prc
publishing research consortium

Text Mining and Scholarly Publishing

Jonathan Clark

By Jonathan Clark, Loosdrecht, The Netherlands,

Jonathan Clark is an independent advisor on strategy and innovation. Prior to starting his own company, he worked with Elsevier for 20 years in various roles in publishing, product management, technology, strategy & business development. Jonathan is a former Chair and Director of the International DOI Foundation.

The Publishing Research Consortium (PRC) is a group representing publishers and societies supporting global research into scholarly communication, with the aim to provide unbiased data and objective analysis.

Our objective is to support work that is scientific and pro-scholarship. Overall, we aim to promote an understanding of the role of publishing and its impact on research and teaching.

Outputs from work supported by the PRC are available from the website:
www.publishingresearch.net

The founding partners are The Publishers Association, the Association of Learned and Professional Society Publishers, and the International Association of Scientific, Technical & Medical Publishers. Corresponding partners include The Association of American University Presses and the Professional / Scholarly Publishing Division of the Association of American Publishers.

Contents

Acknowledgements

The author gratefully acknowledges the help and advice of the following people in preparing this briefing paper:

Mark Bide
Geoffrey Bilder
Bob Campbell
Judson Denham
Maximillian Haeussler
Michael Jubb
Marc Krellenstein
David Marques
Cameron Neylon
Heather Piwowar
Carlo Scollo Lavizzari
Alan Singleton
Graham Taylor
Anita de Waard

1 Introduction

What is text mining? How does it relate to data mining? Why do people want to do text mining? How does it work? What do publishers need to do to support text mining?

There are many questions swirling around the topic of text mining of the scholarly literature. At the time of writing, text mining is generating a frenzy of debate in the scholarly publishing world. There is the usual misunderstanding, over-enthusiasm and unrealistic expectations that are associated with technology hype.[1]

There is no universally agreed definition of text mining. This is partly because it is being used by different communities for different purposes. Each community has its own preferred scope and definition of text mining. This can lead to disagreements over where information extraction finishes and text mining starts or the difference between text mining and data mining. This paper aims to disentangle the topic and to clarify the underlying issues for the general, non-expert reader.

The scope of this paper is the text mining of scholarly journals and books. The word "text" is used to describe the content of these sources and "publishers" refers to scholarly publishers only. The focus is on what scholarly publishers can do to make their content more machine-accessible, although it is hoped that others will find this paper helpful, for instance for those who are responsible for making policy in this area, whether it be at trade or government level, researchers not currently engaged in the area, librarians, and other interested parties.

2 What is Text Mining?

Fundamentally, text mining is the indexing of content. Words that are part of a fixed vocabulary are found within a text and extracted to create an index that shows where in the text each word was found. The index can be used in the traditional way to locate the parts of the text that contain those words. The index can also be used as a database and analysed to discover patterns: for example, how often certain words occur. In simple terms, text mining is the process that turns text into data that can be analysed.

Bad indexes are nothing more than a basic keyword search. They show where a word occurs in a text but there is no guarantee that this word has the same meaning as the word that was searched for. It's just a word. For example, "substrate" is a word that has many different meanings depending on the scientific discipline.

Good indexes point to the meaning of a passage of text. They take the context around the word into account. This is what text mining aims to do: to extract the meaning of a passage of text and to store it as a database of facts about the content and not simply a list of words. One could say that text mining is smart indexing.

BRIEF HISTORY OF TEXT MINING

Hearst[1] gives an excellent overview of the early development of text mining. Text mining has its roots in computational linguistics and information retrieval. Large bodies of text were used to develop and test better text analysis algorithms. Some of these early algorithms helped to create tools that greatly improved information retrieval. Hearst was one of the first to suggest the use of large text collections to discover new facts and trends by applying data mining techniques to texts. He called this Text Data Mining.

1 "Untangling Text Data Mining", Marti A. Hearst, Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, June 20-26, 1999 (invited paper) doi:10.3115/1034678.1034679.

1 Gartner Hype Cycle is a graphic representation of the maturity and adoption of technologies and applications. http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp

Text mining is not new. Humans have read and extracted meaning from written works ever since they learned to write. They have created indexes, summaries and classifications from the facts they extracted. The availability of digital texts allows machines to do text mining faster and often more consistently than humans ever could.

Automatic indexing typically finds all occurrences of an index term. Text mining finds the terms and relates them to the context and meaning of the sentence or paragraph that contains the terms. Unsurprisingly, finding two terms in the same sentence is a much better indicator of true relevance than finding the same two terms in, say, an entire journal article.

Text mining extracts meaning from text in the form of concepts, the relationships between the concepts or the actions performed on them and presents them as facts or assertions. These facts are stored in a relational database that can then be used for analysis. The analysis of the concept database is often called data mining. There is an implicit assumption that the extracted concepts do have meaning.

STATISTICS VS SEMANTICS

It appears logical that the only way a computer can read like a human is if it can be taught to understand human grammar - this is the semantic approach to computational linguistics. The idea that a computer can teach itself to read by using statistical methods seems far fetched. And yet huge progress has been made in recent years. One of the best examples of how far statistical techniques have come is Google Translate. Google collects as many manually translated books and documents as it can find. Translation queries are then compared against all these volumes. It might make mistakes, but this statistical approach still beats all semantic methods of translation.[1]

1 "How Google Translate works", David Bellos, http://www.independent.co.uk/life-style/gadgets-and-tech/features/how-google-translate-works-2353594.html

Text mining uses tools and techniques developed as a result of research in computational linguistics. This is the scientific field that studies the use of computers to process and analyse language.

Some people have defined text mining as the discovery of new knowledge from a large body of natural language text using computational algorithms to extract semantic logic. This author believes that it is helpful to distinguish between text mining as the extraction of semantic logic from text, and data mining which is the discovery of new insights. The knowledge that is extracted during text mining is not new and it is not hidden. That information was already known to the author of the text, otherwise they could not have written it down.[2] There is value in text mining alone, for instance in enriching scholarly publications or helping readers keep up with the literature. Data mining holds great promise for discovery but it is not the only reason to do text mining.

