Data Scraping to the rescue: settling Brazilian public communication of science's debate?

Marcelo Pereira[1]
Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

Roberto Takata
Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

Introduction

Public communication of science is increasingly relevant, not only to democratize knowledge, but as an imperative duty and necessity for science, to guarantee visibility, legitimacy, resources and citizens' trust in scientific institutions.

In this sense, universities and their scientists have increasingly sought to interact with the general public. They have dedicated more time and resources to holding events aimed at non-specialist audiences, dialoguing with civil society, producing audio and video content, serving traditional media and new digital media, in short, in these and in many other public communication of science activities.

One of the challenges to better understand public communication of science is to understand how the scientific production of this domain is situated in semantic terms. This allows us to understand which are the main subjects related to the public communication of science, and, in this way, to monitor, evaluate and induce policies, eventually filling the gaps and promoting dialogues between areas that may be necessary.

Thus, the aim of this research is to compare the public communication of science outputs in six public Brazilian universities: UFMG, UFU, CEFET-MG and UFOP, in the state of Minas Gerais, and Unicamp and UFSCar, in the state of São Paulo. In order to do so we conducted an exploratory analysis of data obtained from public platforms with information about academic outreach activities in those institutions.

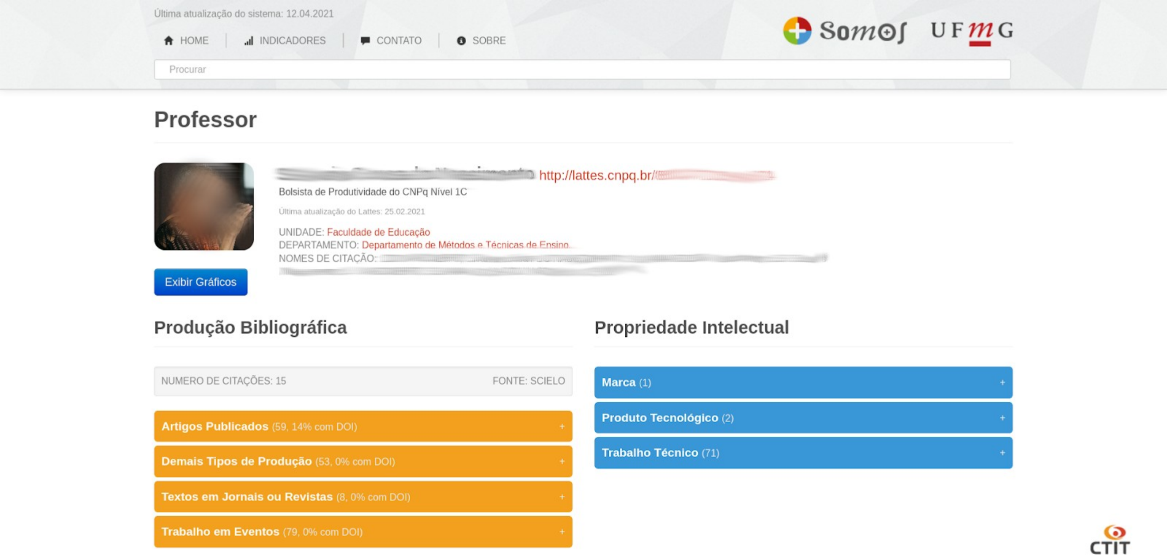
Using data scraping techniques, we collated keywords used by researchers in their papers and the description of science outreach activities and projects, so as to generate a word frequency table and word cloud.

Our hypothesis is that meaning is given by context, which can be revealed by the occurrence and frequency of keywords. If the words that occur in conjunction with the key phrases are similar between the different terms, there is no significant semantic difference between the terms.

[1] Optional contact information (e.g., Corresponding author, email: mymail@myserver.xy.)

Method

We scraped the data available in “SOMOS”, which is a web platform that gathers information about research and researchers from a given brazilian University. Most of the data come from “Lattes”, wich is a nationwide scientists database. There are, currently, about eight brazilian Universities adopting SOMOS, which allows for some comparisons. We resort to web-scraping because there are still no public APIs available for science databases in Brazil.



