Illustrating the value of Open Access for text mining

Sylvain Massip, Charles Letaillieur

Illustrating the value of Open Access for text mining

Sylvain Massip and Charles Letaillieur

Open Access – a growing opportunity

- There are many arguments in favour of **Open Access to scholarly publication** in particular allowing a better appropriation of research by practitioners and interested citizens.
- In 2018, **Open Access** represented about **28%** of total published scientific articles[1]
- While the cost of publication is widely studied[2],there are not many reports on the value that open access to scholarly publication can bring
- **The aim of this study is the analysis of the value that Open Access brings for the special case of text mining tools.**

2014 Ebola's outbreak – an emblematic case study

- In 2015, three Liberian healthcare professionals argued that with Open Access, they would have been warned and many lives would have been saved[3]. In response, several authors pointed that Open Access is not enough, and that **research and discovery tools** were also needed[4].
- **Revisiting this case, we study the weak signals that Open Access combined with the appropriate text mining tools could have detected.**

Materials and methods

- Corpuses of **open-access articles** containing "**ebola**" were retrieved from PubMed
- Each entry was constituted of the following:

title|source|abstract|fullText|TextminedTerms

	Fulltext	Abstract	TextMined
Mean	2377	116	16.8
Median	2207	111	12

Table 1. Statistical description of the corpus retrieved from pubmed

Increased classification power

- Using a corpus of **3872 entries**, we apply a supervised Machine Learning approach to **Abstract** and **Fulltext** in order to predict the **TextMinedterms** as a multilabel problem.
- **Tf-IDF vectorization** was combined with **Gradient Boosting classification** from SciKit Learn:

	Precision	Recall	F1
FullText	0.630	0.622	0.626
Abstract	0.324	0.308	0.316

Table 2. Prediction scores for Textmined Terms

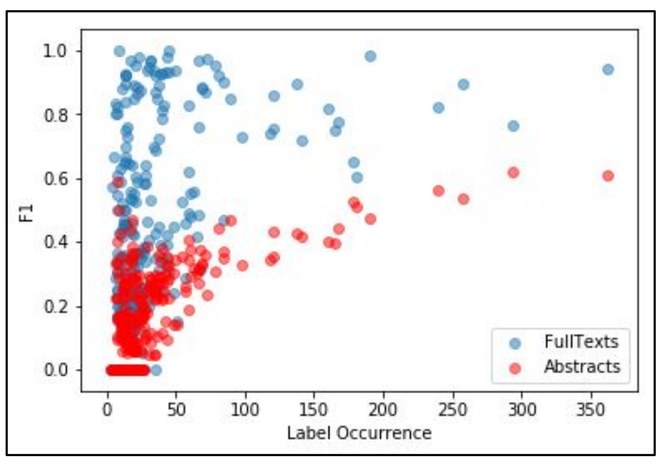

Figure 2. Fulltext vs. Abstracts prediction score

Chronology of Ebola in West African Countries

- We use Elastic Search / Kibana to fine filter the presence of **Liberia**, **Sierra Leone** or **Guinea** in fulltext corpus and in abstract/title corpus
- A **weak signal** can be found **before 2014** in the **fulltext corpus only**.

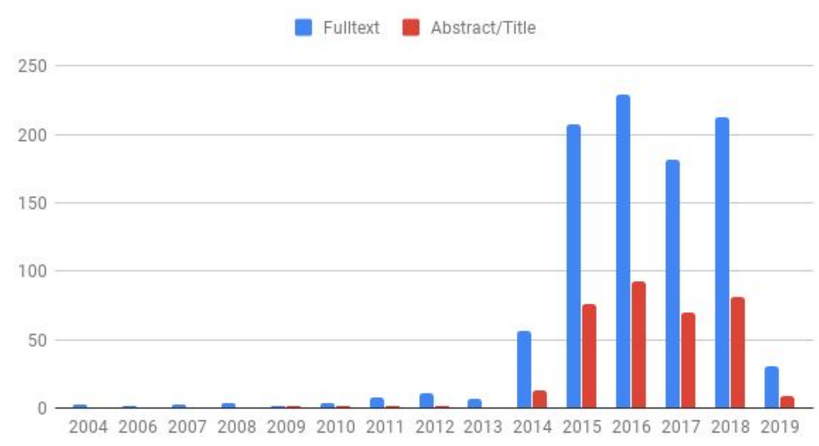

Chart 1.Fulltext vs. Abstract mentioning West African countries through years

"Remanent" Tag Cloud

A tag cloud was drawn from the TextMinedTerms which appear in Fulltext, and not in abstract or title.

This allows interesting concepts to emerge:

- co-occurrent diseases (influenza,hiv)
- environmental conditions (water, temperature)

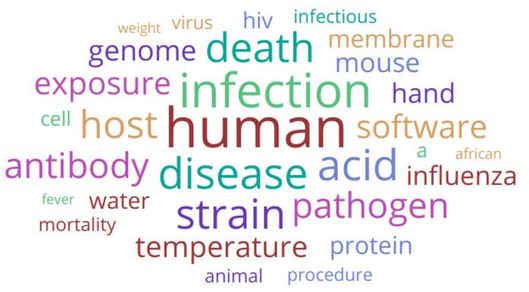

Figure 3. Full text only Tag cloud

Conclusions

Based on the ebola litterature, this initial study exemplifies the **extra value** obtained when **Textmining is applied to Fulltext** rather than **Abstracts only**:

- A much stronger **classification power** is obtained: it becomes possible to predict PubMed's Textmined Terms.
- The **chronology** of the 2014 outbreak is observed, and **an increase in the number of articles dealing with ebola in West Africa is observed before the event.**
- By drawing a tag cloud of the TextMined terms that appear in the FullText only, it is possible **to extract concepts** that it would be interesting to investigate.

A more comprehensive study will be needed to assess more quantitatively these examples, and to see how these results generalize to other case studies.

Contact

OpSciDia
www.opscidia.com
Sylvain Massip
sylvain.massip@opscidia.com
0033 6 28 30 59 20
Charles Letaillieur
charles.letaillieur@opscidia.com
0033 6 50 41 50 99

References

1. Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., … Haustein,S. "The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles." *PeerJ*, *6*, e4375 (2018) https://doi.org/10.7717/peerj.4375
2. Van Noorden, R. "Open access: The true cost of science publishing." *Nature*, *495*(7442), 426–429 (2013). https://doi.org/10.1038/495426a
3. Dahn, B., Mussah, V., & Nutt, C. (2015, April 7). Yes, We Were Warned About Ebola. The New York Times. Retrieved from https://www.nytimes.com/2015/04/08/opinion/yes-we-were-warned-about-ebola.html
4. Crotty, D. (2019). Access Alone Isn't Enough: Revisiting Calls for Discovery, Infrastructure, Technology, and Training. Retrieved from https://scholarlykitchen.sspnet.org/2019/01/31/access-alone-isnt-enough-revisiting-calls-for-discovery-infrastructure-technology-and-training/