3 What is Data Mining?

Data mining is an analytical process that looks for trends and patterns in data sets that reveal new insights. These new insights are implicit, previously unknown and potentially useful pieces of information. The data, whether it is made up of words or numbers or both, is stored in relational databases. It may be helpful to think of this process as database mining or as some refer to it "knowledge discovery in databases". Data mining can be used to mine any database, not just ones created using text mining. Data mining for scientific research is well established in fields such as astronomy and genetics.[3]

2 "Untangling Text Data Mining", Marti A. Hearst, Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, June 20-26, 1999 (invited paper) doi:10.3115/1034678.1034679.

3 See for example: http://www.astro.caltech.edu/~george/aybi199/Djorgovski_DMintro.pdf and http://www.ebi.ac.uk/luscombe/docs/imia_review.pdf

4 Why do Text Mining?

Broadly speaking there are (so far) four main reasons to embark on text mining: to enrich the content in some way; to enable systematic review of literature; for discovery or for computational linguistics research.

4.1 Enriching the Content

Text mining can be used to improve the indexing of text. This is especially useful for publishers since they can create indexes more efficiently (machine-aided indexing). They can also add value in a digital environment by creating links from articles to relevant further reading. For example, mentions of gene sequences can be identified in articles and linked directly to databases such as GenBank.[4] This use of text mining is widespread and predicted to grow quickly. 46% of publishers in a recent study[5] reported that they currently text mine their own content, of the ones that do not, a further 30% will start doing so within a year of the study.[6] Third party tools have also been developed to improve the reading experience, such as Utopia Docs which identifies named entities within PDFs and build links out to related web resources in the life sciences.[7]

4.2 Systematic Review of Literature

Text mining can help a scientist to systematically review a much larger body of content and do it faster. There is considerable demand for this kind of text mining in the corporate environment: why pay biologists to read biology papers when machines can do it for them, and they can concentrate on doing research instead? Furthermore, text mining can help researchers keep up with their field and reduce the risk that they miss something relevant.

4.3 Discovery

Text mining is used to create databases that can be mined for discovering new insights. Many people, especially in the pharmaceutical world, believe that there is huge promise here and to a large extent this is driving the hype around text mining. Scholarly texts are written to communicate factual information or opinions and so it seems to make sense to try to extract this information automatically. However, there are very few published examples that show new insights as a direct result of data mining. One example identifying new therapeutic uses for thalidomide is often quoted.[8] It is not clear what can be considered as a new insight. Is it the discovery of some sort of association between a gene and the literature surrounding a particular disease, or is it only an insight if the association is verified in the lab? It is probably more useful to think of text mining as machine-aided research tool that can open up additional sources of information for use in research rather than as some sort of holy grail.

4.4 Computational Linguistics Research

Text mining itself is the subject of research into text mining. There is considerable work worldwide in the field of computational linguistics dedicated to improving the extraction of meaning from text. Text mining is the raw material for this research. This area appears to be driving a very large part of the current activity in text mining. Around half of the publishers recently surveyed have been approached by researchers in this field requesting permission to mine their content.[6] This is also the area where the

4 GenBank http://www.ncbi.nlm.nih.gov/genbank

5 "Journal Article Mining, A research study into Practices, Policies, Plans.....and Promises". Commissioned by the Publishing Research Consortium by Eefke Smit and Maurits van de Graaf, Amsterdam May 2011 http://www.publishingresearch.net/documents/PRCSmitJAMreport20June2011VersionofRecord.pdf. Hereafter noted as Smit E. & van de Graf, M.

6 See Elsevier's Article of the Future http://www.articleofthefuture.com and Smart Content http://river-valley.tv/smart-content-at-elsevier/; Content Enrichment from Wiley http://river-valley.tv/content-enrichment-a-publishers-perspective/; Ref Enhanced articles from Royal Society of Chemistry http://www.rsc.org/Publishing/Journals/ProjectProspect/

7 Utopia Docs from Lost Island Labs, a spin-out from University of Manchester, UK http://getutopia.com.

8 "Generating hypotheses by discovering implicit associations in the literature: A case report for new potential therapeutic uses for Thalidomide," Weeber et al. J Am Med Inform Assoc. 2003 10 252-259.

most progress is being made in developing new tools and techniques. The research in this area is often challenge driven. For example, BioCreative[9] sets text mining tasks as challenges that are relevant to biomedicine and that stimulate the community to develop new methods. Challenges like this have also resulted in the development of tools for researchers in other scientific disciplines such as Reflect for the life sciences.[10]

5 Text Mining and Meaning

Text mining employs Natural Language Processing (NLP) to extract meaning from text using algorithms. One approach is essentially rule-based and looks for predefined words and patterns of words from which meaning can be inferred. The fundamental idea is that if the computer can be taught the rules of grammar and usage then it will be able to derive meaning from the words it recognises. For instance, the computer can be instructed to recognise the words in this sentence: "insulin is a peptide hormone" and to identify the parts of speech so that it can derive meaning i.e.: insulin belongs to the class of peptide hormones and it can infer meaning e.g.: insulin belongs to the class of proteins (because it already knows that peptide hormones are proteins).[11]

This approach can be very successful but it has limitations. For instance, longer sentences can be highly ambiguous when processed this way. Moreover, authors are always bending and stretching the rules of grammar to fit their needs.

A different approach using statistical methods is becoming increasingly popular and the techniques are improving steadily. These methods use the frequency and location of words to reveal concepts. A simple example would be to determine the most common or least common terms in a text and then identify other terms that occur together with these. This technique has been used to build auto-classification tools. In the case, the system is first "trained" using a small sample set of documents whose classification is known. It then uses the patterns that it learnt to classify new documents.

6 Text Mining and Discovery

It is generally accepted that mining of large data sets such as gene sequences can be useful to discover new insights that would not be possible using any other technique. The computer can analyse huge quantities of data and look for trends and patterns that are not obvious and have never been seen before. But how can this also be applied to a large collection of documents?

