Quantitative Analysis of Texts: Presentation and Operationalization of the Technique via R Interface on Twitter

Análise Quantitativa de Textos: Apresentação e Operacionalização da Técnica via Linguagem R no Twitter

Steven Dutt-Ross
Breno de Paula Andrade Cruz

ABSTRACT

A Quantitative Text Analysis allows texts to be analyzed in the light of Statistics - a word cloud is an example used to communicate the results of statistics through an image. This text aims to present this technique for data collection and treatment through the R interface by presenting a tutorial. In this way, public data of Unirio's official account is used as a didactic example for presentation as possibilities for data collection and treatment on Twitter. Initially, it discusses Method and Technique in Science in the light of online information and then advances in the R interface entries and outputs, they are: (a) Preparation of the database through Twitter; (b) Frequency Results - word cloud and absolute frequency of terms; (c) Bigrams and trigrams - statistical association of words; (d) Word Correlation Network - word co-occurrence graph; (e) Cluster Analysis - statistical association of words; and (f) Next Word Prediction. This text is relevant to contribute to the training of researchers in Public Administration and Business by presenting a free interface (R) that can be used to generate results or badges for empirical research.

Key-Words: Quantitative Text Analysis; Research Technique; Quantitative Research; R Interface; Twitter.

RESUMO

A Análise Quantitativa de Textos permite que textos sejam analisados à luz da Estatística – a nuvem de palavras é um exemplo utilizado para comunicar por meio de uma imagem os resultados de frequência estatística. O presente texto tem como objetivo apresentar esta técnica para coleta e tratamento de dados por meio da linguagem R ao apresentar um tutorial para replicá-la. Desta maneira, são utilizados como exemplo didático dados públicos da conta oficial da Unirio para apresentar as possibilidades de coleta e tratamento de dados no Twitter.

Submitted: 07/01/2020
Accepted: 09/22/2020

Steven Dutt-Ross
steven.ross@uniriotec.br
PhD in Production Engineering – Universidade Federal Fluminense
Doutor em Engenharia de Produção – Universidade Federal Fluminense
Rio de Janeiro/RJ – Brazil

Breno de Paula Andrade Cruz
brenocruz@gastronomia.ufrj.br
PhD in Business – Fundação Getúlio Vargas
Doutor em Administração – Fundação Getúlio Vargas
Rio de Janeiro/RJ – Brazil

RESUMO

Inicialmente discute-se Método e Técnica na Ciência à luz de informações online para depois avançar nos *inputs* e *outputs* da linguagem R, são eles: (a) Preparação do banco de dados pelo Twitter; (b) Resultados de Frequência - nuvem de palavras e frequência absoluta dos termos; (c) Bigramas e trigramas – associação estatística de palavras; (d) Rede de Correlação de Palavras – gráfico de coocorrência de palavras; (e) Análise de Cluster – associação estatística de palavras; e (f) Previsão da Próxima Palavra. Este texto é relevante por contribuir na formação de pesquisadores do campo de Administração Pública e de Empresas ao apresentar uma linguagem gratuita (R) que pode ser utilizada para gerar resultados ou *insights* para pesquisas empíricas.
Palavras-Chaves: Análise Quantitativa de Textos; Técnica de Pesquisa; Pesquisa Quantitativa; Linguagem R; Twitter.

Science and Data Collection in the online environment

The increasing development of digital platforms that make virtual social networks viable has significantly affected the communication process in recent years - whether in public or private spaces. Freire and Freire (2019a) highlight the importance of understanding the amount of information available in the virtual environment when making a counterpoint between Data Science and Information Science. Specifically, the authors highlight the link between Data Science (mainly digital) and scientific communication. The importance of focusing on social networks is given, among other reasons presented by Cezar and Suaiden (2017), due to the fact that the conception of the identity of individuals encompasses their participation in social networks, mainly because the society of information prints a complex interactive pattern. This interaction often occurs mediated by digital platforms.

There are tools that can assist the market in decision making and the production of knowledge in academia. Historically, software such as Excel, Word, SPSS and Stata have helped researchers to produce knowledge in their fields; and recently, increasingly sophisticated software has emerged - including for qualitative research. But they usually have one feature in common: they are licensed by companies and need to be paid. The R language appears as a free possibility for researchers to program and develop statistical analysis - including via digital platforms with API (Application Programming Interface) - as is the case with Twitter.

R is a computer language among its users and shows a strong emphasis on the treatment of statistical data. It is possible to program a code to extract information from a database. In addition to being free and having free code, the R language allows you to reproduce the results as well as new libraries (library) are developed every day - the user network is large, so it is easy to get help and answer questions about the use of the language. If a researcher in London builds a code, a researcher here in Brazil can reproduce everything that has been done if the database is available; or else replicate the model with new primary data. Most cutting-edge research is done in R or Python, as it can be seen in recent studies published in Nature - Le Lan et al (2020), Virtanen et al (2020), Prat et al (2020) e Benítez-Cabelo et al (2020).

More attentive researchers who are interested in techniques such as the one presented here may wonder why using the R language and not the Iramuteq ® or Alceste ® software. Iramuteq ® uses the R language to perform its analysis; however, it works on the classic version of R (version 3.1.2 of 2014). By not using the most current version of R (version 4.0.1 3 of 2020), Iramuteq ® becomes obsolete and outdated by 6 years with regard to word processing. On the other hand, Alceste has an owner (copyright). More than that, both Alceste ® and Iramuteq ® do not offer the outputs that we present here as resources, such as (i) forecasting the next word, (ii) the co-occurrence network of the terms, (iii) bigrams and (iv) trigrams. In summary, although Iramuteq ® is free as the R language, it is outdated regarding our proposal.

The spontaneous publication of individuals or legal entities on digital platforms such as Twitter enables a large amount of information that can be worked on in R. The data are present in several digital platforms and are the result of the communicational explosion that emerges from Web2 (FREIRE; FREIRE, 2019b). It is interesting to notice that information overload on the web does not occur only in popular digital platforms such as Facebook and Twitter - this also happens in the process of sharing scientific production, as pointed out by Cassotta et al. (2017).

As a strategy to keep pages and websites up to date, many organizations use pages on social digital media such as Twitter, Facebook, Instagram and WhatsApp to communicate with customers, suppliers and other stakeholders. This interaction seems to be more effective than before - even if they have problems that can damage a company's image or reputation (CRUZ, 2017). Several studies point out the use of data collected in the online environment and to do Science in different fields

of knowledge (HOGAN, 2017; GRANELLO; WHEATON, 2011; LEFEVER, DAL; MATTHÍASDÓTTIR, 2006; CANTRELL; LUPINACCI, 2007).

It should be brought to attention that our goal is not to discuss data mining. Data mining creates the corpus (our database of words); and in Quantitative Text Analysis data mining is only the first step, as our focus is to present statistical outputs from words, such as the word cloud, bigrams, trigrams, next word prediction, cluster analysis and network co-occurrence of terms. Readers who seek greater depth in techniques related to data mining can consult the works of Wu et al. (2008), Hand and Adams (2015) and Wu et al. (2014)

In addition, we highlight that our proposal in this text is not to discuss whether we should adopt Quantitative Content Analysis (NEUENDORF; KUMAR, 2016) or other qualitative methods: we present the Quantitative Analysis of Texts from public data on Twitter - not featuring as Netnography. And, specifically with regard to qualitative research in the area of Public and Business Administration, the text by Cruz and Ross (2018) signals an important reflection for some qualitative studies that use data collected in the virtual environment and are classified as Netnography. The authors emphasize the methodological rigor of Netnography as a Method and criticize studies that only collect data on digital platforms and classify them as Netnography. Thus, the technique we will present is neither (i) a method (ii) nor a data collection technique that makes part of a netnographic study.

Netnography is a qualitative method that implies immersion in a virtual community and interaction between the researcher and the investigated virtual community (KOZINETS, 2010). The proposal here is to exemplify the use of a data collection technique (with a purely quantitative bias) that can be used to complement data analysis in quantitative and qualitative studies; or, additionally, help generate insights from an existing database.

Considering the discussion about method and technique, we consulted the Dictionary of Philosophy (JAPIASSÚ; MARCONDES, 1996) to make our central argument more robust: we are presenting a technique and not a method. Thus, we have:

> **Method** - set of rational, rule-based procedures that aim to achieve a specific objective. For example, in Science, the establishment and demonstration of a scientific truth (p. 181).

> **Technique** - set of practical rules or procedures adopted in a letter in order to obtain the intended results. (...) In a sense derived mainly from modern science, practical application of theoretical scientific knowledge to a specific field of human activity (p. 257)

We adopted the technical perspective. However, it is not simply a data collection technique (such as using a focus group). It is a data collection technique since it works with data mining, but it is mainly a data analysis technique due to the presentation of the concepts of bigrams, trigrams and n-grams, Word Correlation Network and Cluster Analysis. Thus, Quantitative Analysis of Texts is a technique that involves the collection of qualitative data on a digital platform of public content to perform a quantitative analysis of these contents by presenting a correlation analysis between the terms used by users.

The presence of the digital platform Twitter is a reality in the communication process among people and organizations themselves, and between people and organizations. Created in 2006, by allowing the sharing of texts, photos, videos and especially the use of hashtags (# symbol), it has become an important communication tool all over the planet, since it is used by politicians (AUSSERHOFER; MAIREDER, 2013), companies (CULOTTA; CUTLER, 2016), NGOs (GUPTA; RIPBERGER; WEHDE, 2016), celebrities and anonymous people who become celebrities.

In this sense, the analysis of the posts allow the assessment of how this type of interaction between different people and organizations in society happens on Twitter. In Brazil, for example, politics and television are always hot topics on Twitter (SANTINI et al., 2020). But there are also other issues that deserve to be highlighted - such as the fight against Coronavirus. For a public or private educational institution, for example, it is interesting to understand what students and society write about the organization because it is possible to (re) think the image and reputation through institutional communication.

