Text and Data Mining

Reed Elsevier perspectives

Digital publishing at Reed Elsevier

We have three main publishing business:

Legal and Professional – *LexisNexis*

Provides legal, tax, regulatory and news & business information and analysis to legal, corporate, government

B2B – *Reed Business Information*

Provides information and marketing solutions to business professionals and publishes over 100 business magazines with market leading positions in many sectors.

Science and Medical – *Elsevier*

- Elsevier is the world's largest digital publisher of scientific and medical information, our main platform Science Direct carries 11mn articles and book chapters
- Elsevier serves more than 30 million scientists, students and health and information professionals worldwide.
- The Lancet – the world's leading general medical journal is our best known publication.

What do text miners want to do with content?

- **Create apps –** use content to develop research support tools.
- **Index the content –** identify entities and relationships mentioned in the content in order to facilitate the curation and indexing of the content.
- **Research the content –** conduct research in which the content itself is a source to be analyzed – e.g. for human intent detection, for customer sentiment analysis, for understanding relationships between chemical entities.

Who is text mining content published by Elsevier?

29 text mining agreements at present
(10 of which are with European individuals or institutions)

Higher education/research institutions

6 agreements

Corporates

3 agreements

Corporates breakdown:
1 pharma company,
1 engineering conglomerate,
1 digital publisher.

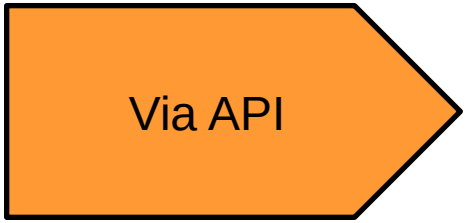
App developers

20 agreements

App developer breakdown:
8 computer science academics,
3 life science academics,
1 health provider,
8 commercial app developers (all SMEs).

How are we giving text miners access?

Under an Elsevier text mining agreement we allow the transfer of content to the user's system via one of two delivery mechanisms:



Real-time, one-document-transferred-at-a-time access to XML/plain full-text via an API.

By ConSyn, a content syndication service that allows text miners to compile a corpus of full-text articles and have it delivered to their systems, typically as a downloadable ZIP file.

These delivery mechanisms have set up and running costs. To recoup these costs our policy to date has been to charge the value add of the services to commercial users but not to non-commercial ones.

Why is content delivery managed this way?

1. Platform Stability

- Providing access to content for text miners via mechanisms like APIs or ConSyn, which are separate from our main online platform ScienceDirect (SD), is critical to maintaining a first class service for our 11 million worldwide readers.

- Those millions of readers expect to be able to access the 2,500 peer review journals and 11,000 books we hold on the SD platform whenever they need to. This corpus represents 20-25 per cent of published scientific research and contains over 11 million journal articles and book chapters. At the moment we are able to deliver 99 per cent up time for the SD platform.

- Allowing multiple users to systematically download from SD on the industrial scale required by text miners would create such a burden for our systems that it could crash the SD platform or, at the very least, slow its performance to the extent that regulars readers would find it harder to access the service. We created APIs and ConSyn to bypass our core platform to avoid this problem.

Why is content delivery managed this way?

2. It's more efficient for genuine text miners

- Text miners tell us that downloading content from our SD platform is not efficient for them either and that they would prefer to receive our content via a different route i.e. APIs or ConSyn.

- Text miners do not want to waste time downloading vast numbers of documents, which could take them weeks or even months to execute. They also often require specially formatted plain/XML text.

- Downloading 1000s of articles would put a strain on the text miner's system too, as well as ours, so receiving documents through an API, or if dealing in 100,000s of articles, via a ZIP file, is preferable for all parties.

Why is content delivery managed this way?

3. Knowing our users

- Giving access via APIs/ConSyn allows us to confirm the bona fides of text miners.

- This is critical because we are transferring, sometimes in a single ZIP file, a substantial proportion of our published content. The piracy risks are obvious.

- To prevent content being illegally redistributed by a rogue user, our standard subscription agreements prohibit systematic substantial retention of articles on the user's systems. Text mining requirements and therefore text mining agreements necessarily differ in that respect (substantial retention is needed by the miner). For that reason we need to understand the individual text miners credentials and purpose before we agree to transfer content.

Next steps – encouraging and simplifying text mining

- **Continue to enable mining for users at subscribed institutions where their purpose is non-commercial without charge.**

- **Actively encourage text mining requests from researchers and institutions to our single point of contact** - universal.access@elsevier.com

- **Set up a pilot automated licensing system.** Researchers at institutions involved in the pilot will have access to a self-service process that gives access to their institution's subscribed Elsevier content through APIs. Under this system it will not be necessary for Elsevier to consider requests to text mine on a case by case basis. The bulk of requests will be considered pre-approved, an automatic licence generated, and access provided through the automated system.

Concerns

An exception would override legitimate restrictions and make online piracy easier

- To prevent platform collapse or content being illegally redistributed by a rogue user, our standard subscription agreement prohibits systematic crawling and systematic substantial retention of articles on the user's systems.
- With our technical and legal defences removed a rogue user operating from within a subscribed institution claiming a text mining purpose could come onto our SD platform and suck out thousands of PDF versions of articles and redistribute them.
- The exception would override these legitimate contractual restrictions and prohibit us from deploying technical measures (e.g. throttling back to prevent crawling or closing access to the relevant IP addresses when we detect suspicious systematic downloading).
- Russian and Chinese based websites offering pirated content are a large and growing problem for the science publishing sector.

Summary

Licensing is working well at Elsevier and will improve further

- Demand to text mine Elsevier content is being met, with no extra charges in the vast majority of cases (26 out of 29 current customers have text mined without charge).
- Additional services to support text mining are likely to be offered as Elsevier improves its understanding of the needs of text miners.

It's early days still

- Text mining demand is embryonic – low numbers at the moment.

The prospect of an exception is high risk to us

- Major risk of increased piracy and unauthorised redistribution.
- Platform stability would be threatened.
- Risk of chilling effect on future service innovation

Text Mining Primer

- **API** (application programming interface) – An interface for a software program that enables interaction with other software, similar to the way a user interface facilitates interaction between humans and computers.
- **Entity** —In text mining, an entity may refer to a group of words, code, statistics or anything else in the document that can provide information. For Elsevier's users, entities of interest often include such things as chemical names, genes, proteins or sequences.
- **Text mining** – The process of deriving information from articles by extracting word patterns and other relationships that could lead to new discoveries.