One illustrative example of how this might work is as follows. According to the published literature there is no relationship between deforestation and hurricanes. No amount of text mining will reveal these or similar words in the same context. Moreover, scientists working in deforestation are unlikely to meet hurricane experts. However, there are sentences to be found that link deforestation to increased silt in rivers. Furthermore, there are sentences that relate increased silt in rivers to seas becoming shallower due to silt run-off. And shallow seas have been linked to an increase in hurricane formation. A new question emerges from this analysis: is it possible that deforestation could lead to more hurricanes?

The train of causal relationships linked otherwise unlinkable articles and discovered a new hypothesis that can be researched further. This example is based on work by Swanson.[12] It is worth looking into

9 BioCreative http://www.biocreative.org
10 Reflect http://reflect.ws
11 This is also an example of a triple, where "insulin" the subject, "peptide hormone" is the object and "is a" is the predicate. See Glossary for explanation of triple.
12 "Two medical literatures that are logically but not bibliographically connected" Don R. Swanson Journal of the American Society for Information Science Volume 38, Issue 4, pages 228–233, July 1987 DOI: 10.1002/(SICI)1097-4571(198707)38:4<228::AID-ASI2>3.0.CO;2-G.

his research that started with human mining in the 1980s and led to the development of a software tool called Arrowsmith.[13]

One of the problems encountered in trying to link documents from different scientific fields is the use of jargon and of terms that have specific meanings in each field. This can make it difficult for experts in each field to understand each other. It could be that the real value of text mining will be to remove arcane jargon that can only be understood by domain experts so that articles from one scientific field become easily accessible to researchers from a different field.

GETTING STARTED WITH TEXT MINING

For Publishers

The first step for a publisher who wishes to make their content available for text mining is to decide the terms and conditions under which they will do so. The simplest approach for non-commercial, research use would seem to be a standard license such as CC-BY-NC. Alternatively, they could use the STM template[1] to add a clause to existing institutional agreements. Commercial use would most likely require a separate, negotiated license.

The next step is to make all the relevant people aware of what kind of text mining is possible, and under what conditions. It is especially important that the rights and permissions departments know who to foward such requests to. Social media is a good way to alert researchers.

Access to the content must be arranged. In some cases, one-time delivery of a package of digital content may be sufficient but it is more likely that some sort of feed is required so that the body of mined content can be updated with newly published material. The simplest way to achieve this is to grant access to the digital platform and allow automatic crawling of the articles. In some cases however this may not be straightforward, for instance when the platform is run by a third-party such as HighWire, Atypon, MetaPress, or Ingenta,

PDF content is sufficient for many text mining purposes but it much more useful to offer both the PDF and HTML contents if that is available. Content in XML format, and preferably using a widely-used format such as the PubMed DTD is the most useful.

The publisher will need to provide some basic technical support.

For Researchers

The are a large number of resources to help the researcher who is interested in doing text mining. For the biomedical field there is a good wikipedia entry[2] and an excellent paper by Cohen and Hunter.[3]

In general, the researcher should start by defining the goal of their text mining - what should the system do? What is the use case? The next step is to find examples of the text that will be mined and to review it in detail. As Cohen and Hunter note: Scientists without linguistic training are often amazed at the range of linguistic possibilities for expressing even the most apparently simple concepts. Another key question to tackle early on is how the text mining will be evaluated.

Unfortunately, there is no similar guide to obtaining the necessary rights and permissions for the content that is needed. Perhaps the best strategy is to search blogosphere and find someone with direct experience

1 STM Statement on Text and Data Mining and Sample License http://www.stm-assoc.org/text-and-data-mining-stm-statement-sample-licence/
2 http://en.wikipedia.org/wiki/Biomedical_text_mining
3 Cohen KB, Hunter L (2008) Getting Started in Text Mining. PLoS Comput Biol 4(1): e20. doi:10.1371/journal.pcbi.0040020, http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.0040020

13 Arrowsmith http://arrowsmith.psych.uic.edu/arrowsmith_uic/index.html

7 How to do Text Mining

7.1 Selecting the Sources

The first ingredient needed for text mining is the content, the text that will be mined. The choice of which and how much content is needed depends on the reasons for doing the text mining: publishers enriching their own content; systematic review of literature; discovery or computational linguistics research

In the case of a publisher mining their own content it is usually straightforward - everything that is available. Most publishers with a digital workflow store all their content in a single data warehouse so mining all of it is easier than setting up rules to extract only certain sources.

For systematic review of literature the body of content will usually be determined by the scientific field of study and include all published items that are relevant to that field or a statistically representative sample of the content. It is important that the content is refreshed regularly to keep track of the latest publications.

In the case of mining for discovery of new, previously hidden insights it is harder to define the content needed for mining. Knowledge discovery is the search for hidden information. By definition the location of this information is unknown (if it were known then keyword searching would find it). Hence the need is to start looking as widely as possible in the largest set of content sources possible. This need is balanced by the practicalities of dealing with large amounts of information, so a choice needs to be made of which body of content will most likely prove fruitful for discovery. Text mines are dug where there is the best chance of finding something valuable.

Researchers in computational linguistics have different content needs. Their research goals are usually to do something new and interesting in text mining. Researchers who want to develop new tools and techniques need a consistent body of content that is large enough to demonstrate that their tools are effective. There is no consensus over how large is large enough in this context, but the general feeling is that as large as possible is best. Other computational linguist researchers are hunting for new applications of text mining and are looking for new content sources that have never been previously mined.

7.1.1 Problems with Sourcing Content

Extracting concepts from a body of text requires that words and combinations of words are recognizable in the text - so the text must be machine-readable. In addition, the machine must know the location of the concepts in the text - so the text must be structured. The machine can be taught to recognize sentences but it needs help to identify titles, section heading and so forth. The more structure there is, the more meaning can be extracted by the machine. Knowing certain words appear in the same sentence is a great step forward, but knowing these words also appear in a heading or a title, for example, is much more useful since this implies a relationship beyond the words themselves. For example, a section heading “protocols” would imply that the section contains information about the protocols used. This might allow information on protocols used to be collected from many studies and compared.

It follows that the entire body of text to be mined must be structured in the same way. If the text is in XML then it must conform to the same Document Type Definition (DTD). The DTD contains the explanation of the tags used in the XML and other structural elements. It is needed for the computer to be able to interpret the XML.

