

LIA@SimpleText2021: automatic query extraction from press outlets for mining related readable passages in scientific literature

Malek Hajjem¹, Eric SanJuan¹

¹LLA, Avignon Université

Abstract

We have developed a theoretical and software tool to assist science journalists in writing popular science and technology news articles. The public's appropriation of the scientific debate and its consequences on political communication, which is often polarized or partisan, has led us to experiment with multinomial modeling in order to be able to automatically extract the associations of terms characteristic of a newspaper article with general circulation on a given period. To effectively compute these associations we compute the latent Dirichlet distributions (LDA) to a set of multinomials. We use recent implementations available in R and nonparametric correlation tests to verify the significance of associations automatically extracted by the model or formulated as hypotheses by the researcher. The detection of the vocabulary and systematic associations of a political communication is necessary for the study of the possible propagation of the ideas induced in the media, classic (such as the written press), encyclopedic (such as Wikipedia) or specialized (such as DBLP).

Keywords

Language Model, Technical and Scientific Watch, Latent Dirichlet Allocation

1. Introduction

With the growing polarization of the scientific debate, partisans tend to irrigate the terms of the political debate by emphasizing, imposing, diverting or redefining certain themes. This results in the multiplication of semantic shifts, the use of neologisms or sometimes the creation of new terms specific to a group or a personality. The hardening of opinions seems to favor the reinforcement of an existing bias and to multiply the injunctions of governments with regard to scientific research. This increases the difficulty of the comparative study of recent scientific publications as the terms used can echo calls for projects.

In this work, we use probabilistic language models to reveal the underlying concepts as strong multi-term associations to help clarify the links between scientific language and citizen debate by following the methodology presented in [1]. This approach attempts to model the terms of the debate as word probability distributions. It will thus be possible to objectify the proximity of topics in scientific publications from different communities or their circulation in the press. For example, we can calculate indices of adequacy between the themes of a newspaper and the

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania

✉ malek.hajjem@gmail.com (M. Hajjem); eric.sanjuan@univ-avignon.fr (E. SanJuan)

🌐 <https://lia.univ-avignon.fr/> (M. Hajjem); <https://termwatch.es/> (E. SanJuan)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

content of the summaries of a bibliographic database such as DBLP.

Such language modeling that generates an ordered set of notions, concepts and entities, also allows to automatically define temporal queries for social media monitoring. Thus projecting these themes in another media space makes possible, for example, a diachronic follow-up of the themes through the press titles disseminated on social media [2].

2. Methodology

It is made up of two stages described below:

1. For each scientific press article, we compute multiple language models using the word exchangeability property. Scientific and technological news published in the non-specialized press are distinguished from other types of texts such as stories, argumentative texts by their property of exchangeability between their terms: we can change the ordering of their words without disturbing their overall meaning. This seems to be essentially based on the choice of nominal values, their frequency and the associations thus created. This induces the existence of a latent explanatory variable of the dependencies between words. The approximate computation of this variable by latent Dirichlet allocation (LDA) makes it possible to objectify the language elements and the recurrence of the amalgams specific to each editorial line. This in an unsupervised way, only by analyzing their press releases to the exclusion of any other textual, terminological or semantic resource.
2. We then search for associations of terms characteristic of a set of news by using non-parametric tests. The computation of probabilistic models is done by journal and by period. The production of newspaper articles is expensive and too imbalanced between publications. There are also terms that are almost exclusively used by a single media outlet. The search for a single model encompassing all newspapers ran into these two pitfalls. Rather than complicating the probabilistic model, we thus preferred to proceed by subsets of articles and use the Wilcoxon correlation test to determine whether the associations are characteristic or not of the studied sub-corpus. The results of these tests are comparable to each other because we have chosen the same number of dimensions $k = 30$ for all the models sought.

2.1. Probabilistic model

We consider press articles where terms, topics and concepts have been previously discussed with scientists. We assume that the terms to be included in the news article are chosen according to an internal editorial line targeting specific groups of readers. Our approach to revealing the target audience relies on the exploration of significantly correlated sets of terms based on a generative language model and nonparametric statistical tests.

We use a word bag representation where we have chosen to ignore the word order in the news article and focus on the co-occurrences and frequencies of the terms. Indeed, tech news articles differ from argumentative or narrative texts. They are short and often rely on a reduced number of statements. These hypotheses allow us to apply the concept of exchangeability

highlighted by De Finetti ¹ [3] which allows us to assume that existence of a multinational latent variable explaining the dependencies between terms.

2.2. Statistical significance of term associations

For each scientific press article, the bag of words representation induce a multinomial model for the frequency probabilities of the words. The exchangeability hypothesis induces the existence of a latent variable with the same law which is sufficient to determine the probabilities of co-occurrence between words. The LDA approach makes it possible to calculate this latent variable approximately if we know its dimension. The difficulty is that there is no way to infer this dimension as a prior. We used the two measures of Griffiths and Deveaud [4] to compute posterior estimates of the quality of the model obtained for each number of dimensions. In our tests, we found an optimal number that was still slightly less than 30. We then chose to set this parameter to 30, which makes robust the comparison of results of Wilcoxon's non-parametric statistical correlation test.