Professor

[redacted] http://lattes.cnpq.br/[redacted]

Bolsista de Produtividade do CNPq Nível 1C

Última atualização do Lattes: 25.02.2021

UNIDADE: Faculdade de Educação
DEPARTAMENTO: Departamento de Métodos e Técnicas de Ensino
NOMES DE CITAÇÃO: [redacted]

Produção Bibliográfica

NUMERO DE CITAÇÕES: 15 FONTE: SCIELO

Artigos Publicados (59, 14% com DOI) +
Demais Tipos de Produção (53, 0% com DOI) +
Textos em Jornais ou Revistas (8, 0% com DOI) +
Trabalho em Eventos (79, 0% com DOI) +

Propriedade Intelectual

Marca (1) +
Produto Tecnológico (2) +
Trabalho Técnico (71) +

Figure 1: Screen capture of a researcher’s page in SOMOS platform.

We collected keywords related to “science communication” from the scientific production of six Universities. By scientific production, we mean papers published in scientific journals and reviews, presentations in conferences, and outreach activities. Word lists will be analyzed by natural language processing techniques for comparison of text semantics.

It follows that we compared the sets of keywords of each University using two different similarity indexes: the Jaccard index and the Cosine Similarity.

The Jaccard index, also known as the Jaccard similarity coefficient, is a statistic used for gauging the similarity and diversity of sample sets. It measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

Figure 2: The Jaccard index equation

The Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space. It is defined to equal the cosine of the angle between them, which is also the same as the inner product of the same vectors normalized to both have length 1. For text matching, the attribute vectors *A* and *B* are usually the term frequency vectors of the documents. Cosine similarity can be seen as a method of normalizing document length during comparison.

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2} \sqrt{\sum\limits_{i=1}^{n} B_i^2}},$$

Figure 3: The cosine similarity equation

where A_i and B_i are components of vector A and B respectively.

The difference between the two indexes used in this study is that Jaccard index takes only unique set of words, while cosine similarity takes total length of the group of words.

Results

The results obtained by both calculations are similar, as we can see on Figures 4 and 5. In general, what determined the similarity was the size of the vectors, but we can see some nuances. For example, UFMG is more similar to Unicamp than to UFSCAR, although we have collected more keywords from the latter. Or even, UFOP, is more similar to UFSCAR than UFMG, according to the Cosine of Similarity.