If the XML is in different formats then it must be normalised or converted to a single format before it can be mined. This presents two problems to the researcher. They must either limit their search to content that is structured in the same way or they must copy the source material and normalise it themselves. There is no standard DTD for scholarly publishing, each publisher has their own and although these are

often freely available to all, no standard has emerged. Copying of content that is subject to copyright requires the clearing of rights and permissions to do this. For these reasons the body of text that is most used by researchers for text mining is PubMed. Note that PubMed allows unrestricted text mining to abstracts only. It is all in a consistent DTD, it is relatively large, and a signed license is not needed (although users implicitly agree to abide by NLM terms & conditions).

The lack of availability of XML content has led many researchers to use PDFs as their source material. This is certainly not ideal because the PDFs require conversion before they can be used for text mining and this conversion can introduce errors. However, as researchers point out: PDFs are better than nothing at all. If the HTML is available alongside the PDF then text mining can be done more successfully. Note that if the content sources are available in a common format, then normalisation is not required. So if access is given through an API then text mining can be done directly using the curator's (usually the publisher's) platform (database) without the need to copy the content to a separate location.

The researcher must take into account the technical and the rights aspects of each source they have selected to mine. Clearly, the more content providers there are, the more time-consuming this is. DTD conversions must be written and tested for each source with a different DTD. The rights have to be cleared as well. In some cases, such as PubMed, these are implicit, but other rights holders will require a click-through license, and some will require written license agreements.

Most publishers report that all requests from bone fide researchers to mine their content are approved. Some publishers provide an XML feed in common formats.

ORPHAN WORKS

Orphan works amongst scholarly journals are virtually unknown. Ownership of journals and their titles do change regularly but the introduction of ISSNs in 1971 makes identification relatively easy. Moreover, the vast majority of scholarly journals were launched within living memory. Orphan works can be a problem with older books with multiple authors, especially pre-1970 before ISBNs were introduced. The older the book, the harder it is to accurately identify it, and therefore know for certain who holds the rights. A British Library study that investigated mass digitisation of books[1] showed 21 of a sample of 101 books published between 1970 and 2010 were orphan works, but this study was not limited to scientific literature. Estimates for STM books suggest a figure of under 10%.

1 "Seeking New Landscapes: A rights clearance study in the context of mass digitisation of 140 books published between 1870 and 2010", Barbara Stratton, British Library 2011, http://pressandpolicy.bl.uk/imagelibrary/downloadMedia.ashx?MediaDetailsID=1197

It is not always clear however who the rights holder is, nor how to contact them to seek permission. Even when it is clear, the process may be time-consuming to approach them all when the body of content is large and from diverse sources. Furthermore, those responsible for granting rights and permission may not be familiar with text mining and may not know how to deal with the request. It is known however that several organizations, including PLS, CrossRef, and CCC, are working on enabling services in this area.

The problem for researchers wishing to mine across sources from multiple publishers is well illustrated by the list assembled for the Genocoding Project.[14] The list highlights the differences in policy and procedures for each publisher. In many cases, permission was eventually obtained, usually through direct contact with publishing staff.

Even when permission is granted, the access terms and conditions may be problematic. Some publishers request that text mining crawlers leave 5 or 10 second delays between successive article downloads so that the crawlers are not treated as robots and blocked automatically. This sounds reasonable enough until the scale is taken into account. A collection of one million articles would take 4-8 months of continuous downloading.[15]

14 The Genocoding Project http://text.soe.ucsc.edu

15 Assuming download times of 5-10 seconds per article, and ignoring downtime. For comparison 1.4-1.6 million scholarly journal articles are published worldwide each year.

Some publishers make their content available on third-party hosting systems such as HighWire Press. This adds an extra complicating step since even after permission has been granted, the hosting service must be set-up to allow access to the text miner.

7.2 Information Extraction

The second ingredient for the text mining recipe is extraction tools. There are numerous tools available, many of them open source.[16] More are under development as researchers strive to extract more meaning from text. These tools interpret the meaning of text, identify and extract out concepts and the relationships between the concepts. There are also many commercially available tools and tutorials on how to use them.

There are two basic steps needed to extract information. Firstly, the tools must be able to recognise structures in the text, for example sentences. They must recognise words that are things, and sets of words that describe a concept. They recognise verbs that suggest relationships between concepts, and they recognise morphological variants of words. In this first step, the text is sliced, diced, chunked and tagged into structured formats ready for extraction into a structured database. In essence, this first step deconstructs the human language of text and reconstructs it for machines.

The second step is to extract information from the text. This can be as simple as recognising terms such as gene sequences, but more often includes recognising named-entities such as names of people, and fact extraction such as the relationships between terms or entities. The results are presented as facts or assertions and collected in a database. The starting point for this step is usually a template that describes the generalised form of the information to be extracted. An example would be a certain pattern of words that occur in a certain way. The skill is to design templates that will extract meaning in the form of facts from the text. A simple example from chemistry is a pattern where two chemicals, A and B are found close to the phrase “reacts with”.

It is also possible to have the computer analyse a body of text for patterns and suggest these as templates. This can be very powerful. For instance, there are applications that extract patterns from a document or set of documents, which are then used to drive a sophisticated “more like this” search, which is based on meaning rather than keywords.

One of the easiest ways to see text mining in action is to paste a sample of text into the Document Viewer of OpenCalais.[17] It is a simple, yet effective, demonstration of the basic process described above.

The choice of which tools to use depend on the goal of the text mining effort. Text mining applications will often bundle a set of tools together to provide out-of-the box functionality.

Commonly used tools are as follows:

- Named entity tools recognise words as entities based on dictionaries.
- Word disambiguation tools identify the correct meaning when there are multiple possible meanings for the word.
- Part-of-Speech (POS) taggers recognise and mark words as corresponding to a particular part of speech. In other words, it identifies nouns, verbs, adjectives, adverbs and so on. This is important in text mining because it allows relationships between concepts to be determined.
- A parsing tool determines the grammatical structure of the text and turns it into something that is machine readable. In other words it identifies sentences and phrases.
- Stemming is used to remove common endings to words reducing them to the same stem throughout the text e.g.: “gene” and “genes”. Stop word removal takes out words like “the” and

16 See for example the following lists of text mining tools :http://arrowsmith.psych.uic.edu/arrowsmith_uic/tools.html; http://www-nlp.stanford.edu/links/statnlp.html; http://www.nactem.ac.uk/software.php.

