Advancing Text and Data Mining

Libraries, Researchers, and Elsevier working together

Gemma Hersh
9th June 2015
Outline

• What is text mining?
• How does text mining work?
• Elsevier and TDM
• Addressing concerns
• Industry Initiatives
What is Text Mining?
What is text and data mining (TDM)?

- Process of deriving information from large amounts of text or data content
- Computers automatically extract information like trends, word patterns and other types of relationships
- Uses programs which learn about the text in order to extract the right information (known as natural language processing)
- Potential to lead to new discoveries and resources
Why researchers use TDM

Two main use cases:

1. Answering a specific research question
   - How long does it take for concepts in STM literature to reach general media?
   - What is the relationship between the research and consulting commitments of economics and finance professors?
   - What are the characteristics of subjects in social psychology experiments?

2. Building a new data resource for the community
   - An HIV mutation database for which mutations found in literature are mapped to the underlying sequence database
   - A database on growth and alimentation of fishes, and a fish classification to identify new species for aquaculture
   - A database with the electrophysiological properties of diverse neuron types
A research example

http://neuroelectro.org
How to text mine
TDM can automate the process of discovering relationships between interesting things

Pre-processing | Name entity recognition | Relationship extraction
First you want good, clean content…

**CONTENT**

**INTERESTING THINGS…**
- Enzyme
- DNA
- Response
- Toxin

**RELATIONSHIPS…**
- Enzyme
- Response
- DNA
- Toxin

Pre-processing  |  Name entity recognition  |  Relationship extraction
Published literature?

- The information source that researchers seek already exists.

- Researchers can build on existing results published in scholarly articles.

- It’s a curated and trusted source of information.

- There is a lot of it, it’s cross discipline and it covers archives as well.

Source: Effect of low frequency repetitive transcranial magnetic stimulation on kindling-induced changes in electrophysiological properties of rat CA1 pyramidal neurons. [http://dx.doi.org/10.1016/j.brainres.2015.02.023](http://dx.doi.org/10.1016/j.brainres.2015.02.023)
But online, data is fragmented and of varying quality...

Content on the web ...
...has different formats
...has different structure
...is in different places
...is of variable quality

This is where publishers can help

More on this later...
Next comes identifying the interesting things

Pre-processing

Name entity recognition

Relationship extraction
“The Airbus A380-800 has been a first to use new technology over the skies of London.”
but even with a simple example its not that easy…

Alternative names…
Airbus A380-800
Airbus A-380-800
Airbus A 380-800
Airbus A380/800
Airbus a380
A380-800
…

Abbreviations, irregular spelling, mis-spelling…
Arbus A380-800
Airbus A380800
…
Finally we pull out the relationships...

- **CONTENT**
- **INTERESTING THINGS…**
  - Enzyme
  - Response
  - DNA
  - Toxin
- **RELATIONSHIPS…**
  - Enzyme
  - Response
  - DNA
  - Toxin

Pre-processing  Name entity recognition  Relationship extraction
…though it looks easy, it’s very difficult…

“The **school** has been a **proponent** of using **technology** in lessons. **Every pupil** in the **school** **has** their own **iPad**…”

proponent (school, technology)

has (pupil, iPad)

in (pupil, school)
A typical text mining workflow

**Define the problem**

- What are you wanting to learn?
- What relationships do you need to know?
- What content do you want to cover?

