Key Issue
Text mining, copyright and the benefits and barriers to innovation

TORSTEN REIMER
Programme Manager,
Digital Infrastructure
JISC

Do you want to cure cancer? It doesn't matter whether your research is about solving one of the grand challenges of humanity or addressing a more humble question – your first step is likely to be looking at what others have done before. Due to the ever-increasing number of scholarly publications (about 1.5 million new articles published every year), building up an overview of any field of study is an extremely time-consuming process. In prominent topics such as cancer research, it is even more difficult: for the last ten years alone, the UK PubMed Central (UKPMC) database lists 312,308 citations with the word 'cancer' in the title – browsing them at the leisurely pace of 85 per day will take you about ten years.[1] And by that time, ten years' worth of new articles on cancer will have appeared. To make such a search even more complex, relevant articles may not feature the keyword 'cancer' and critical information may be hiding in a footnote within a completely unrelated publication. There is huge potential for advancing knowledge by systematically identifying, analysing and cross-referencing existing research, but the work required is prohibitively time-consuming and expensive. Unless we use machines to help us – and that is where text mining comes into play.

What is text mining?

Text mining refers to the process of automatically (i.e. using software) extracting information from text with the aim of generating new knowledge. It does this by structuring the input text in such a way that computers can analyse it and derive patterns within the structured data. Text mining is similar to data mining but is used on texts written by and for humans, whereas data mining looks at information that is already structured, usually sitting in a database, and not necessarily generated by humans (it could be data generated by a research instrument such as a telescope). Text mining offers much more than just a full-text search over a document and also goes beyond a normal web search in that it is not only used to identify a document or passage but to extract relations and create new knowledge from them.

Examples of text mining use

Text mining can be used in all areas of research, but it is particularly prominent in the biomedical field. For example, immunologists used text-mining tools to infer that the drug thalidomide could treat several diseases it had not been associated with before. They used a two-step approach in which they first identified papers on thalidomide that contained concepts relating to immunology. One concept they were particularly interested in was thalidomide's ability to inhibit a chemical, Interleukin-12, involved in the launch of an immune response. A second automated search for diseases treatable through blocking the action of Interleukin-12 revealed several not previously linked with thalidomide, including chronic hepatitis and a type of gastritis.[2]

Using text-mining tools and natural language processing it is not only possible to identify such patterns but also opinions. An area where this can be particularly useful is the analysis of social media. In market research, for instance, a company could monitor Twitter to find out whether customers like the flavour of a new product, or to identify criticisms of the product. An academic study into the London Riots of 2011 found that social media was mostly used by residents to organize their response to the riots and not primarily by rioters.[3]

Copyright and the cost of text mining

Despite its potential, text mining is not without problems. Some are technical (not all file formats are equally suitable), whilst others relate to supporting researchers in the application of techniques and tools. However, the most difficult issue may very well be copyright. In order to analyse publications you have to make a digital copy and then process it – and for that, permission of the rights holder is required. In theory this is straightforward, but how would you actually do this with the aforementioned 312,308 articles on cancer?

The Wellcome Trust has recently looked into this in more detail. It identified 15,757 full-text articles in UKPMC, published since 2000, that mention 'malaria'. About half of these articles are open access (OA), which means they can be text mined without having to ask for permission, but the remaining 7,759 articles are not. They were, in fact, published in 1,024 different journals. The Wellcome Trust have calculated that it would take around two thirds of a working year to negotiate permissions for those articles, representing a cost of £18,630 if a newly-qualified researcher were employed for the purpose.[4] In other scenarios the actual cost would be even higher, as UKPMC contains a comparatively high proportion of OA publications.

> " … it is not difficult to see why some argue that copyright is a significant barrier to innovation."

Looking at this example, it is not difficult to see why some argue that copyright is a significant barrier to innovation. This is not just a view held by text mining researchers. In 2010, the Prime Minister commissioned *A Review of Intellectual Property and Growth* because of a concern that 'the current intellectual property framework might not be sufficiently well designed to promote innovation and growth in the UK economy.'[5] One of the recommendations of this independent report undertaken by Professor Hargreaves was to grant an exception that would allow text mining for non-commercial research. The government broadly endorsed the Hargreaves report and its publication has been followed by a consultation undertaken by the Intellectual Property Office (IPO).