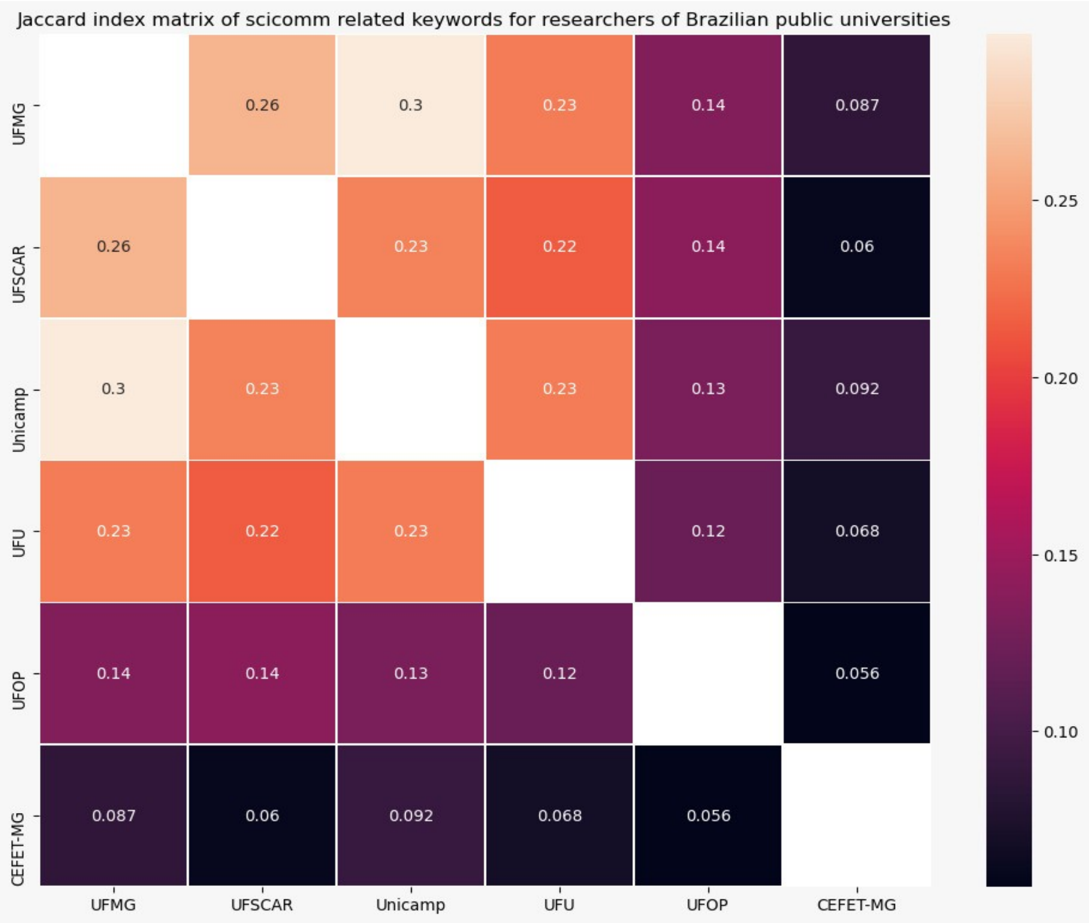


Figure 4: Jaccard index matrix of scicomm related keywords used by Brazilian universities researchers.