**Define your approach**

- Choose your tools:
  - Off the shelf (open source)
  - Outsource to a specialized company

- Write an algorithm that will extract what you want

- Test the algorithm to ensure it works
  - Use corpus and tree-banking to help

**Access & extract**

- Gain access to the published literature
- Download all the content

- Retest your algorithm
- Run your tools and analyze the error reports
- Re-run if needed

**Analyze & build**

- Process extracted data
- Analyze to see if relationships exist
- Write your results in a paper
- Build the database of extracted values
Elsevier and Text Mining
Supporting TDM at Elsevier

2006 Started supporting *ad-hoc* TDM access requests
- Low - but increasing level - of interest from early adopters

2012 First content mining policy published
- Developed new APIs & rolled out better TDM technical solutions

2013 Pilot (~30 academics)
- Discover two main user cases for TDM and input to refine our TDM technical solutions

2014 Publish updated TDM policy
- Automatically give access to subscribed users
- Launch a developers portal to facilitate access

2015 Launch STM Open Access Corpus
- Valuable open access tool to help researchers refine and develop their tools in order to TDM more effectively
TDM Pilot Learnings – Researcher Challenges

Technical
- Obtaining necessary infrastructure
- Having to deal with different formats from content providers
- Sourcing and understanding TDM technology

Functional
- Fine-tuning pipeline, curating output, representing output meaningfully

Logistical/Legal
- Gaining access to the needed content
- Gaining permission to mine the content
TDM Pilot Learnings – Library Challenges

**Expertise**
- Understanding specific TDM-based projects well enough to assess implications & offer advice to patrons

**Legal**
- Understanding and tracking what is allowed for what resources
- Negotiating permissions with multiple providers
- Ensuring academic freedom is protected

**Financial**
- Concerns about any additional costs
- Understanding how TDM affects usage figures for the library
Elsevier’s TDM policy

- Researchers at academic institutions can text mine subscribed content on ScienceDirect for non-commercial purposes via the ScienceDirect APIs
- Access is granted to faculty, researchers, staff and students at the subscribing institution

Text mining output can be shared publicly under these conditions
- May contain "snippets" of up to 200 characters of the original text
- Proprietary notice
- Must include DOI link to original content

Open access content
- Text and Data mining permission are determined by the author’s choice of user license.
- This information is detailed in the individual articles
How to get access?

1. Register their project at [http://dev.elsevier.com](http://dev.elsevier.com)

What do researchers need to do?

<table>
<thead>
<tr>
<th>My Projects</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>My Registered Websites/Applications</strong></td>
</tr>
<tr>
<td>Registered on</td>
</tr>
<tr>
<td>01 Apr 2014</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>My Text Mining Projects</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Registered on</strong></td>
</tr>
</tbody>
</table>

**Register a new site**

- Register your website or application here.
- [Register a New Site](#)

**Register a new text mining project**

- You can register new text mining projects here.
- [Register a New Text Mining Project](#)
How to get access?

1. Register their project at http://dev.elsevier.com

2. Accept a simple registration form
How to get access?

1. Register their project at [http://dev.elsevier.com](http://dev.elsevier.com)
2. Accept a registration form
3. Obtain an Elsevier API Key
TDM Access: What do Institutions need to do?

• TDM access clause will be part of standard ScienceDirect subscription agreement for new academic customers and upon renewal

• For existing agreements, an add-on contract amendment is available – just contact your Elsevier Account Manager

• After signing institutional agreement/amendment, access to our API key registration page for your researchers will be enabled for your institution’s IP address range

• We facilitate access for researchers at non subscribing institutions too
Addressing concerns
What is an Application Programming Interface (API)?

- An interface for software programs that enables interaction with other software.
- APIs make the TDM process more efficient for vast quantities of data.
- They enable publishers to maintain site performance, availability and reliability for their (human) users.
- Deliver extracted text in an optimal format, preferred by text miners.