The Value and Benefits of Text Mining study

In order to inform its input into the consultation and to provide evidence and information of the benefits, risks and barriers of text mining to the academic sector, JISC commissioned a study into the value and benefits of text mining. In particular the report looked at costs, benefits, barriers and risks applying to text mining; these are the general themes that emerged:

- costs include access, transaction, entry, staff and infrastructure costs
- benefits include: efficiency; unlocking hidden information and developing new knowledge; exploring new horizons; improved research and evidence base; and improving the research process and quality
- broader economic and societal benefits were also highlighted, such as cost savings and productivity gains, innovative new service development, new business models and new medical treatments
- barriers and risks: consultees in general felt that there were significant barriers to uptake of text mining in UK further and higher education. These include: legal uncertainty, orphaned works and attribution requirements; entry costs; 'noise' in results; document formats; information silos and corpora-specific solutions; lack of transparency; lack of support, infrastructure and technical knowledge; and lack of critical mass[6]

Benefits of text mining

The '*Value and Benefits'* report contains several case studies that exemplify the benefits of text mining and includes a valuation. In the case of a biomedical researcher, it showed that text mining made it possible to increase the coverage of literature surveyed by a factor of 4.17, thereby identifying significantly more links between publications that would otherwise have been unnoticed. Even more time could have been saved in this case by using text mining to automatically summarize research papers. The report concluded that by using such an automatic summarization approach, the higher education sector could save over £370m per year in staff time – if the Hargreaves recommendations were implemented. The authors also calculated that a JISC service that uses text mining to help researchers identify and then access different journal archives could, if used across the sector, save £59.9m worth of academic time. Text mining can generate additional savings and free up more staff time for research. Perhaps even more importantly, it also has huge potential to contribute to innovation by generating new research findings.

"… the higher education sector could save over £370m per year in staff time – if the Hargreaves recommendations were implemented."

Barriers and risks to text mining

In the light of those benefits, it is worth looking at the barriers in more detail, especially with regards to the order in which they must be addressed so as to facilitate a broader uptake of text mining in academia. It is notable that even researchers in expert centres who don't suffer from problems such as lack of support or infrastructure and who have the skills to overcome technical issues are restricted by legal issues in their application of text mining. This also applies to content for which researchers, usually through their libraries, have already paid licence fees, as these fees do not include the right to mine content such as academic journals. Researchers and text mining experts who were consulted said that with the exception of open access publications and content for which they own the rights, text mining was often not practical because of the need to negotiate licences. This would suggest that without a change in the legal situation, it may simply not be possible to satisfactorily address the other barriers to the uptake of text mining.

Following their analysis of the text mining market, the authors of the JISC study have concluded that there is evidence for a market failure – which would justify a legal intervention to guarantee that the right conditions exist for text mining:

'The technological developments underpinning text mining are relatively recent and hence were not envisaged in previous consideration of the impact of copyright. However, because the process of text mining involves the production and storage of copies of material that may be subject to copyright, there is a new conundrum: the market intervention of copyright – originally intended to protect creative producers – may be inhibiting new knowledge discovery and innovation.'[7]

It is hoped that the IPO consultation will inform a reform of UK copyright law that allows text mining to fulfil its potential for driving innovation and growth.

References and notes

1. UK PubMed Central: http://ukpmc.ac.uk/ (accessed and searched 3 May 2012).
2. JISC, *Text Mining*: http://www.jisc.ac.uk/publications/briefingpapers/2008/bptextminingv2.aspx (accessed 4 May 2012).
3. JISC, *Social media 'not to blame' for inciting rioters*: http://www.jisc.ac.uk/news/stories/2011/12/riot.aspx (accessed 4 May 2012).
4. JISC, *The Value and Benefits of Text Mining to UK Further and Higher Education*, 2012: http://bit.ly/jisc-textm (accessed 4 May 2012), 27-28.
5. Hargreaves, I, *Digital Opportunity: A Review of Intellectual Property and Growth*, 2011, 1.
6. JISC, ref. 4, 17.
7. JISC, ref. 4, 47.

Key Issue © Torsten Reimer

E-mail: t.reimer@jisc.ac.uk

To cite this Key Issue:

Reimer, T, Text mining, copyright and the benefits and barriers to innovation, *Insights*, 2012, 25(2), 212–215, doi: 10.1629/2048-7754.25.2.212

To link to this Key Issue:

http://dx.doi.org/10.1629/2048-7754.25.2.212