Another example considering the perspective of public administration is to understand the assessment of the Ministry of Health and President Jair Bolsonaro by Twitter users amid the coronavirus crisis. Although DataFolha released partial results on the evaluation of the president and the former ministers (DATAFOLHA, 2020), it must be considered that the behavior of users on Twitter is different from those who do not use the platform. Thus, uploading hashtags (highlighting a subject on social media) is a strategy to draw media attention to issues that groups, social movements and users consider important.

Thus, the objective of this study is to structure and systematize phases and the step by step of Quantitative Analysis of Texts conducted in the R language as a data analysis technique. Specifically, we have: (a) the use of public data from the account of the Federal University of the State of Rio de Janeiro (Unirio) on Twitter to present real examples of the operationalization of the technique - not requiring formal authorization from the institution; (b) the discussion on the importance of updating the techniques in conducting studies in the field of Public and Business Administration with the use of a free programming language such as R. The next section presents the operationalization of Quantitative Analysis of Texts.

The Operationalization of Quantitative Text Analysis

In order to exemplify the use of this technique in this work, we focused on the metropolitan region of Rio de Janeiro and four federal institutions of higher education: Federal University of Rio de Janeiro (UFRJ), State University of Rio de Janeiro (UERJ), Federal University of the State of Rio de Janeiro (Unirio) and Universidade Federal Fluminense (UFF). In a total of more than 34 thousand publications until December 2019, we worked with the intentional sample (purposive sample) and chose Unirio (account @comunicaUNIRIO), with 4,768 publications, because we are familiar with the institution – and it facilitates the interpretation of results in application of the technique.

PHASE 1 – DATA COLLECT VIA R-TWEET

There are two ways to capture data from the Internet: via data scraping and via API. In data scraping a routine is created to capture data from the Internet. However, this scraping of data is often illegal because it is a cloning of data that can be commercialized (and there is a market for that). The API is a set of routines and programming standards that allows accessing an application / platform. R-Tweet is an external system that consults data on the Twitter platform through integration via API. As it is allowed by Twitter, there is no illegality or ethical conflict for the researcher (the data can be consulted and used).

R-Tweet is an R library to access the Twitter API (KEARNEY, 2019) and that makes it possible to download up to 3,200 recent tweets from a given account. For

example, the capture of more than three thousand posts from Unirio was carried out through this package. For Kearney (2019), twitteR is different from R-Tweet due to the possibility of searching for users, keywords and the capture of status (such as Facebook considering the status update).

In this initial data collection we focused only on the main publications. In other words, all retweets and replies given by account followers or Twitter users to a post have been deleted. We also used the twitteR package (GENTRY, 2015) - a package to capture the corpus on Twitter and work in the R language (hence the capital letter R). These texts were saved in text format (.txt) and can be downloaded here[1]. Then, Unirio's Twitter information was used. Data collection took place on January 3, 2020, and the corpus capture for the @comunicaUNIRIO account totaled 3,140 posts as of October 14, 2015.

PHASE 2 – TEXT CLEANING

We can observe that Unirio's Twitter used 8,984 different words in the last 3,140 posts between 2015 and 2019. The data was captured in the beginning of 2020 and we collected all posts in the period between 10/2015 and 12/2019. However, as can be identified in Figure 1, comments appear with emojis, prepositions and connectors that must be eliminated in order to actually be able to make an association between the words. If this cleaning does not occur, these words will appear in the following stages in the first positions of the rankings and may neglect other words or expressions that are really important (RAULJI; SAINI, 2016; SCHOFIELD; MAGNUSSON; MIMNO, 2017).

After downloading the data, a routine to clean these data was performed. This is because we have many words with little meaning, such as the connectors "de", "da", "que". As a result, it was necessary to adopt the procedure to delete them. Thus, the nouns, verbs and adjectives remained; pronouns, articles, numerals, prepositions, conjunctions, interjections and adverbs were eliminated. In addition, we also have words or expressions from the web that have no meaning and are the result of Twitter hyperlinks. This category was named as computational language and words like "https", "http", "www" and "#" were also excluded.

1 https://bit.ly/3jCCndJ

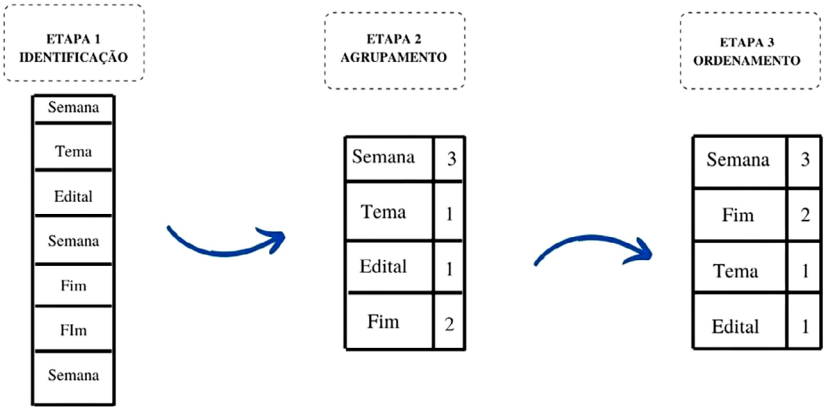
Figure 1 Data Bank example – 2015/2019

	Date	Tweet
1	2019-12-16 13:53:00	@sleepinthdirt Vem comigo!
2	2019-12-16 13:43:00	Hoje tem filézin de frango com purê de batata. Amanhã tem carne ensopada ou almôndeg...
3	2019-12-13 19:49:00	@PAlinhares @uff_br UNIRIO, PÔ!!!!!
4	2019-12-13 10:09:00	#pracegover Imagem com fundo azul comemorativa do Dia Nacional do Cego. Mulher ceg...
5	2019-12-11 13:53:00	Minicurso 'Treinamento para participação em atividades com animais vertebrados': inscriçõ...
6	2019-12-10 15:21:00	Os espetáculos de fim de ano do Projeto de extensão Teatro em Comunidades Redes de Te...
7	2019-12-10 15:19:00	Gente, vai até sexta (13) a exposição: "Egito: A antiguidade com olhares de modernidade". ...
8	2019-12-06 15:42:00	O documentário "Corpo de Baile" que conta os 80 anos de história da dança no Theatr...
9	2019-12-04 13:29:00	Trago verdades!!!!! Este vídeo da série "Saúde é Vida" do NIS/UNIRIO tá demais. Precis...
10	2019-12-03 13:37:00	@LuceldoMino @taisdeverdade Nossa, parece delicioso! (SQN)
11	2019-12-03 12:51:00	Você define os seus limites! https://t.co/9bEbvHkZeT
12	2019-12-03 12:20:00	Hoje às 13h30 vai rolar mesa-redonda: Cuidados intermediários, rede de atenção e...
13	2019-12-02 14:04:00	02/12 - Dia Nacional do Samba. https://t.co/oRnvl46nSr
14	2019-12-02 13:06:00	Hoje tem escondidinho de frangoooo. Terça tem carne moída ou hamburguer de soja. N...

Source: Research data.

In order to make this data cleaning operational, we converted all the texts into the Tidy format (WICKHAM, 2014) - a database with each word on a line. After this step, we checked the frequency of each word. To demonstrate this format, we divided this method into 3 steps: Step 1 - Identification of words; Step 2 - Grouping the same words; Step 3 – Counting equal words and ranking them. Figure 2 shows a database in Tidy format.

Figure 2 Steps to insert the data in Tidy format



Source: Authors' elaboration

Table 1 shows the 10 most frequent words before and after this cleaning procedure. In other words, considering these 10 words, only "dia" could be a word that brings some meaning - it could be the "prato do dia no bandejão" or "dia de" an event at the institution. Through these results, we highlight the importance of eliminating meaningless words. As it can be viewed, other words are identified after this cleaning and the list of banned words can be viewed here on this link[2].

Table 1 10 most frequent words in Unirio's Twitter between 2015 e 2019 before and after the cleaning step

Before cleaning		After cleaning	
Word	**Frequency**	**Word**	**Frequency**
t.co	2515	Unirio	819
HTTPS	2466	confira	301
De	2295	saiba	264
E	1593	inscrições	249
A	1130	vai	221
Da	1051	escola	166
O	1006	palestra	158

2 https://bit.ly/3kGo7BZ

Dia	985	semana	150
Do	880	tema	147
Unirio	819	edital	138

Source: Data collection of the account @comunicaUNIRIO

We realized that after eliminating some words and expressions the word Unirio jumped from the 10th position to the 1st position in the ranking. Initially analyzing the terms that appear in Table 1 after eliminating some words, we can see that they are associated with the context of a university as there are words such as school, lecture, public notice, know and theme. Thus, the use of the words “confira”, “saiba”, “inscrições”, “palestra” and “edital” also suggests that Unirio’s official Twitter is used for sharing the institution’s calendar. Possibly, the widespread use (150 times) of the word “week” has to do with Academic Integration Week - SIA (a major event at the university). It is interesting to notice that the words associated with the Scientific Initiation Day - JIC (another major event of the institution) are not among the top ten. The word cloud is shown in Figure 3.

Figura 3 Word cloud after data cleaning

Source: Data collection of the account @comunicaUNIRIO

STATISTICAL ANALYSIS

R is a language used by statisticians from all over the world to present simple or sophisticated models through their libraries - and in this methodological article two relevant libraries are being used: Quanteda and Tidetext. Therefore, it is used by researchers who perform quantitative research with theoretical and methodological robustness, using other libraries such as OpenNLP, Rweka, RcmdrPlugin.temis, tm, languageR, koRpus, RKEA, lsa and maxent. Thus, the analysis of texts presented here is quantitative (and not qualitative). The results presented are the results of statistical methods programmed in R.