APIs are standard across industries:

- Twitter: [https://twitter.com/tos](https://twitter.com/tos)
- PubMed Central: [http://www.ncbi.nlm.nih.gov/pmc/about/copyright/](http://www.ncbi.nlm.nih.gov/pmc/about/copyright/)
Practicalities

1. Registration form
   - Automatic process
   - No additional liability
   - Aligned with institutional e-amendment
   - Provides guidelines on reuse
   - Can offer one-to-one support
   - Removal of project description in response to feedback
Practicalities

2. Mining images
   - We always have done on request
   - Now do automatically
   - Terms of reuse, as we aren’t always the rightsholder
Practicalities

3. TDM outputs

- No copyright claim on original research findings
- We do have output requirements when our material is quoted as follows, in line with standard citation practices:

  • To distribute the TDM Output externally, which may include a few lines of query-dependent text of individual full text articles or book chapters which shall be up to a maximum length of 200 characters surrounding and excluding the text entity matched (“Snippets”) or bibliographic metadata.

  • Where Snippets and/or bibliographic metadata are distributed, they should be accompanied by a DOI link that points back to the individual full text article or book chapter.

  • Where images are used you should clear the rights for reuse with the relevant copyright owner and/or rightsholder.

  • TDM Output should include a proprietary notice in the following form: “Some rights reserved. This work permits non-commercial use, distribution, and reproduction in any medium, provided the original author and source are credited.”

  • Open access content usage terms defined by license
4. Flexibility

- We consider all requests and help all researchers with any queries
- Continued evolution in response to feedback: project descriptor, license outputs, images

http://www.elsevier.com/tdm
Developing NLP Tools

- Elsevier has provided a freely-available and redistributable corpus.

- For each article in the corpus we provide:
  - XML source
  - A plaintext version for easier text mining
  - Several versions with different annotations. (speech tags, sentence breaks, simple noun and verb phrases, root forms of words, syntactic constituents parse trees, Wikipedia concept identification, and discourse analysis)

- Kick-started the process of manually creating test sets,
  - Commissioned a treebank of 10 full-text articles to be used as a default test set.

110 articles
Open access with a CC-BY license

10 domains
agriculture, astronomy, biology, chemistry, computer science, earth science, engineering, materials science

10 test sets

http://elsevierlabs.github.io/OA-STM-Corpus/
Industry initiatives
Aligned with recent STM Declaration on TDM

Supporting cross-publisher TDM

Addressing two issues:

1. **Problem:** Researchers want to be able to access full text content from multiple publishers’ sites for OA or subscribed content in a consistent way.

   **Solution:** Common API (protocol) for requesting machine readable full text from many different publishers.

2. **Problem:** Researchers want to know whether text and data mining is allowed, and if not, get permission.

   **Solution:** Licensing information embedded in article metadata and a registry for supplemental text and data mining terms and conditions (licenses).

http://www.crossref.org/tdm/index.html
Status of CrossRef TDM

• BETA version is live

Participating Publishers

• AIP
• IoP
• HighWire
• Walter de Gruyter
• Elsevier
• KAMJE
• PLOS

• Journal of Zhejiang University Science
• Wiley
• Springer
• APS
• BMJ

13.5 Million
Articles with full test links and license information

Capacity for
36 Million+
DOI records by mid 2015

FREE
Public service
Copyright Clearance Center Text and Data Mining Pilot

• Pilot service to help researchers access and download XML content in a standardized way.
• Commercial service aimed at researchers working in corporate settings

The service is designed to be used in conjunction with existing enterprise text mining software and methods.
Elsevier and text mining

- Actively collaborating with researchers & institutes to facilitate TDM
  - Funding postdocs at the University of Bologna
  - Funding interns at the University of Melbourne
  - Joint research grant with the University of Manchester
  - Funding an NLP Centre of Excellence at Indian Institute of Technology, Bombay
  - Sponsoring academic conferences, e.g.
    http://www.clips.uantwerpen.be/clin25/home

- Enabling access with an updated policy framework

- Providing separate channels for machine-to-machine and human access to content.

- Investing in our platforms to support researchers to text mine and launched a developers portal

- Providing technical support and services beyond the basic content access.

- Working on Natural Language Processing tools; Providing resources such as an open access STM corpus and tree-banking

http://www.elsevier.com/tdm
Thank you!

g.hersh@elsevier.com
@gemmahersh