17 OpenCalais http://viewer.opencalais.com.

"a". Tokenization breaks up text into pieces called tokens, usually words or symbols. This is a little more complicated than it sounds, for instance, it may be desirable to retain punctuation for a word like "isn't". ". Tokenization throws away punctuation and would normally turn this into two tokens "isn" and "t". Rules have to be developed to decide the correct tokens to use.

- Sentiment analysis tools extract opinions, emotions and sentiment words from text and classifies them as positive, negative or neutral. The simplest approach uses a so-called Bag-of-Words method to scan the text for frequency of positive and negative words. More sophisticated tools use other semantic analysis and natural language processing techniques. The National Centre for Text Mining (NaCTeM) has a test site for a basic sentiment analysis tool that is open to all to try out.[18]

The result of text mining is a database of facts that have been extracted from the text.

7.2.1 Problems with Information Extraction

The availability of text mining tools and freely available video tutorials on how to apply them has made text mining significantly easier in recent years. Biomedicine and chemistry lead the way in the development of text mining tools in academic research. However, there can be errors in the information extraction steps. Entities may be missed, or relationships extracted where none existed in the text (false positives). Text mining tools and applications are assessed according to common evaluation standards, such as "precision", "recall" and "F measure", and using standard content sets, sometimes known as a gold-standard corpus.

Alongside the tools, appropriate vocabularies and semantic models, such as dictionaries, thesaurii, taxonomies, and ontologies must be acquired or built, and maintained. Domain experts are usually needed for this.

The information extraction is a multi-step process and relatively complex. It is not always obvious to domain experts how the results of text mining have been distilled from the underlying text. This is less of a problem when the text mining is being done as part of computational linguistics research, but it may be significant when the goal is to create databases that can be mined for discovering new insights in specific scientific disciplines. The recent JISC study notes that this could discourage researchers from using text mining. It is also expensive and there is often no clear business case since the outcome is uncertain.[19]

8 How can Publishers Support Text Mining?

Scholarly publishers report that the number of requests for text mining permissions from researchers is still relatively low, but most of them do expect this number to grow. What can publishers do to make text mining of their content easier?

As a rights holder the publisher must give permission for text mining. This can be done in a number of ways. Permission can be included in an access license agreement with, for instance, an institution. The International STM Association have produced a model clause for this purpose.[20] Some publishers have established a process for individual researchers to obtain permission to text mine with some restrictions,[21] while others do not support text mining yet. Some organisations such as PubMed, allow unrestricted text mining without permission, although note that this applies to the abstracts only.

18 NaCTeM Sentiment Analysis Test Site http://www.nactem.ac.uk/opminpackage/opinion_analysis

19 "*The Value and Benefits of Text Mining*", JISC, March 2012 http://www.jisc.ac.uk/publications/reports/2012/value-and-benefits-of-text-mining.aspx

20 STM Statement on Text and Data Mining and Sample License http://www.stm-assoc.org/text-and-data-mining-stm-statement-sample-licence/

21 See for example Elsevier http://www.elsevier.com/wps/find/intro.cws_home/contentmining or Springer: http://www.springeropen.com/about/datamining/

The Pharma-Documentation-Ring (P-D-R) recently updated their sample license to grant text and data-mining rights for the content to which each of the P-D-R members subscribes.[22]

In their recent study Smit and van der Graaf[23] report that over 90% of the publishers in their survey grant permission for the research-focused mining requests they receive. 32% of the publishers allow for all forms of mining without permission, mostly under their Open Access policy.

Permission is granted under defined terms and conditions of use that are usually detailed in the license. This could be a standard creative commons license or one designed specifically for a particular purpose.

The process of obtaining or granting permissions for text mining is daunting for researchers and publishers alike. Researchers must identify the publishers and discover the method of obtaining permission for each publisher. Most publishers currently consider mining requests on a case by case basis. As text mining grows, publishers may find it challenging to cope with a large number of requests.

One way forward would be a clearing house for permissions that provides a single point of contact for researchers and publishers. There is an initiative from the Publishers Licensing Society (PLS) to develop this concept further.

Some sort of license is needed to verify permissions and to enable access to the content for the purposes of text mining. Ideally, this would be a standard "click-through" license that is simple and easy to implement. A machine-readable license would enable every article with a DOI to have this license associated with it which would greatly simplify the whole process. A researcher would accept and receive a software certificate or API key that would work across all content with a DOI.

The period of time that a license should cover will depend on the text mining need. For computational linguistics research, often a one-time access will be sufficient. All that is needed is a large enough body of text to work on. However, for systematic literature reviews and data mining, it is clear that access will be needed over an extended period as new content is added all the time.

Permissions and licensing is only a part of what is needed to support text mining. The content that is to be mined must be made available in a way that is convenient for the researcher and the publisher alike.

Content may be delivered as a single delivery (a so-called "data dump") or online access may be granted. Publishers may choose to allow robot crawling of their digital content, possibly with restrictions. Many content platforms have security features enabled that shut off access if an unauthorised robot attempted to systematically download content. In this case, the researcher's robot needs to be allowed as a exception. It is also possible to use API (application programming interface) to allow real-time access using standard formats. APIs also allow third-party applications to be built on top of the content.

CrossRef has proposed using DOI content negotiation to implement a standard API for accessing full text on the publishers web sites. There would also be a way for researchers to accept click-though licenses and receive software certificates or keys that show that they have accepted the license. The publisher would then use their own platform and tools to verify that the user was coming from a legitimate source and to deliver the content via their standard, built-in DOI-based mechanism.

PDFs can be used as a source for text mining, and this may be an easy way for many publishers to support text mining. It would be more useful for the researcher the HTML is made available alongside the PDF. More useful still is content that can be delivered in a more structured format such as XML. It is even better if that XML can be delivered already in an already widely-used format such as the PubMed DTD.