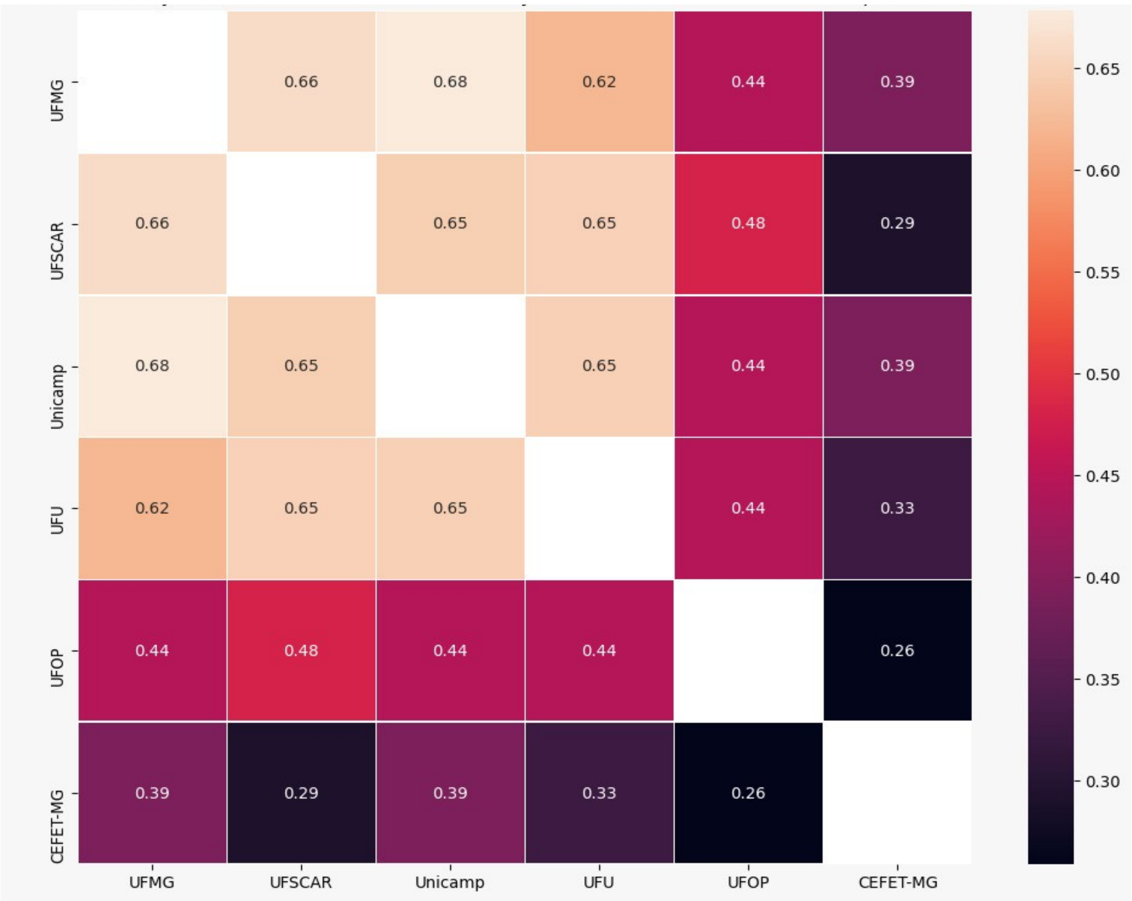


Figure 5: Cossine similarity index matrix of scicomm related keywords used by Brazilian universities researchers.

The closer similarity between UFMG (from Minas Gerais) and UFSCar and Unicamp (both from São Paulo), rather than between UFMG and other Universities from the same state suggests that spatial proximity could be not so crucial to research similarity (keywords associated with science communication). UFMG, UFSCar, and Unicamp are among the largest universities in Brazil and display higher rankings both in national and international evaluation panels (e.g. Ranking Universitário da Folha, QS, and ARWU World University Ranking). A possible question is whether they are more connected with the international research community on science communication. Another possibility not explored in this preliminary study is that there could be more collaboration between larger universities researchers than between minor ones. An analysis of paper co-authorships could shed light on that hypothesis.

A factor that must be examined more carefully in a deeper analysis is the bias that must be present in the database caused by lack of registration and/or misregistration. This effect can be especially important in smaller institutes, where a greater engagement in registration by few researchers can influence the data and lead to wrong analysis.

A common theme among all keyword sets is education and related terms, which suggest a strong association between the two fields. Further analysis could make clear if other terms more directly associated with science communication such as journalism, museum communication, popularization, literacy, awareness, engagement, appropriation, among others (Buchi & Trench, 2016; Barata et al. 2018) are also relevant, being masked by the prominence of educational terms, or if they are of minor importance in Brazilian science communication research.

To conclude, we would like to reinforce that this type of analysis can bring important understandings of how different universities approaches science communication. This method would benefit from the implementation of public API's that could allow easier and broader access to science databases, and certainly a limitation that must be taken into account is the lack of correct registration by scientists.

References

AWRU: Academic Ranking of World Universities. Retrieved 15 December 2019, from http://www.shanghairanking.com/

Barata, G.; Caldas, G. & Gascoine, T. 2018. Brazilian science communication research: national and international contributions. An. Acad. Bras. Ciênc. 90 (2, suppl.1): 2.523-42. https://doi.org/10.1590/0001-3765201720160822

Bucchi, M. & Trench, B. 2016. Science communication and science in society: a conceptual review in ten keywords. Tecnoscienza 7 (2): 151-68.

Cosine similarity. Retrieved 24 August 2021, from https://en.wikipedia.org/wiki/Cosine_similarity

Jaccard index. Retrieved 24 August 2021, from https://en.wikipedia.org/wiki/Jaccard_index

Overview of Text Similarity with Python. Retrieved 21 March 2020, from https://towardsdatascience.com/overview-of-text-similarity-metrics-3397c4601f50

Pedregosa et al., Scikit-learn: Machine Learning in Python. JMLR 12, pp. 2825-2830, 2011.

QS World University Rankings. Retrieved 15 December 2019, from https://www.topuniversities.com/university-rankings

RUF: Ranking Universitário Folha. Retrieved 15 December 2019, from https://ruf.folha.uol.com.br