Thus for each sub-corpus, we force the calculation of a multinomial variable with 30 explanatory modalities of the dependencies between terms. The more the terms are associated, the more their probabilities of being associated with the same modalities will be high. The Wilcoxon test makes it possible to compare nonparametrically two aligned series of probabilities. The statistic gives the degree of correspondence and the p-value the probability that this statistic is not significant. If it is not possible to compare probability vectors between corpora, it is however possible to compare the results of tests carried out on the same number of dimensions.

3. Experiments and results

The search for terms and characteristic associations for each corpus is done interactively using the annotated R program (Rmarkdown) available online ² includes the following features:

1. Selection of the sub-corpus by journal and period.
2. Calculation of the language model.
3. Association test between two terms and exploration of their context.
4. Find and explore all of the terms significantly associated with a target term.
5. Query of the ElasticSearch database and evaluation of the overlap between the content of the press article and that of the specialized database.

We applied this approach to the 13 tech articles from The Guardian shared within the simple text workshop [5] for task 1³ consisting in mining the DBLP database to retrieve relevant citations (short readable sentences from abstracts).

Table 1 shows for each article the nouns w which maximize the probability of being associated to one of the 30 LDA topics:

$$\max \{p(w|i) : 1 \leq i \leq 30\}$$

¹<https://journals.openedition.org/msh/6793>

²<https://guacamole.univ-avignon.fr/pubiutdev/sanjuan/simpletext/LDAsimpletext.rmd>

³<https://www.irit.fr/simpleText/>

Table 1

Queries extracted from SimpleText task 1 journal papers

paper	queries	significant correlations	# passages
1	assistant biases digital	assistant digital	16
2	confidential privacy		13
3	smart speaker surveillance	smart speaker	13
4	discovery diseases drug ...	discovery drug ...	17
5	crispr editing gene	editing gene	16
6	driving self	driving self	15
7	conspiracy misinformation theories	conspiracy theories	14
8	cryptocurrency financial markets	financial markets	17
9	forensics		8
10	disrupting humanoid robots	humanoid robots	11
11	drones		7
12	data patient	data patient	7
13	advertising digital marketing	digital marketing	14

and, when existing, the pairs of nouns which topic distribution is significantly correlated. These results have been shared with the workshop to provide queries to retrieve related passages. The total length of extracted passages using elasticsearch BM25 is also reported.

4. Conclusion and perspectives

We have developed a statistical tool to assist the journalist in extracting characteristic terms from a popular science article. This extractions had been identified in [6] as being the main obstacle to automatic short text contextualization. This lock having been lifted, we will subsequently study the overlap or permeability between popular articles and bibliographic reference databases.

References

- [1] M. Shams, A. Baraani-Dastjerdi, Enriched LDA (ELDA): combination of latent dirichlet allocation with word co-occurrence analysis for aspect extraction, *Expert Syst. Appl.* 80 (2017) 136–146. URL: <https://doi.org/10.1016/j.eswa.2017.02.038>. doi:10.1016/j.eswa.2017.02.038.
- [2] N. Peladeau, E. Davoodi, Comparison of latent dirichlet modeling and factor analysis for topic extraction: A lesson of history, in: T. Bui (Ed.), 51st Hawaii International Conference on System Sciences, HICSS 2018, Hilton Waikoloa Village, Hawaii, USA, January 3-6, 2018, ScholarSpace / AIS Electronic Library (AISeL), 2018, pp. 1–9. URL: <http://hdl.handle.net/10125/49965>.
- [3] E. Lehrer, D. Shaiderman, Exchangeable processes: de finetti’s theorem revisited, *Math. Oper. Res.* 45 (2020) 1153–1163. URL: <https://doi.org/10.1287/moor.2019.1026>. doi:10.1287/moor.2019.1026.
- [4] R. Deveaud, E. SanJuan-Ibekwe, P. Bellot, Accurate and effective latent concept modeling

for ad hoc information retrieval, *Document Numérique* 17 (2014) 61–84. URL: <https://doi.org/10.3166/dn.17.1.61-84>. doi:10.3166/dn.17.1.61-84.

- [5] L. Ermakova, P. Bellot, P. Braslavski, J. Kamps, J. Mothe, D. Nurbakova, I. Ovchinnikova, E. SanJuan, Text simplification for scientific information access - CLEF 2021 simpletext workshop, in: D. Hiemstra, M. Moens, J. Mothe, R. Perego, M. Potthast, F. Sebastiani (Eds.), *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part II*, volume 12657 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 583–592. URL: https://doi.org/10.1007/978-3-030-72240-1_68. doi:10.1007/978-3-030-72240-1_68.
- [6] O. Hamzaoui, T. Jiménez, C. Lagier, E. SanJuan-Ibekwe, Contextualisation du discours politique, *Document Numérique* 22 (2019) 63–84. URL: <https://doi.org/10.3166/dn.22.1-2.63-84>. doi:10.3166/dn.22.1-2.63-84.

A. Online Resources

The source code is available at

- guacamole.univ-avignon.fr,