22 ALPSP, P-D-R and STM Joint Press Release http://www.stm-assoc.org/2012_09_12_PDR_ALPSP_STM_Text_Mining_Press_Release.pdf.
23 "Journal Article Mining, A research study into Practices, Policies, Plans.....and Promises". Commissioned by the Publishing Research Consortium by Eefke Smit and Maurits van de Graaf, Amsterdam May 2011 http://www.publishingresearch.net/documents/PRCSmitJAMreport20June2011VersionofRecord.pdf.

Publishers can choose to text mine their own content in order to make it more useful to researchers. A number of STM publishers have used text mining to enrich content and to build linked data applications.[24] Text mining by publishers could also be used to create mash-ups where content from related sources is presented together, or to power semantic searching. Publishers could go a step further so that users would have the ability to do semantic queries on concepts extracted from the text. This could potentially be of great value to researchers since they would be saved the trouble of designing, building and testing their own text mining tools.

8.1 Challenges with Text Mining

One of the challenges that publishers face is how to support text mining and other digital content syndication in today's real-time economy. Researchers expect real-time interactions and immediate access in the same way as any internet resource. For instance, Twitter will issue API keys immediately upon completion of a web form.

Publishers, as noted earlier, are willing to grant requests from bone fide researchers to mine their content. The problem is how to verify the credentials of a researcher when they walk up to the text mining front door? It is now possible to uniquely identify researchers using the new ORCID identifier[25] but this system is not designed for authentication.

Fear of the legal consequences has led to a very cautious, conservative approach to text mining from all concerned. There are significant legal uncertainties surrounding text mining and there is no consensus on how best to deal with them. The recent Hargreaves report[26] recommended that text and data mining be excepted from UK copyright, but this would not remove the legal uncertainties.[27]

Open Access models support text mining as long as the terms and conditions allow for systematic or bulk download of content, and the subsequent re-use of the content. However, even this route is not without its challenges. The results of text mining be attributed to the source material, as many licenses such as CC-BY demand.

LEGAL UNCERTAINTIES

Copyright law is clear that permission is required from the copyright holder in order to do text mining that involves the reproduction of a substantial part of a copyright-protected work. There is uncertainty as to what happens when the rights holder is unknown or has disappeared, so-called orphan works. When permission is given, it is important to consider what are known as "derivative works". Text mining of content may frequently result in the creation of databases of facts or raw data extracted from the sources mined. It is not entirely clear whether any resultant database is protected separately as a derivative work. Furthermore, if new knowledge arises from the data mining of that database, is that work derivative? Some licenses cover derivative work but require attribution of the source. It is not always clear how this attribution should be done in practice.

The greatest challenge for publishers is to create an infrastructure that makes their content more machine-accessible and that also supports all that text-miners or computational linguists might want to do with the content.

Finally, it is worth noting that some publishers are pursuing advanced authoring tools that would make it easy for authors to insert semantic information at the time of writing. This has the potential to remove the need for text mining completely.

24 See Elsevier's Article of the Future http://www.articleofthefuture.com and Smart Content http://river-valley.tv/smart-content-at-elsevier/; Content Enrichment from Wiley http://river-valley.tv/content-enrichment-a-publishers-perspective/; Ref Enhanced articles from Royal Society of Chemistry http://www.rsc.org/Publishing/Journals/ProjectProspect/

25 See ORCID http://about.orcid.org and "ORCID: a system to uniquely identify researchers", by Haak et al, Learned Publishing, Volume 25, umber 4, October 2012 , pp. 259-264(6) DOI: http://dx.doi.org/10.1087/20120404

26 *"Digital Opportunity, A Review of Intellectual Property and Growth, An Independent Report"* by Professor Ian Hargreaves, May 2011, http://www.ipo.gov.uk/ipreview-finalreport.pdf

27 *The Value and Benefits of Text Mining*, JISC, http://www.jisc.ac.uk/publications/reports/2012/value-and-benefits-of-text-mining.aspx

9 Case Studies and Illustrative Examples

The following case studies are example of real-world applications of text-mining.

9.1 SureChem (http://www.surechem.com)

SureChem is a search engine for patents that allows chemists to search by chemical structure, chemical name, keyword or patent field. Typical uses for SureChem are to check if particular compounds have been protected (and thus may or may not be patentable), or to identify new or unexplored types of compound which may be candidates for research projects. It is available as a free service with some restrictions on use, and as a paid service for individual and enterprises.

SureChem is based on text mining of the chemical patent literature. The tools identify chemical names in the full text of patents and translates them into structures that can be searched. The text mining process is described in some detail in a three-part blog series by James Siddle.[28]

SureChem uses as its sources a normalized and curated database of patents from IFI Claims® and also MedLine. The first step in their process is to annotate the text using an entity extractor. The extractor uses a combination of statistical and dictionary methods to identify and tag chemical names, and to extract the names so that then can be converted into chemical structures.

Many patents are digitized using Optical Character Recognition (OCR) and this introduces errors in the chemical names that have to be corrected. The structures are standardised and then mapped back into the patent database.

The SureChem database is accessible in three ways. There is a web portal, a web service using API and as an in-house database. The latter is valuable for pharmaceutical companies for example who want to do their searches privately behind their firewall.

The text mining in SureChem is not especially complex, and certainly not cutting edge from a tools and technique perspective. It is however not an easy task to mine this data with sufficient accuracy to be able to sell the service. It is an excellent example of how public data can be mined and value created. SureChem customers pay for the service of being able to search patents using chemical structures and not for the content.

SureChem are looking to add other sources of data, for instance journal articles and to extend into biology and perhaps further.[29]

9.2 BrainMap.org

BrainMap is a database of published functional and structural neuroimaging experiments. The database can be analysed to study human brain function and structure.

Many scientific articles have resulted from analysis and data mining of this database that could not have been performed in any other way.[30] A suite of tools have been built on top of the database that help researchers in this field. It has also provided semantic data that can be used as ontologies and dictionaries for text and data mining tools in neuroimaging.