Bigrams (Word pairs /partnerships)

The n-grams are the adjacent elements in a sequence of words mined in a text and are recognized and generated statistically. Thus, they can be classified by their number of combinations: a bigram for two combinations, a trigram for three combinations, a tetragram for four combinations and so on until the polygams are reached.

A bigram is a sequence of two adjacent elements in a sequence of symbols (tokens). A Bigram is an n-gram for n = 2. With a bigrama we try to answer the following question: 'What words are used together more often?'. This strategy can be used for any database or official Twitter account or other platform that contains public data. In order to build it, it is necessary to use the unnest_tokens function of the tidytext package (Silge, Robinson, 2016). This function divides a table into one token per row. Annex 1 (Tutorial for the Replication of the Quantitative Text Analysis Technique) presents all the codes used in this article. Then, it is possible to go through all the steps to replicate the technique's step by step.

Notice that Unirio was the first word in the word cloud (Fig. 3), but it does not appear in bigrams and trigrams because Fig 3 is the visual representation of the simple frequency of each word. Bigrams and trigrams, on the other hand, seek word associations and can not be considered in the simple frequency of words.

Regarding Unirio's institutional communication, we seek to verify the association of words and the frequency of this association. We built all possible bigrams from the 3,140 Unirio posts. After creating the bigramas, we checked the frequency of each one. After this step, we ordered by frequency: the ten most frequent bigrams are shown below in Table 2. Notice that the 'pós-graduação' bigram can be analy-

zed only as an expression of a term separated by the hyphen. Thus, when cleaning the text, if we remove hyphen from 'pós-graduação' and 'mesa-redonda' (which suggests the debate and not the physical format of the table), a different result can be presented and this suggests a limitation of the technique.

Table 2 The top ten bigrams in Unirio 2019's tweets

Primeira palavra	Segunda palavra	Frequência
restaurante	escola	52
pós	graduação	46
villa	lobos	45
mestrado	profissional	42
fique	ligado	38
inscrições	abertas	37
vai	ser	26
mesa	redonda	24
iniciação	científica	23
quintas	culturais	22
Unirio	musical	22
auditório	vera	21
aula	inaugural	21

Source: Data collection of the account @comunicaUNIRIO

It is interesting to observe that advertising the school restaurant is the most common activity at Unirio. After the presentation of the Menu, we can the concern with the research agenda with a great frequency of the words "pós-graduação", "mestrado profissional", "mesa redonda" e "iniciação científica".

There, it is possible to observe the institution's calendar (to advertise its events) by the use of word partnerships "fique ligado", "inscrições abertas", "Auditório Vera", "aula inaugural", "série Unirio". Finally, we can also observe a cultural calendar by the use of word pairs such as "quintas culturais","Unirio musical" and "artes cênicas".

The same procedure presented above is used for the construction of Trigrams and other polygrams and can also be replicated through Appendix 1. Thus, the programming is the same and one word joins others through the identification and selection of n- grams via statistics in R. Table 3 shows the results for the year 2019 the trigrams found in Unirio's account.

Table 3 The top ten trigrams in Unirio 2019's tweets

Primeira palavra	Segunda palavra	Terceira palavra	Frequência
sala	villa	lobos	21
série	Unirio	musical	18
auditório	vera	janacopulos	14
projeto	quintas	culturais	13
auditório	tércio	pacitti	10
infecção	hiv	aids	9
série	villa	lobos	9
villa	lobos	aplaude	9
instituto	villa	lobos	8
ter	vcs	aqui	8
auditório	vera	janacópulos	7
série	vitrine	musical	7
siga	curta	confira	7
vai	rolar	palestra	7
continue	acompanhando	aqui	6
enfermagem	alfredo	pinto	6
alfredo	pinto	eeap	5
hospital	universitário	gaffrée	5

Source: Data collection of the account @comunicaUNIRIO

Knowing the context or the theory is important to analyze the n-grams that arise from the results indicated by the Quantitative Text Analysis technique. For example, a reader who does not know the reality of Unirio may possibly not under-

stand the first trigram “Sala Villa Lobos” and others like “Série Unirio Musical”. If one of the authors is part of the institution, we can assume that both the room and the series are linked to the Bachelor of Theater course. Thus, we could use this interpretation to suppose that there is a group of trigrams that is related to ‘Arte’. It is also possible to suppose that in a university with 25 undergraduate courses, perhaps the courses in Theater, Nursing and Medicine are those with greater prominence in the official communication of the institution.

When we refer to Medicine and Nursing as courses highlighted in the official communication of Unirio on Twitter, we consider the trigrams “infecção” “HIV” “AIDS“, “Enfermagem Alfredo Pinto”, “Alfredo Pinto - EEAP” (name of the school of nursing) and “Hospital Universitário Gaffrée “- and all of them related to the Health area.

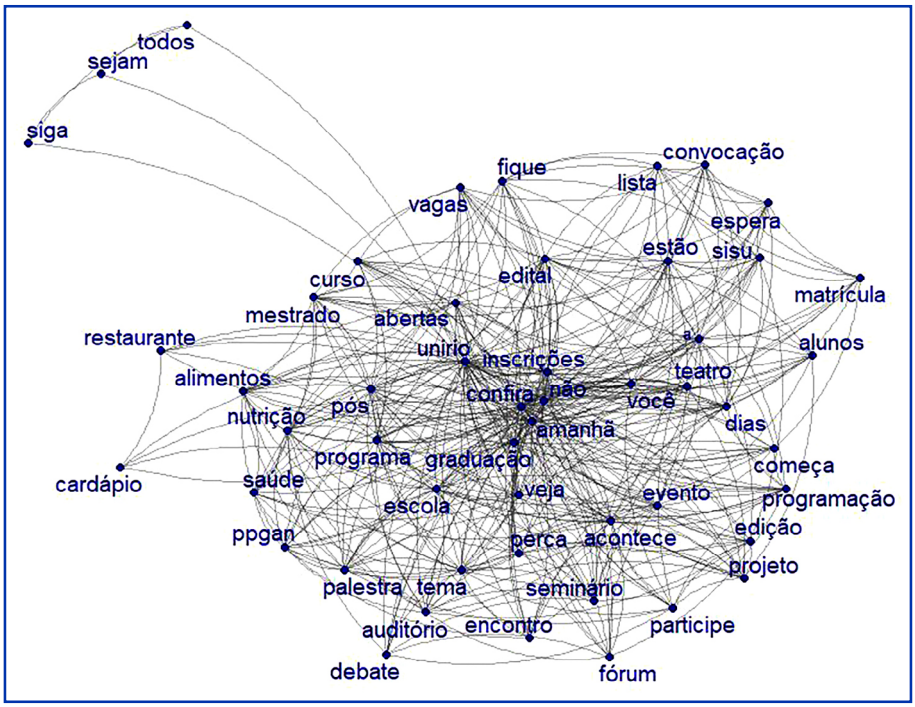
A third group that emerges from a qualitative analysis of these trigrams is ‘Agenda de Divulgação’. The word auditorium that mis contained in the trigram ‘Auditório Vera Janácoulos’ highlights this analysis, as well as the trigrams ‘Vai rolar palestra’ and ‘sala Villa Lobos’. Considering that Twitter is more used by young people (RUMMO et al., 2020) than other previous generations, it is possible to understand this official Unirio account with highlights to the Art and Health events for undergraduate students.

Word Correlation network

Below is a tool to automatically generate a visual summary of unstructured text data. Unlike tools like word clouds, by using this tool we seek to observe the structures of relationships between words - while in the word cloud the largest word represents the highest frequency, in the Word Correlation Network these relationships are determined by an analysis of co-occurrences .

Thus, the algorithm applies a decreasing function to the distance between the pairs of words found, so that words that occur regularly close to each other have a high score (the shorter the distance, the greater the value); but even words that occur at some distance will make a small contribution to the general co-occurrence. To trace a co-occurrence network of words, we use the textplot_network function of the Quanteda Package (BENOIT et al., 2018). Step 3 of the tutorial shows the schedule and Graph 1 shows the results of the Unirio case in this text.

Graph 1 Correlation of Words (Co-occurrencies)



For Graph 1, we see that the words “siga”, “sejam” e “todos” are distant from the others. This is justified because when analyzing Unirio’s posts, we verified that there is a greeting message: “sejam todos bem vindos”; and the messages ended with the phrase “follow @Unirio” - in reference to the institution’s Twitter account. Therefore, that is why these three words are associated and they have co-occurrence (they always appear with some proximity - when one appears, possibly the others also appear). In this network, it is possible to see that these three words are connected with the word Unirio - which is associated with all the other words; and which shows that the word Unirio is the main word in this network of words (which corroborates the findings of the word cloud).

