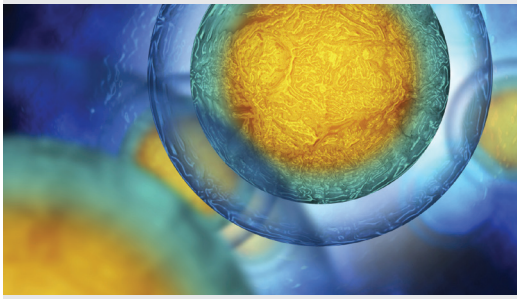
Another study, conducted by publisher Elsevier, compared using abstracts and full-text articles to derive relevant information about drugs and proteins that affect the progression of fibromyalgia. They found 31 relationships in the literature by mining abstracts and an additional 53 relationships when they ran the same search across the full-text articles.[6]

While text mining article abstracts yields some information, there are limitations as to what can be discovered through that process. To ensure that researchers don't miss vital data, discoveries, and assertions, the full text of the article should be mined.

② Obtain XML files versus converting PDFs

Unlike article abstracts, full-text articles are not often readily available from publishers in a format suitable for text mining. When researchers obtain full-text articles through company subscriptions or document delivery, the documents are often provided as PDFs, a suboptimal format for use with text mining software. The burden is then on researchers to convert the PDFs — potentially thousands in a bulk delivery — to XML.

While XML has its advantages, the format presents a challenge: Tasking highly-skilled researchers with converting document formats for input into text mining tools is inefficient and costly, and creates a number of problems with the source documents.

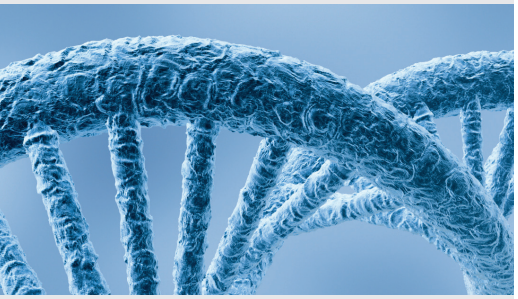


First, to convert PDFs to XML, researchers must do so through software tools, a process which is time-intensive and which creates a number of problems with the document itself, including loss of data and tables, conflation of document sections into a "blob of text," and the addition of bad characters and non-words. The conversion process introduces the possibility of error (e.g., poor character recognition for uncommon fonts) and often removes tags that indicate sections of the article, such as introduction, conclusion, and materials and methods, making the corpus difficult to mine.

While some PDFs have text embedded in the document and apply fonts to render the text readable, they lack the comprehensive metadata and tagging of document sections which are included in XML documents.

All these issues result in an increase in false positives, especially when matches are found within the bibliographic citations and a likelihood that true positives may remain hidden in the text.

Given the problems with converting PDFs, many researchers opt to acquire XML feeds from publishers. This approach, too, can be challenging given the wide variation among publishers in how the data is delivered and the need to then normalize the material (e.g., metadata) into a single standard to text mine against it efficiently. The challenge for the researcher shifts from converting PDFs to XML to converting multiple XML feeds into a common and consistent XML format.



③ Make sure content is licensed for commercial mining

Researchers often negotiate with publishers for the right to text mine content for commercial purposes, as this right is not commonly included in standard subscription agreements. Converting PDFs intended for human consumption into a machine-readable format such as XML results in the creation of additional copies. Creating and storing those reformatted copies typically requires additional permission from the publisher.

The absence of any mention of text mining in the terms of a subscription agreement does not mean that it is permitted. While there is an exception under the law in the United Kingdom for text and data mining for non-commercial research, no such exception exists for research on behalf of corporations.

In addition, researchers must deal with inconsistent licensing terms and fees from multiple publishers. Because text mining projects depend on access to a broad base of content, organizations must work directly with multiple rightsholders for the use of full-text XML articles. This results in varying fee structures, inconsistent terms of use, and ultimately, reduced productivity.

Without a common set of terms and conditions for the licensed use of full-text content across publishers, researchers and 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.

DISCOVER HIDDEN KNOWLEDGE AND ACCELERATE THE PACE OF DISCOVERY

Text mining has enormous potential, but it can only fully accelerate and enrich a company's research and development program when the barriers between researchers and the content they want to mine are lowered. If researchers spend time tracking down full-text articles and obtaining permissions from individual publishers and converting full-text article PDFs to XML before they are able to mine the content, there could be a loss in productivity (as much as 4 to 8 weeks to prepare a corpus), a deceleration of the research and development process, and increased copyright infringement risk.

IMPROVE YOUR TEXT MINING RESULTS WITH FULL-TEXT XML ARTICLES

RightFind® XML for Mining from RightsDirect enables you to make discoveries and connections that can only be found in full text. You can obtain XML-formatted full-text content from publications you subscribe to and discover articles that fall outside of company subscriptions, giving you the most complete full-text article collection for mining. This lets you:

- Go beyond the abstract level to search, download and mine full-text articles in XML format from both company subscriptions as well as unsubscribed published material;
- Gain the peace-of-mind that your text mining projects comply with copyright, minimizing your organization's infringement risk; and
- Reduce the time and costs associated with article conversions, content management, and negotiations with publishers.

1 STM: International Association of Scientific, Technical and Medical Publishers (2015) The STM Report, Fourth Edition. Available at http://www.stm-assoc.org/2015_02_20_STM_Report_2015.pdf

2 Weeber, Marc et al. "Generating Hypotheses by Discovering Implicit Associations in the Literature: A Case Report of a Search for New Potential Therapeutic Uses for Thalidomide." *Journal of the American Medical Informatics Association: JAMIA* 10.3 (2003): 252–259. PMC. Web. 16 Feb. 2015.

3 Elsevier (2015) Harnessing the Power of Content - Extracting value from scientific literature: the power of mining full-text articles for pathway analysis. Available at www.elsevier.com/__data/assets/pdf_file/0016/83005/R_D-Solutions_Harnessing-Power-of-Content_DIGITAL.pdf

4 Enrique Bernal-Delgado and Elliot S Fisher. "Abstracts in high profile journals often fail to report harm." *BMC Medical Research Methodology* (2008); 8:14

5 Catherine Blake. "Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles." *Journal of Biomedical Informatics* Volume 43, Issue 2, April 2010, Pages 173–189.

6 Elsevier (2015) Harnessing the Power of Content - Extracting value from scientific literature: the power of mining full-text articles for pathway analysis. Available at www.elsevier.com/__data/assets/pdf_file/0016/83005/R_D-Solutions_Harnessing-Power-of-Content_DIGITAL.pdf

RightsDirect

A Copyright Clearance Center Subsidiary

RightsDirect, a wholly owned subsidiary of Copyright Clearance Center (CCC), provides licensing and information management solutions that make copyright compliance easy, allowing companies to manage permissions and share the most relevant digital content across borders. Together, CCC and RightsDirect serve more than 35,000 companies and over 12,000 publishers around the globe.

LEARN MORE

Learn how RightFind XML for Mining can help you improve the results of your text mining efforts, increase efficiency, strengthen your compliance program and save money by contacting RightsDirect.

info@RightsDirect.com

+31-20-312-0437

rightsdirect.com