28 "*Mining Patents in the Cloud Parts 1, 2 & 3*", James Siddle: http://www.digital-science.com/blog/posts/mining-patents-in-the-cloud-part-1-the-surechem-data-processing-pipeline; http://www.digital-science.com/blog/posts/mining-patents-in-the-cloud-part-2-amazon-web-services; http://www.digital-science.com/blog/posts/mining-patents-in-the-cloud-part-3-design-issues.

29 "Take My Content Please!", Nicko Goncharoff, http://river-valley.tv/take-my-content-please-the-service-based-business-model-of-surechem/.

30 BrainMap Publications (1994-present) http://www.brainmap.org/pubs/.

BrainMap was conceived in 1988 and the database was built up manually over 20 years by extracting the Functional MRI data from publications. It is developed at the Research Imaging Institute of the University of Texas Health Science Center San Antonio.

Had the computational linguistics technology in 1988 been at the level that it is today, this extraction could have been assisted by text mining and would have saved much time and effort. As such, it is a good example of the kind of research that text mining can enable. Neurosynth is a recent project that uses an automated approach to achieve essentially the same result based on around 5,000 journal articles.[31]

9.3 Relay Technology Management Inc. (http://relaytm.com)

Relay Technology Management Inc. is a company that uses text mining to create information products for pharmaceutical and biotech companies. It is a useful case study since it demonstrates that valuable information can be extracted from abstracts alone without the need for full text.

The source for their text mining are chiefly Medline abstracts and the patent literature. They believe that abstracts contain the necessary information to extract meaning. There is also a trade-off between the effort of mining a much larger body of text (full text would mean 10 to 20 times more text to mine) and the improvement in precisions and recall by mining full-text instead of the abstracts. Abstracts have certainly proven sufficient for them to build a business as commercial text miners.

Relay extracts drug and disease names as entities from the abstracts. They mine for various relationships between these entities but find that relatively simple co-occurrence of a drug and a disease in the same abstract is already very valuable. However, for this to be successful, the entity extraction has to be accurate and precise. They employ a combination of statistical and semantic methods using ontologies to achieve this.

In addition to abstracts, they also mine other freely available sources such as press releases, NIH grant applications and so forth. Relay also license some content sources in order to support specific information discovery needs.

Relay have developed a suite of data mining products that sit on top of their database and provide their customers with trend analysis and alerting services. For instance, customers can be alerted whenever drugs or diseases they are targeting are reported on. Note that since the alerting is based on text mining, this goes way beyond a simple keyword search. The drug and disease entities are identified in all the possible ways that they can be described. The trend analysis looks for example at hot spots of research into certain diseases, or a more general search into drug classes or disease types. Individual researchers and institutions can be associated with certain research trends.

Relay have published two case studies to show how their tools can be used. One identified early stage programmes in bacterial infections (pre-clinical, phase I and II clinical trials) as suitable for licensing. The other identified top research investigator, leading institutions, and emerging disease etiologies for Duchenne Muscular Dystrophy.[32]

31 Neurosynth.org beta http://www.neurosynth.org/
32 Relay Case Studies: http://relaytm.com/software_biotech_pharma_tech_transfer/case-studies/

10 Popular Misconceptions and Confusions

"Text mining" is a misleading term, partly because it is wrong way round. Text mining is the mining of text looking for meaning. Compare this with gold mining, which is the mining (of earth) looking for gold. There is the potential for even more confusion if the text that is mined is made up of numbers rather than letters. This is the case, for instance, when tables in scientific articles are mined for the information they contain. Nevertheless, the mining metaphor is useful since the process is indeed the extraction of valuable information.

The difference between text mining and data mining is somewhat blurred when statistical analysis is used to extract meaning from the text. Indeed, one could argue that from a computer's point of view text mining and data mining are very similar, especially since in a formal sense text is a form and data, albeit unstructured. Statistical text mining tools look for trends and patterns in the words, and use them to extract concepts from the text. It may be helpful to refer to think of this as using data mining techniques for the purposes of mining text.

"Text and Data Mining" is used mostly as a collective term to describe both text mining and data mining. *"Content Mining"* has also been used as an umbrella term to cover the whole field.

"Big data" is another term that is sometimes associated with text mining. This is logical since although it is theoretically possible to mine a single article, it is usually more useful to mine databases of large numbers of articles. The scope of big data goes far beyond scholarly literature of course, and mostly describes data in existing databases, such as research data, that was never text mined.

It is perfectly possible to use PDF files as source material for text mining. Pdf converters available are widely available.[33] Researchers have even resorted to copy and pasting from PDFs. It helps if the PDF are structured or tagged. However, the tools are not very good, and the manual part is time-consuming and error prone. Nevertheless, it may be good enough for research in computational linguistics, especially if it opens up a new source of content that has not been mined before. For instance, the text mining tool, Reflect, was designed primarily to tag proteins in PDFs.[34]

There is no consensus on the *overlap between Computational Linguistics and Natural Language* Processing. Some use the terms interchangeably, others argue that one is a sub-set of the other. No-one agrees on where to draw the line between them. For the sake of clarity, in this paper the term "Computational Linguistics" is used to refer to the scientific field that uses computers to study language. "Natural Language Processing" is used to describe the development of tools and techniques for language analysis.

11 Some further reading

Witten, I.H. (2005) *"Text mining." in Practical handbook of internet computing*, edited by M.P. Singh, pp. 14-1 - 14-22. Chapman & Hall/CRC Press, Boca Raton, Florida.
http://www.cs.waikato.ac.nz/~ihw/papers/04-IHW-Textmining.pdf.
The Stanford Natural Language Processing Group http://www-nlp.stanford.edu/index.shtml.
National Centre for Text Mining (NaCTeM) http://www.nactem.ac.uk.
The Arrowsmith Project http://arrowsmith.psych.uic.edu/arrowsmith_uic/index.html.
"Journal Article Mining, A research study into Practices, Policies, Plans.....and Promises". Commissioned by the Publishing Research Consortium by Eefke Smit and Maurits van de Graaf, Amsterdam May 2011
http://www.publishingresearch.net/documents/PRCSmitJAMreport20June2011VersionofRecord.pdf.