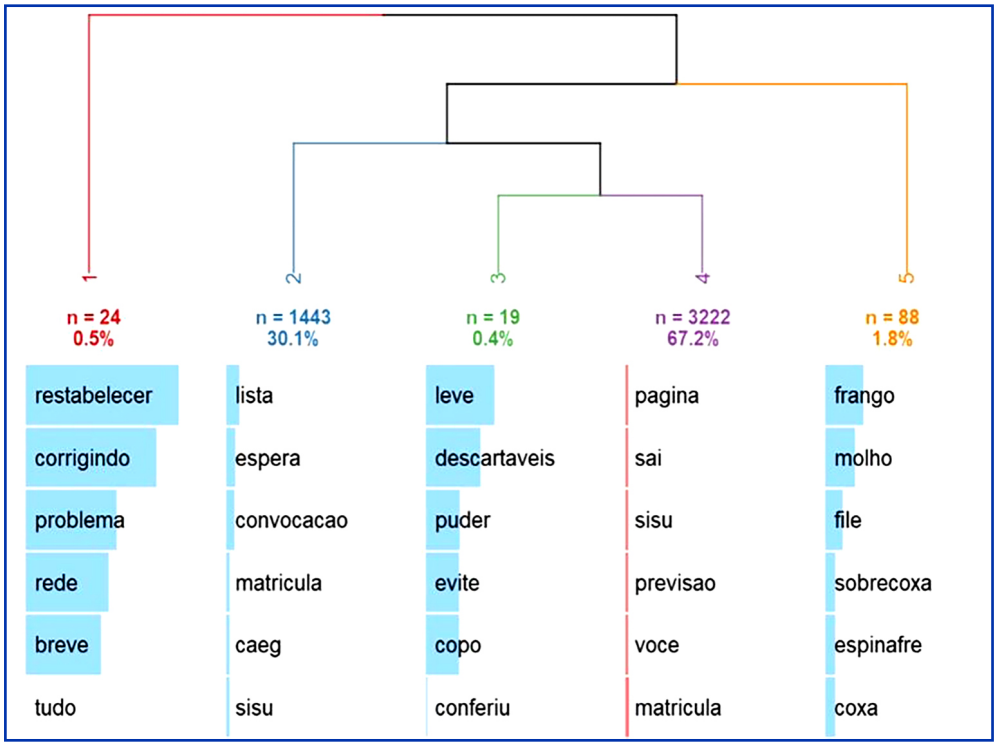
This was the first possible analysis in Graph 1 and it can be extended to other words that appear in a graph like this, which is the result of the technique presented here. For example, the word “restaurante” is related to the words “menu”, “alimento” and “nutrição”. Likewise, the word “aluno” is connected with “matrícula” - which

is associated with the words Sisu, “lista’, “espera” and “convocação”. With regard to empirical studies, the chosen theory or knowledge of the field may help to explain the co-occurrence of words.

Cluster Analysis

Cluster analysis is a statistical technique used to classify elements into groups, so that elements within the same cluster are similar, and elements in different clusters are distinct from each other (MYERS; SIROIS, 2006). To define the similarity - or difference - between the elements, a distance function is used. In this approach, the Euclidean distance was used. In addition, it should be observed that we have adopted the hierarchical method. The result is shown in graph 04 below and its operationalization in Annex 1.

Graph 2 Dendogram with the Cluster Analysis results for Unirio’s Twitter account



From Graph 2 we can perform the interpretation of the groups that were identified in the Cluster Analysis. Each group presents its frequency (n) and its percentage within the database. It is possible to notice that the first five words of each group contribute significantly to the process of naming each group. Thus, we have:

- **Group 1 – Problems:** There are words like 'restabelecer', 'corrigindo' and 'problema'. This indicates that Unirio's tweet communicates problems that happen within the University space. For example, problems with the internet, water shortage, energy outage.
- **Group 2 – Call for new students:** This category is related to the call for new students to join Unirio, highlighting aspects related to enrollment, waiting list and general information about the Sisu list.
- **Group 3 – Being sustainable at Bandejão (the popular name for the restaurant at the university):** In this category, we highlight the great amount of information regarding a more sustainable posture of the users of the university restaurant, since they are asked to bring their own glasses and avoid the increase of waste by the use of disposable glasses.
- **Group 4 - ENEM:** This category presents the words that are related to ENEM (a type of university entrance exam that is taken by high school students), from the registration, waiting list, call and announcement. The peculiarity of the words here does not help to understand the main characteristic that differentiates group 2 from this group. An analysis of feelings via R could be applied to verify if the difference between the groups occurs through the feelings associated with the words of the two groups. This possible similarity of these two groups considering the example used here shows how the theory can help to interpret the results of the Quantitative Text Analysis technique presented here.
- **Group 5 – Prato do dia (Today's special):** This is the most well-defined category, with the description of various food items that are part of the dish that is offered as "today's special" at Unirio's university restaurant.

Based on this Cluster Analysis, we no longer have a subjective analysis of what we can possibly find in Unirio's account and we establish a rational-statistical

analysis without interference of a researcher's personal values, for example. The results of this case presented by this technique can even have managerial implications that can generate some reflection by the managers of that institution regarding the institutional communication. The next section introduces the 'prediction of next word' feature.

Prediction of next word

Google's autocomplete feature is an example of predicting the next word and will be presented here as one of the possible results for Quantitative Text Analysis. Thus, when we perform a Google search, it tries to guess the next word. We can also create a model to predict this from a database. For example, when Unirio's communication uses the word "Unirio", which word could come next? Thus, from Table 4, we show that it is possible to predict the next word that Unirio's official communication would use.

Table 4 Example of prediction of next word considering the word 'Unirio'

Word 1	Word 2	Absolute Frequency	Relative Frequency (%)
Unirio	musical	12	33,33
Unirio	promove	9	25,00
Unirio	recebe	7	19,44
Unirio	inscrições	6	16,67
Unirio	oferece	2	5,56

Source: Elaborated by the authors considering the database.

After the creation of the bigram, a sequence of two adjacent elements, we can use them to predict the next word. For this, it is necessary to fix the first word of the bigram - we did it in Table 4. The modal word (which is more repeated) after the term "Unirio" is "musical". Based on relative frequency, this word is more likely to appear when the term Unirio is used by the institution's official account.

For this method to work, stemming all words (equivalence treatment) is necessary. Lemmatization is the process of grouping the inflected forms of a word so

that they can be analyzed as a single item, identified by the lemma (WACHELKE; WOLTER, 2011) - for example, the word beautiful: beautiful and beauty would be grouped in the same category. In many languages, words appear in various inflected forms. For example, the verb 'andar'can appear as 'andar', 'marchar', 'caminhar' and 'percorrer'. The basic form (the root word), 'andar', which can be found in a dictionary, is called the word lemma. Thus, all of these words have been replaced by "andar" (basic form / root word).

As this research is being conducted in a public university, we replicate the bigrams for the words 'Education' and 'Research' to observe which words could be associated with those terms. The results of Unirio's official account can be seen in Table 5.

Tabela 5 Prediction of next word using the words 'Educação' and 'Pesquisa'

Word 1	Word 2	Frequency	Word 1	Word 2	Frequency
	ambiental	**3**		**economia**	**3**
	infantil	3		produção	3
	tutoria	3		científica	2
	ser	2		acadêmica	1
	popular	2		aids	1
Educação	tutorial	1	**Pesquisa**	bioescritas	1
	contra	1		cace	1
	cultura	1		cultural	1
	estatística	1		veja	1
	excelência	1		alemão	1
	fala	3		debateram	3

Source: Elaborated by the authors considering the database.

In this section the main outputs generated from the Quantitative Analysis of Texts as a research technique were highlighted. If Iramuteq ® brings the word cloud and cluster analysis as outputs, this technique used in this article allowed a further analysis by presenting the bigrams, trigrams, prediction of the next word and

the co-occurrence network of terms of this quantitative research technique programmed in R. All of these open source outputs can be reproduced using the tutorial that is additionally provided.

Conclusions

The availability of data on different platforms and digital media has been a reality in the new context of research in Applied Social Sciences considering the last two decades. Before data collection often involved systematic planning of contact with sources and organizations; nowadays much data can be obtained through public or private portals. In the Transparency Portal (Comptroller General of the Union) it is possible to obtain several data that can be used in different research in Public Administration; on the 'Reclame Aqui' website, it is possible to understand the image of a company by reading consumer complaints; on Twitter it is possible to analyze even the speech of politicians.

Thus, at a time when Brazil suffers from an attempt to wreck public universities and science (for some this wrecking has existed for years), finding new strategies and employing new techniques that allow analysis of a larger group of observations is an efficient strategy to be adopted by researchers. Therefore, we consider it relevant to have presented here the Quantitative Analysis of Texts through the R language using Twitter as a corpus.

It is important to highlight that the objective of Quantitative Text Analysis is not to replace any other qualitative method such as Discourse Analysis, Ethnography, Content Analysis (which also has its quantitative approach); or replace any other data collection technique (Participant Observation, Focus Group or Non-Participant Observation). Our objective was to show that this technique may even be associated with qualitative techniques and methods in order to turn the results of some research more robust. Considering that it is a recent technique that has been used by statisticians, for a predominantly Positivist science like Administration - according to some authors like Carton and Moricou (2018) - Quantitative Text Analysis seems to be interesting to study the same phenomenon with another magnifying glass: that of quantitative research.

Although Quantitative Text Analysis has been presented here through the Twitter platform, it can be used for data that is in the offline environment. For example, the use of this technique in the Legal area has grown - being called Jurimetrics. Trecenti (2015) and Rangel (2014) point out that it is possible to analyze judges' decisions to predict what next decisions might be like, through the quantitative analysis of texts. Some other studies analyze speeches by presidents and other politicians (JOATHAN; ALVES, 2020; EVANS; CORDOVA; SIPOLE, 2014).

Words have power, even more than numbers (but this has often been overlooked in studies with a Positivist bias). By the words it is possible to understand ideologies, values and beliefs of the sender of a message. It is possible, for example, to understand whether a citizen is Leftist or Rightist through his speech and his publications on virtual social networks. For example, by the expression "lugar de fala" it is possible to assume that whoever uses it is more connected to the Left Wing than the Right Wing ideology - and this assumption becomes more robust when we go to the field of Political Science and identify the relationship between this expression and the Left wing, (as we see in the work of Morais, 2018). Therefore, Quantitative Text Analysis can be used in different fields of knowledge.

We are not saying that subjective analyzes that emerge from qualitative techniques are not relevant; on the contrary, we understand its importance even in positivist studies. However, we advocate here the importance of a complementary analysis through the use of Quantitative Analysis of Texts - either to corroborate the findings of a research, or to generate new insights. The text is much more complex than the number (you have to stem and understand the meaning of the phrase - it is a craft technique). However, statistics must advance more and more towards the use of techniques like the one presented here in this methodological article. In the perception of some statisticians, Statistics is also reinventing itself through techniques like this - which demystify that it is a science of numbers only.

Sales and Saião's (2019) discussion of Little Science versus Big Science brings us interesting insights regarding the existing dichotomies in data generation. If in Big Science there is uniformity in data generation, investment and infrastructure, in Little Science (the one that many of us do), the data are heterogeneous and rarely archived for reuse. Thus, the Little Science that suffers from investments and infrastructure (see the recent cut of scholarships for scientific research in the Humanities

in 2020) has to adapt with the existing possibilities. In addition, in this context, the Quantitative Analysis of Texts using public data (online or offline) is a strategy that can be adopted to work with a larger volume of data (however, this does not mean that we will do Great Science).

Initially, we think it is relevant to shed light on the possibility of doing Science by using the data available on different digital platforms. Although our choice for an educational institution as a case is punctual, we understand that Quantitative Analysis of Texts can allow a multi-referential look to its users from the theories used. Therefore, we highlight here some future research that can be conducted using this technique.

From the consumer's perspective, it is possible to analyze the comments of Instagram users about products advertised by digital influencers to verify the relationship of these sponsored actions with a possible aspirational consumption or characteristics of reference groups. In Public Administration, for example, it is possible to seek to understand a multitude of perspectives such as reports of local public management, review of minutes of local councils to understand specificities and even the coherence between the speech of candidates to the positions of the executive and the implementation of what was planned to answer later: has the main agenda been implemented?

Also considering that it seems more and more that activists (political, environmental, social and moral) use Twitter to shed light on a topic, person or organization, what are the co-occurrences of words in hashtags raised in Trending Topics on a topic? In times of fake news and ethical dysfunctions in society, is it possible to identify robots through the content that some accounts publish? These and other questions can be answered directly or tangentially through the application of Quantitative Analysis of Texts. The tutorial built to exemplify the possibility of this technique and the answer to other research questions is presented at the end.

Annex 1

TUTORIAL OF THE PRESENTED TECHNIQUE

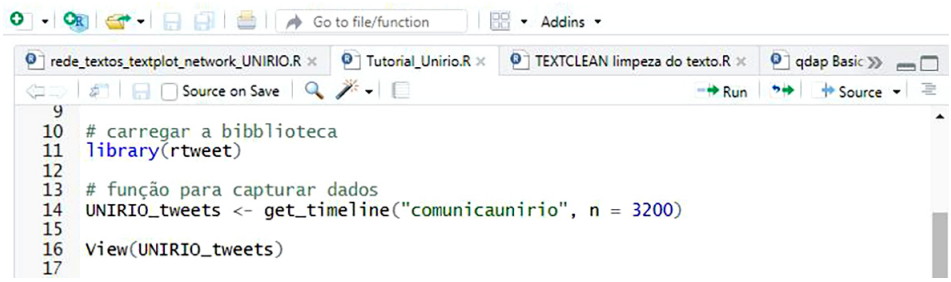
Step 1: Before open R , get the access token for Twitter – The first step is to get the token to access Twitter on the website developer.twitter.com. Get a developer account, browse through developer.twitter.com/en/apps, and click on Create a New App to fill in the form.

Step 2: Browsing through Twitter and searching for the account you would like to collect data from – In this tutorial, we are using Twitter's UNIRIO account. The address is https://twitter.com/comunicaunirio.

Step 3: Opening RStudio and install RTwitter library – To make RTwitter library work, after loading the libraryyou need the key you got on step 2. Look at an example below:

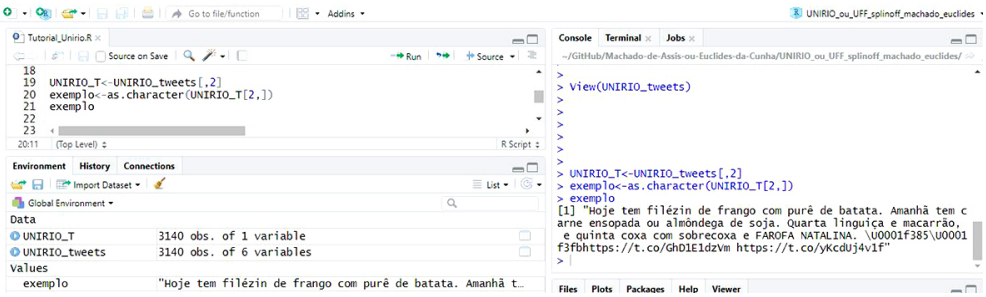
```
## load rtweet
library(rtweet)

## store api keys (these are fake example values; replace with your own keys)
api_key ← "afYS4vbIlPAj096E60c4W1fiK"
api_secret_key ← "bI91kqnqFoNCrZFbsjAWHD4gJ91LQAhdCJXCj3yscfuULtNkuu"

## authenticate via web browser
token ← create_token(
  app = "rstatsjournalismresearch",
  consumer_key = api_key,
  consumer_secret = api_secret_key)
```

In order to collect the tweets from UNIRIO's account , we need to use the code below in R:

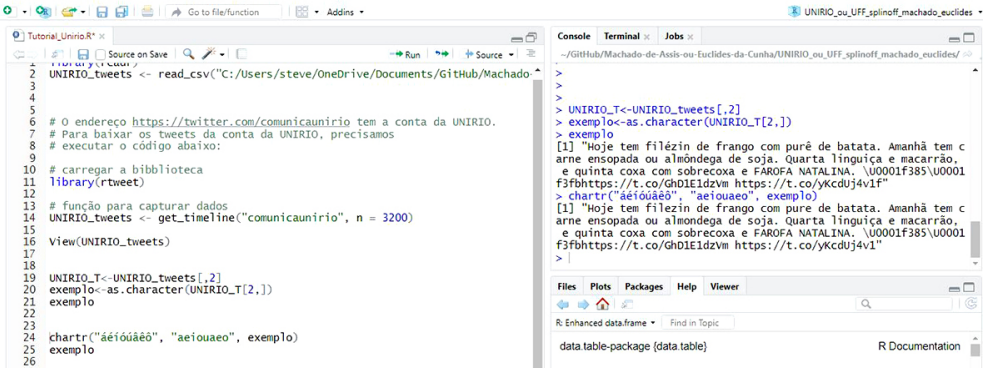


Step 4: Visualizing the Database – The command *View(UNIRIO_tweets)* is used to show database (in order to confirm if the data capture has worked). This command refers to Fig 1 in the text.

Step 5: Filtering only what we need – Considering that in this moment we only need the tweets, we created a new object with them, as it can be seen in the example below.



Step 6: Data cleaning – in order to make the next procedures easier, we remove all the accents (except '~') using the command *chartr*:



To see how this function works, check the example of a text.

> example

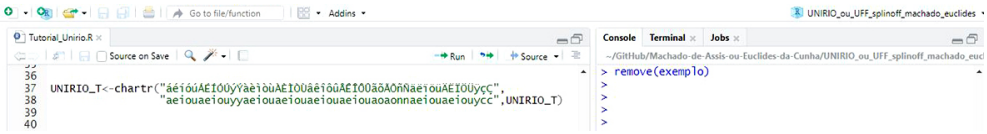
[1] "Hoje tem filézin de frango com purê de batata. Amanhã tem carne ensopada ou almôndega de soja. Quarta linguiça e macarrão, e quinta coxa com

sobrecoxa e FAROFA NATALINA. \U0001f385\U0001f3fbhttps://t.co/GhD-1E1dzVm https://t.co/yKcdUj4v1f"

> chartr("**áéíóúâêô**", "aeiouaeo", example)

[1] "Hoje tem filezin de frango com pure de batata. Amanhã tem carne ensopada ou almondega de soja. Quarta linguiça e macarrão, e quinta coxa com sobrecoxa e FAROFA NATALINA. \U0001f385\U0001f3fbhttps://t.co/GhD-1E1dzVm https://t.co/yKcdUj4v1"

This function removed the accents, and transformed "filézin" into "filezin". Now that we know what this command does, let us use it in the whole database.

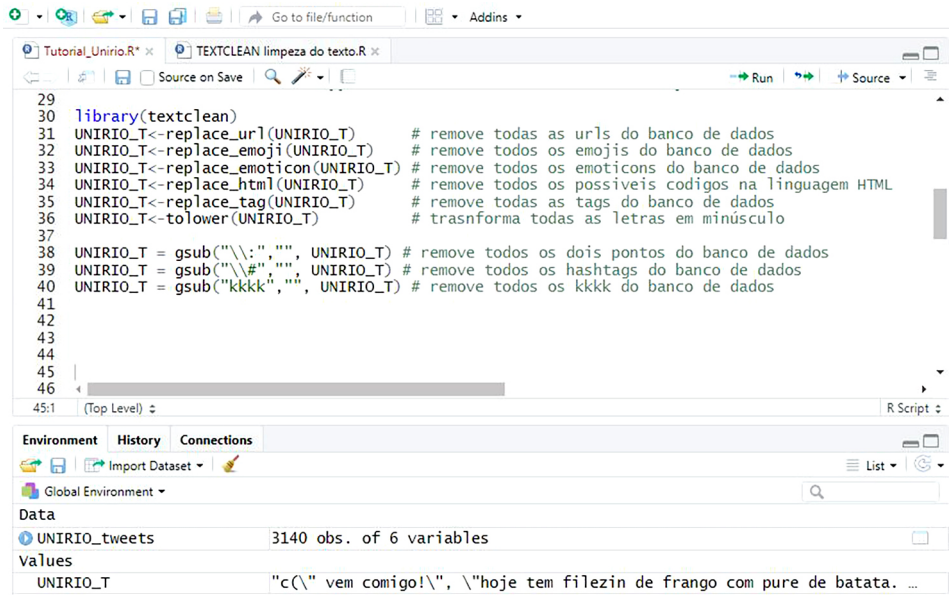


Now it is time to work in the cleaning of data. Besides the accents, we need to establish a pattern for the texts.