33 See here for examples of PDF converter and information extraction: http://en.wikipedia.org/wiki/List_of_PDF_software ; http://www.download32.com/information-extraction-tool-software.html; and a helpful review article "Information Extraction Tools for Portable Document Format", Pitale & Sharma, Int. J. Comp. Tech. Appl., Vol 2 (6), 2047-2051 http://www.ijcta.com/documents/volumes/vol2issue6/ijcta2011020648.pdf.

34 Reflect http://reflect.ws.

12 Glossary

Application Programming Interface (API) is a way for computers to communicate with each other. In the context of text mining & scholarly publishing, API usually refers to a Web API expressed in XML. Web APIs are a way of sharing content and data between two computers without either one having to know anything more than the XML language used in the API and its rules. A simple API might be one that returns the PDF of an article when queried with a DOI.

The *Bag-Of-Words* model is a method of representing a document as a collection of words. It is often used in tools that use the frequency of occurrence of words to classify or compare documents.

Computational linguistics is the scientific field that uses computers to study language. Natural Language Processing (NLP) is an inter-disciplinary field that aims to help machines understand written language in the same way as humans do. NLP is used to build text mining tools that recognise words and linguistic structures, and to extract meaning.

Concepts are sequences of words that represent something meaningful. An example could be “serious side effects” or “minor side effects”.

Content Negotiation is a mechanism that makes it possible to deliver different versions of a document at the same web address. For example, the same document in different languages could be served up into the browser based on the location of the user, even though the URL is the same for all users.

A *Controlled vocabulary* is an organised list of words and phrases. It typically includes preferred and variant terms and covers a defined domain, such as chemistry or life sciences.

Copyright Clearance Center (CCC) is a global rights broker. It provides licenses to academic institutions, businesses and other organizations for the rights to share copyrighted material, while compensating authors, publishers and other content creators for the use of their works.

A *Corpus* is a large and structured set of texts.

Crawling is the process whereby a computer systematically retrieves information, often web pages.

Creative Commons (CC) is a nonprofit organization that has developed free, easy-to-use copyright licenses. CC license provide a simple way to give permissions to share and use creative work. Creative Commons licenses are not an alternative to copyright. They modify copyright terms in standard ways. Two commonly used CC licenses are: Attribution (CC BY) which allows for distribution and remixing of the content as long as credit is given to the copyright holder; and Attribution Non-Commercial (CC BY-NC) which allows the same rights as CC BY but on a non-commercial basis.

CrossRef is a collaborative reference linking service that functions as a sort of digital switchboard. It provides an infrastructure for linking citations across publishers using DOIs. CrossRef is a not-for-profit association of scholarly publishers.

Data mining is the extraction of trends and patterns from data.

A *Dictionary* is a controlled vocabulary that includes the meaning and usage of the words and is ordered alphabetically.

Digital Object Identifier (DOI) name is an identifier of an object such an electronic document. Unlike a URL, it identifies the object itself and not the place that it is located. STM publishers assign DOI names to journal articles and book chapters through CrossRef.

Document Type Definition (DTD) is a list of tags, their meaning and the rules for using them. A valid XML document must conform to the rules of a DTD associated with it. A well-known and widely used DTD in STM publishing world was created by the National Library of Medicine as a common format for medical journal articles. It is usually referred to as the NLM DTD.

An *Entity* is a word or phrase found in text that refers to a real-world thing, for instance drug names, or diseases, or chemical names. Also known as *Named Entity*.

Extensible Markup Language (XML) is a system for tagging a document so that it is readable by computers and humans. The content and structure of the document is described by tags. For instance, <title>Text Mining & Scholarly Publishing</title> could indicate that "Text Mining & Scholarly Publishing" is the title of this document.

F measure is an average of precision and recall, which is sometimes weighted.

A *Gold Standard Corpus* is a corpus of manually annotated texts that are considered correct. It is used for testing NLP tools and can also be used as the basis for machine learning.

The *International Association of Scientific, Technical & Medical Publishers (STM)* is a trade association for academic and professional publishers worldwide. STM members include learned societies, university presses, private companies, new starts and established players.

An *Ontology* is a controlled vocabulary expressed in a formal structured way that is machine-readable.

The *Pharma-Documentation-Ring (P-D-R)* is an association of scientific information departments of twenty-one pharmaceutical companies.

Precision is the fraction of the documents retrieved that are relevant.

Publishers Licensing Society (PLS) represents the interests of publishers in the collective licensing of photocopying and digital copying. Through an agent, they issue collective licences to organisations across the UK.

Recall is the fraction of the relevant documents that are successfully retrieved.

Remixing in the context of Creative Commons is taken to mean an alternative form of the content. Some have chosen to consider text mining as a form of remixing.

Resource Description Framework (RDF) is the standard set of specifications that provide a way for information to be encoded with meaning and thus create the semantic web. RDF uses triples to describe the meaning.

The Semantic Web is used to describe a future where information on the web includes descriptions of the meaning of that information. In this way, the information can be read by humans and computers.

A *Taxonomy* is a collection of controlled vocabulary terms organised in a hierarchical structure. Each term has a parent-child relationship to other terms. For instance, western gorillas belong to the genus Gorilla in the Hominidae family.

Text mining is the extraction of meaning (facts and opinions) from a body of text.

A *Thesaurus* is a collection of controlled vocabulary terms that are grouped and linked together according to the similarity of meaning.

Triples are a language or syntax to express facts in a standard format, namely: <subject> <predicate> <object>. The predicate (sometimes known as a property) expresses a relationship or modifies the subject in some way. Note that these terms are not used in strictly grammatically correct, so the object may be more of a target or a value. For example, given the following sentence: "My bottle of Meursault went down very well with the grilled turbot last night", the following triples could be derived with the help of some dictionaries: <Meursault> <is a> <white wine>; <Meursault> <is located in> <Burgundy>; <Meursault> <goes well with> <turbot>.

A *Triplestore* is a database of triples that allows for storage and easy retrieval of triple using queries.

Publishing Research Consortium
www.publishingresearch.net
February 2013