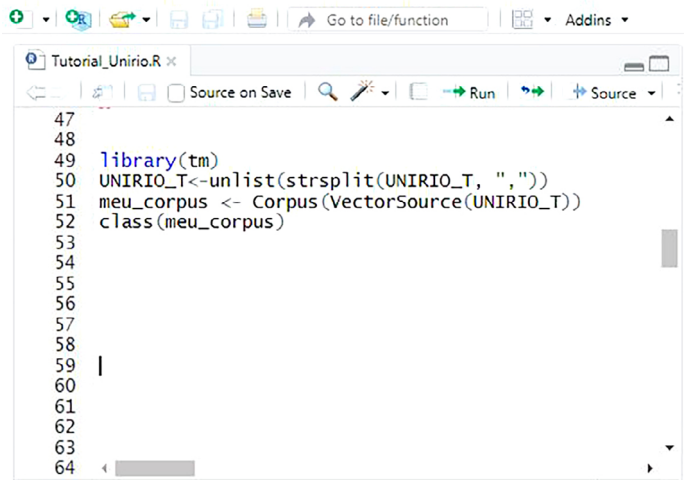
Step 7: Data patterns – We use the *textclean* library. The library is loaded with the command "*library(textclean)".* After this, we remove some elements that are not considered text from the tweet.

- To remove all the urls from database and other information, we use the commands below, for each of them – separated here by a semicolon that is not considered in the coding.

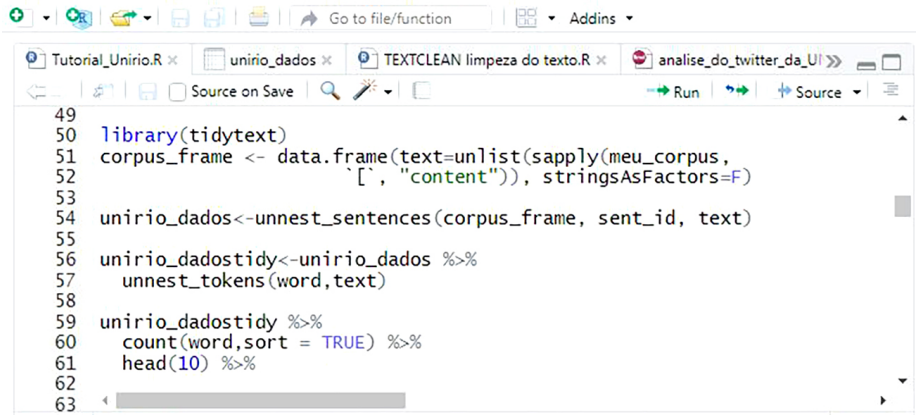
UNIRIO_T<-replace_url(UNIRIO_T); UNIRIO_T<-replace_emoji(UNIRIO_T); UNIRIO_T<-replace_emoticon(UNIRIO_T); UNIRIO_T<-replace_html(UNIRIO_T); UNIRIO_T<-replace_tag(UNIRIO_T); UNIRIO_T<-tolower(UNIRIO_T); UNIRIO_T = gsub("\\:","", UNIRIO_T); UNIRIO_T = gsub("\\#","", UNIRIO_T).



Step 8: Transforming the database into a Corpus – After this, we need to transform this object into a Corpus. In Corpus Linguistics, these text database are objects of research named Corpus. Corpora is the plural of corpus – a set of linguistic data that belong to the oral or written use of language and that can be processed by computer (IBPAD, 2020).

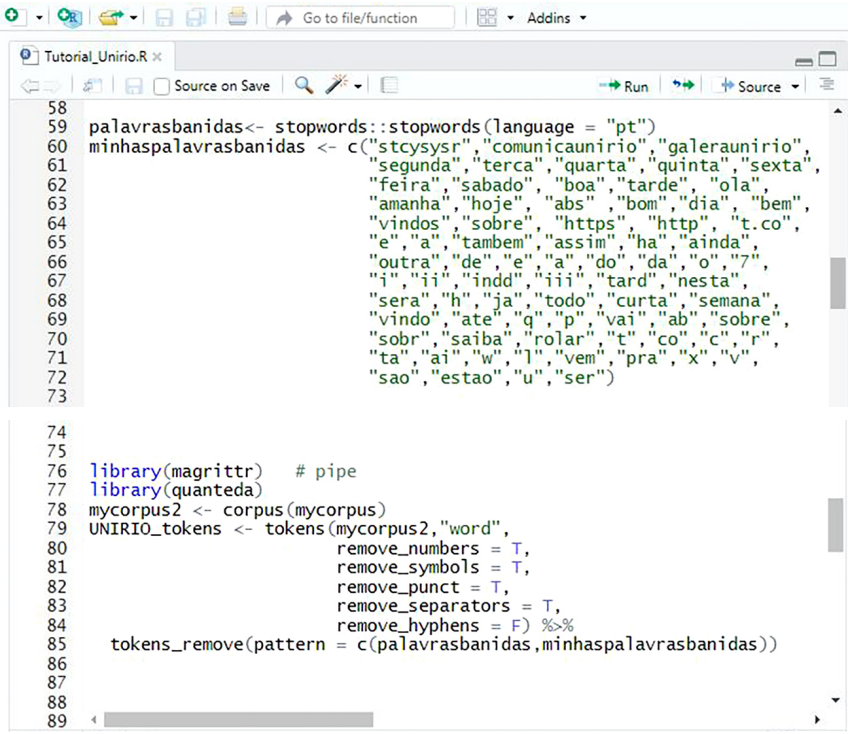


We use the *tidytext* library to divide the database into sentences and then words. A tweet is a text with the limit of 280 characters. We need to break in into smaller pieces. This way, we can create a database of sentences and a database of words. The functions we use are *unnest_sentences* e *unnest_tokens*. The database of words will be very useful to help creating the table of the most frequent words. With the function *count()* we can count the objects. The command sort=TRUE helps to organize it in sequence. Finally, the command head() show us the tem most frequente words – which was previously presented by Table 1.

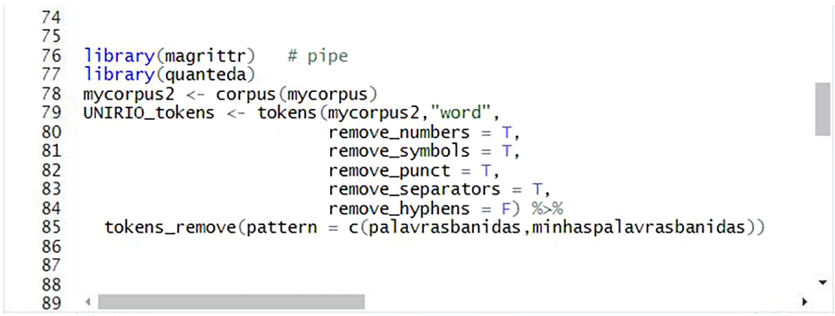


Step 9: Removing the stopwords – Our next step is to remove the stopwords. One of the main forms of corpus pre-processing is to filter out useless data. In natural language processing, useless words (data) are called stopwords. Stopwords is a commonly used word (such as “o”, “a”, “um”, “uma”) that a search engine has been programmed to ignore, both when indexing search entries and when retrieving them, as a result of a search query. In this tutorial, we have two groups of stopwords. The first are those words that have already been defined by the scientific community (pre-defined). The second groups were the words that we define as those that add little value (such as: “bom dia”,”boa tarde”, “quarta”,”quinta”,”sexta”).

To remove *stopwords* we use the function tokens_remove() from the Quanteda package. Find below na example of how to remove them. In this example, we remove the words that were pre-defined as stopwords and those we defined ourselves.



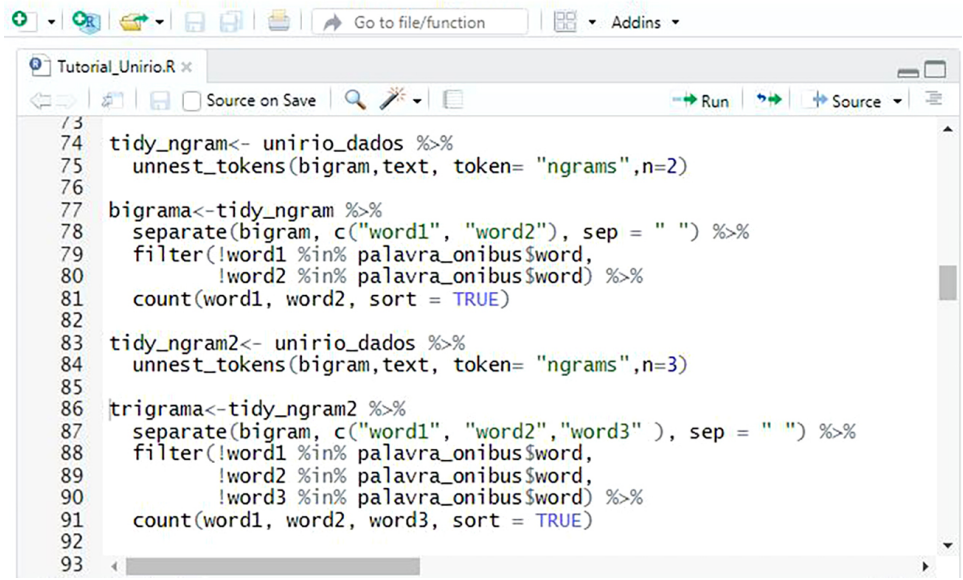
Step 10: Creating the word cloud – As we already have a database with words, we can create a word cloud (Fig 3). In order to create the cloud, we need to build the following code:



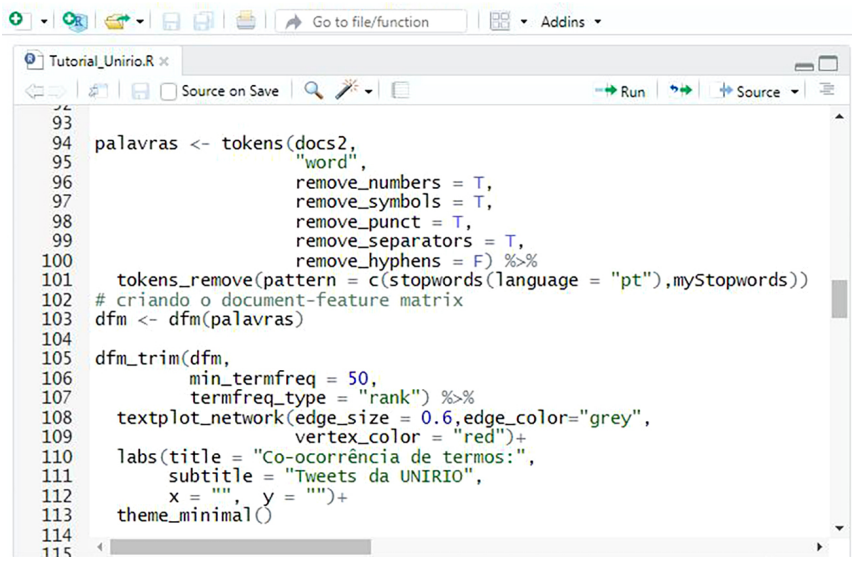
Notice that: **Wordcloud** is the function to generate the word cloud; **Words** is the function to define the object with words; **Freq** is the parameter to define the size of

the words (it will be the frequency here); **Max.words = 100** defines the 100 more frequente words.

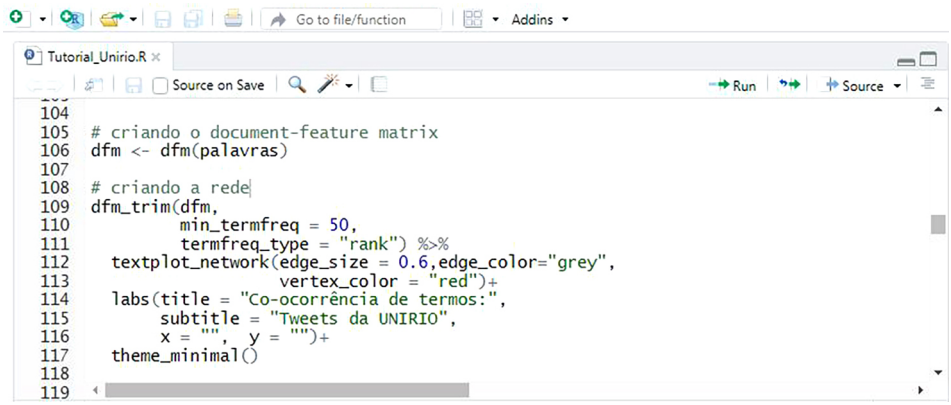
Step 11: Creating bigrams e trigrams – In order to generate bigrams and trigrams (Tables 2 e 3), we need to use the function unnest_tokens(). This function "breaks" the sentences in consecutive sequences of words, called n-gramas. If you define n=2, you choose the bigram, if you define n=3, it will be a trigram. The code to generate n-gramas is as follows:



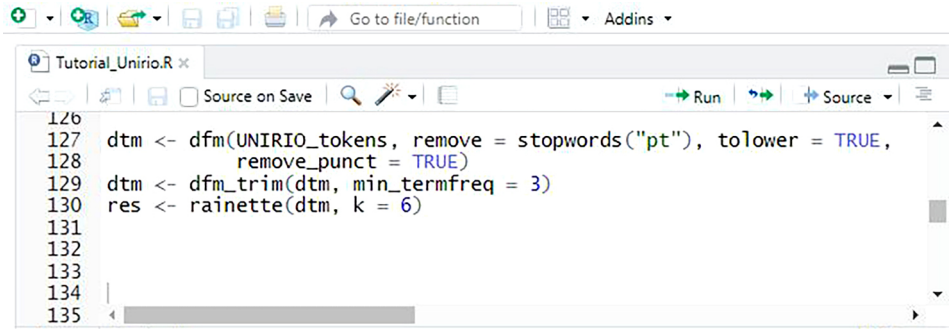
Step 12: Creating DFM (document-feature matrix) – Before creating the wprd co-occurrences netwrok and the *cluster*, we need to create the *document-feature matrix* - DFM. Para executar análises estatísticas, precisamos extrair uma matriz que associa valores para determinados recursos a cada documento. Usamos a função dfm() do pacote Quanteda para produzir essa matriz. "Dfm" é a abreviação de matriz de recursos do documento (*document-feature matrix*) e sempre se refere a documentos em linhas e "recursos" como colunas. Para criar ao DFM, precisamos do seguinte código:



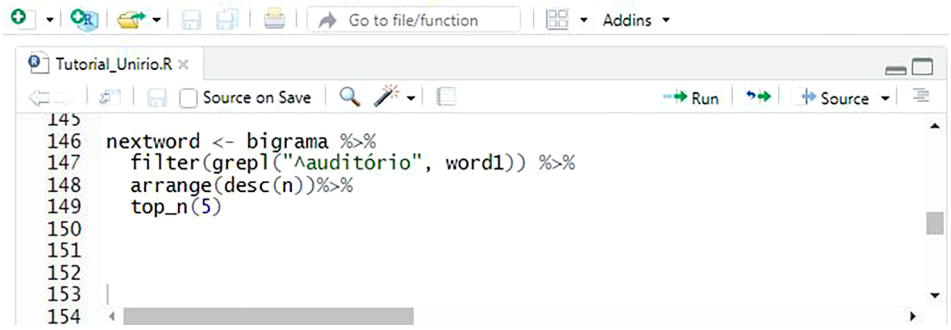
Step 13: Creating word co-occurrence network – In order to create the word co-occurrence network (Graph 1) we use the function textplot_network() from the library Quanteda. Here we can see the possibilities of changing the title, colors and lines.



Step 14: Creating clusters – Here we need a DFM and the rainette function from Rainette library. In the example below, we are creating clusteres in dfm and k=6 shows that we want 6 clusters. Graph 2 presented its results.



Step 15: Creating next word – the prediction of next word is based on the Bigram and had its results presented in tables 4 and 5. This function is not presente in any package; it is necessary to create it. In order to create it we need the following code:



Notice that **filter** is the function to select the words that start with 'auditório'; grepl is a function to find matching, that is, coincidences; Top_n (5) is to show the five words with the greatest association with the audience; ^ is a regular expression. An anchor, that is, the words/expressions that begin with the word 'auditório'; this function selects the words/expressions that begin with the word 'auditório', and show the next word in the bigram.

Annex 2 – Translation of the words and expressions used in the research

acadêmica – academic
alimentos – food
alunos – students
amanhã – tomorrow
ambiental – environmental
Auditório Vera Janacopulos – Vera Janacopulos auditorium
Auditório Tércio Pacitti – Tércio Pacitti Auditorium
Artes cênicas – Scenic Arts, Performing Arts
aula inaugural – inaugural class, opening class
Bandejão – popular name for the restaurant at the university
breve – soon
bioescritas - biowritings
cardápio – menu
científica – scientific
começa – starts, begins
confira – check it out
conferiu – checked
continue acompanhando aqui – follow us to keep informed
convocação – call
copo – glass / cup (in the case - disposable cups)
corrigindo – correcting
coxa – thigh (chicken thigh)
cultura – culture
curta – enjoy, like
curso – course
debate – debate
debateram – debated
descartáveis – disposable
dia – day
economia – economy / economics

edital – announcement
edição – edition
encontro – meeting, get together
ENEM – a type of university entrance exam that is taken by high school students
enfermagem Alfredo Pinto – Alfredo Pinto nursing
escola – school
espinafre – spinach
estatística – statistics
evento – event
excelência – excellence
evite – avoid
fala – speech
filé – filet
fique ligado – Be on to it, stay tuned
fórum – forum
frango – chicken
hospital universitário Gaffrée – university hospital
infecção – infection
iniciação científica – scientific initiation
inscrições – registration
inscrições abertas – open registration
infantil – for children / for kids
leve (verb levar) – take
lista – list
lista de espera – waiting list
matrícula – enrollment
mesa redonda – round table
mestrado profissional – professional Master program
molho – sauce
não perca – Don't miss it
nutrição – nutrition
oferece – offers
página – page
palestra – talk, lecture

participe – participate
popular – popular
pós-graduação – post graduate programs
prato do dia –todays' special
problema – problem, issue
produção – production
programa – program
promove – promote, advertise
projeto quintas culturais – cultural Thursdays project
puder (Se puder) – If you can
recebe – receives, welcomes
rede – network
restabelecer – restore
restaurante – restaurant
saiba – get to know
sala Villa Lobos – Villa Lobos room
saúde – health
Sejam todos bem vindos – You are all welcome
semana – week
seminário – seminar
série Unirio musical – musical Unirio series
série vitrine musical –
siga – follow
sobrecoxa – upperleg (chicken upperleg)
teatro – theater
tema – theme, topic
tutoria – tutoring
tutorial – tutorial
tudo – everything, all
vagas – vacancies, spots
vai rolar palestra – There's gonna be a talk
vai ser – it's going to be
veja – look
você – you

References

AUSSERHOFER, J.; MAIREDER, A. National Politics on Twitter: Structures and topics of a networked public sphere. ***Information, Communication & Society*****,** v. 16, n. 3, 2013. DOI: https://doi.org/10.1080/1369118X.2012.756050

BENITEZ-CABELLO, A.; ROMERO-GIl, V.; MEDINA-PRADAS, E. *et al. Exploring bacteria diversity in commercialized table olive biofilms by metataxonomic and compositional data analysis.* Scientific Reports, v. 10, n. 11381, 2020. DOI: https://doi.org/10.1038/s41598-020-68305-7. Acesso em: 18 jul. 2020

LE LANN, L.; JOUVE, P.; ALARCÓN-RIQUELME, M. *et al. Standardization procedure for flow cytometry data harmonization in prospective multicenter studies*. Sci Rep, v. 10, n. 11567, 2020. DOI: https://doi.org/10.1038/s41598-020-68468-3. Acesso em: 18 jul. 2020

CANTRELL, M. A.; LUPINACCI, P. Methodological issues in online data collection. *JAN*, October, 2007. DOI: https://doi.org/10.1111/j.1365-2648.2007.04448.x

CASSOTTA, M. L. J.; LUCAS, A.; BLATTMANN, U.; GODOY VIERA, A. F. Recursos do conhecimento: colaboração, participação e compartilhamento de informação científica e acadêmica. *Informação & Sociedade: Estudos*, v. 27, n. 1, 25 abr. 2017.

CARTON, G. e MOURICOU, P. Is management research relevant? A systematic analysis of the rigor-relevance debate in top-tier journals (1994–2013). *M@n@gement,* v. 20*,* n. 2, 2017, p. 166-203.

CEZAR, K. G.; SUAIDEN, E. J. O impacto da sociedade da informação no processo de desenvolvimento. *Informação & Sociedade: Estudos*, v.27, n.3, p.19-29, 2017.

CRUZ, B. de P. A.; ROSS, S. D. Caminhos Sinuosos: Os Deslizes nos Estudos em Administração Pública e de Empresas. *RAEP*, v. 19, n. 2, 2018, p. 200-242. DOI: http://dx.doi.org/10.34181/rgb.2019.v2n2.p72-94.52

CRUZ, B. de P. A. Social Boycott. *RBGN,* v. 19, n. 63, 2017. DOI: https://doi.org/10.7819/rbgn.v0i0.2868

CULOTTA, A.; CUTLER, J. Mining Brand Perceptions from Twitter Social Networks. *Marketing Science*, v. 35, n. 3, 2016. DOI: https://doi.org/10.1287/mksc.2015.0968

DATAFOLHA. Coronavírus. Disponível em: http://datafolha.folha.uol.com.br/. Acesso em: 01 jul. 2020.

EVANS, H.; CORDOVA, V.; SIPOLE, S. Twitter Style: An Analysis of How House Candidates Used Twitter in Their 2012 Campaigns. *PS: Political Science & Politics*, v. 47, n. 2, 2014, pp. 454-462, 2014. DOI: https://doi.org/10.1017/S1049096514000389

FREIRE, G. H. DE A.; FREIRE, I. M. Ciência de dados e Ciência da Informação. *Informação & Sociedade: Estudos*, v. 29, n. 3, 30 set. 2019a.

FREIRE, G. H. DE A.; FREIRE, I. M. "As redes são estruturas comunicativas. *Informação & Sociedade: Estudos*, v. 29, n. 2, 2 jul. 2019b.

FREIRE, G. H. DE A.; FREIRE, I. M. Sobre a interdisciplinaridade da Ciência da Informação. *Informação & Sociedade: Estudos*, v. 28, n. 3, 28 dez. 2018.

GENTRY, J. twitteR: R Based Twitter Client. 2015. *Disponível em:* https://CRAN.R-project.org/package=twitteR

GRANELLO, D. H.; WHEATON, J. E. Online Data Collection: Strategies for Research. *Journal of Counseling & Development*, December, 2011. DOI: https://doi.org/10.1002/j.1556-6678.2004.tb00325.x

GUPTA, K.; RIPBERGER, J.; WEHDE, W. Advocacy Group Messaging on Social Media: Using the Narrative Policy Framework to Study Twitter Messages about Nuclear Energy Policy in the United States. *Policy Studies Journal*, August, 2016. DOI: https://doi.org/10.1111/psj.12176

HAND, D. J. e ADAMS, N. M. *Wiley StatsRef: Statistics Reference Online*, 2015. DOI: https://doi.org/10.1002/9781118445112.stat06466.pub2

HOGAN, , B. Online Social Networks: Concepts for Data Collection and Analysis. In Fieldng, N.G., Lee, R., & Blank, G. (eds). *The Sage Handbook of Online Research Methods*, Second edition. Thousand Oaks, CA: Sage Publications, 2017, p. 241-258.

JOATHAN, I.; ALVES, M. O Twitter como ferramenta de campanha negativa não oficial: uma análise da campanha eleitoral para a Prefeitura do Rio de Janeiro em 2016. *Galáxia*, n. 43, 2020, p. 81-98, Apr. 2020. DOI: https://doi.org/10.1590/1982-25532020141565.

KEARNEY, M. W. rtweet: Collecting and analyzing Twitter data. *Journal of Open Source Software*, v. 4, n. 42, 2019. DOI: 10.21105/joss.01829

KOZINETS, R. V. *Netnography: doing Ethnographic Research Online***.** Sage Publications: London, 2010.

LEFEVER, S.; DAL, M.; MATTHÍASDÓTTIR, A. 2 Online data collection in academic research: advantages and limitations. *British Journal of Educational* Technology, June, 2006. DOI: https://doi.org/10.1111/j.1467-8535.2006.00638.x

MYERS, L. e SIROIS, M. J. Spearman Correlation Coefficients, Differences between. *Encyclopedia of Statistical Sciences*, 2006. DOI: https://doi.org/10.1002/0471667196.ess5050.pub2

MORAIS, L. B. V. *As aporias do lugar de fala: como a política identitária afetou a esquerda.* Dissertação (Mestrado em Ciências Política) – Faculdade de Ciências Sociais, Universidade Federal de Goiás. Goiânia. Goiás, p. 187. 2018.

NEUENDORF, K. A. e KUMAR, A. Content Analysis. *The International Encyclopedia of Political Communication*, 1–10, 2016. DOI: https://doi.org/10.1002/9781118541555.wbiepc065

PRAT, C.; MADHYASTHA, T. M.; MOTTARELLA, M. *et al. Relating Natural Language Aptitude to Individual Differences in Learning Programming Languages.* Sci v. 10, n. 3817, 2020. DOI: https://doi.org/10.1038/s41598-020-60661-8. Acesso em: 18 jul. 2020

RAULJI, J. K.; SAINI, J. R. Stop-word removal algorithm and its implementation for Sanskrit language. *International Journal of Computer Applications*, v. 150, n. 2, p. 15-17, 2016

RANGEL, R. C. A jurimetria aplicada ao direito das famílias. *Revista Síntese Direito de Família,* São Paulo, SP: Síntese, v. 15, n. 86, 2014.

R CORE TEAM. R: *A language and environment for statistical computing*. R Foundation for Statistical Computing, 2018, Vienna, Austria. URL https://www.R-project.org/.

RUMMO, P.E.; CASSIDY, O.; WELLS, I.; COFFINO, J. A.; BRAGG, M. A. Examining the Relationship between Youth-Targeted Food Marketing Expenditures and the Demographics of Social Media Followers. *Int. J. Environ. Res. Public Health*, v. 17, n. 3, 2020, 17. DOI: https://doi.org/10.3390/ijerph17051631

SALES, L. F.; SAYÃO, L. F. A grande a a pequena Ciência: análise das diferenças na gestão de dados de pesquisa. *Informação & Sociedade: Estudos*, v. 29, n. 3, 30 set. 2019.

SANTINI, R. M.; SALLES, D.; TUCCI, G.; FERREIRA, F. e GRAEL, F. Making up Audience: Media Bots and the Falsification of the Public Sphere. *Communication Studies,* 2020. DOI: https://doi.org/10.1080/10510974.2020.1735466

SCHOFIELD, A.; MAGNUSSON, M.; MIMNO, D.. Pulling out the stops: Rethinking stopword removal for topic models. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Vol 2, Short Papers. 2017. p. 432-436.

TEIXEIRA, D.; AZEVEDO, I. Análise de opiniões expressas nas redes sociais. *Revista Ibérica de Sistemas e Tecnologias de Informação, v.*8, n.12, 2011, p. 53-65.

TRECENTI, J. A. Z. *Diagramas de influência: uma aplicação em Jurimetria*. 2015. 120 f. Dissertação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2015.

VIRTANEN, P.; GOMMERS, R.; OLIPHANT, T. *et al. SciPy 1.0: fundamental algorithms for scientific computing in Python.* Nat Methods, v. 17, p. 261–272, 2020. DOI: https://doi.org/10.1038/s41592-019-0686-2. Acesso em: 18 jul. 2020

WACHELKE, J.; WOLTER, R. Critérios de construção e relato da análise prototípica para representações sociais. *Psic.: Teor. e Pesq.,* v. 27, n. 4, p. 521-526, Dec. 2011. DOI: http://dx.doi.org/10.1590/S0102-37722011000400017

WICKHAM, H. "Tidy data." *Journal of Statistical Software***,** v. 59, n. 10, 2014, p. 1-23.

WU, X., KUMAR, V., ROSS QUINLAN, J. et al. Top 10 algorithms in data mining. *Knowl Inf Syst*, 14, 1–37, 2008. DOI: https://doi.org/10.1007/s10115-007-0114-2

WU, X.; ZHU, X.; WU, G.; DING, W. Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, v. 26, n. 1, pp. 97-107, Jan. 2014, doi: 10.1109/TKDE.2013.109.