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

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

Steven Dutt-Ross
Breno de Paula Andrade Cruz

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. 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.

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 statistical statistics through an image. This text aims to present this technique for data collection and treatment through the R interface through a tutorial. In this way, public data of Unirio's official account is used as a didactic

Recebido em: 01/07/2020
Aprovado em: 22/09/2020

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

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

ABSTRACT

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.

Ciência e Coleta de Dados no Ambiente *On-line*

O crescente avanço das plataformas digitais que viabilizam as redes sociais virtuais fez com que o processo de comunicação fosse significativamente afetado nos últimos anos – seja no espaço público ou no privado. Freire e Freire (2019a) destacam a importância de se entender a quantidade de informações disponíveis no ambiente virtual ao fazerem um contraponto entre Ciência de Dados e Ciência da Informação. Especificamente, os autores destacam o link entre a Ciência de Dados (principalmente os digitais) e a comunicação científica. A importância de se debruçar sobre as redes sociais se dá, entre outras razões apresentadas por Cezar e Suaiden (2017), pelo fato de que a formação da identidade dos indivíduos passa pelas suas participações nas redes sociais em função da sociedade da informação imprimir um complexo padrão interativo. E, essa interação, muitas vezes ocorre mediada pelas plataformas digitais.

Existem ferramentas que podem auxiliar o mercado na tomada de decisão e a produção de conhecimento na academia. Historicamente, *softwares* como Excel, Word, SPSS e Stata ajudaram pesquisadores a produzirem conhecimento em suas áreas; e, recentemente, *softwares* cada vez mais sofisticados surgiram – inclusive para as pesquisas qualitativas. Mas eles geralmente têm em comum uma característica: são licenciados por empresas e precisam ser pagos. A linguagem R surge como uma possibilidade gratuita para pesquisadores programarem e desenvolve-

rem análises estatísticas – inclusive via plataformas digitais com API (*Application Programming Interface*) - como é o caso do Twitter.

O R é uma linguagem de computação entre seus usuários e tem forte ênfase no tratamento de dados estatísticos. Nela se programa um código para se extrair informações de uma base de dados. Além de ser grátis e ter código livre, a linguagem R permite reproduzir os resultados bem como todos os dias são desenvolvidas novas bibliotecas (*library*) - a rede de usuários é grande, por isso, é fácil conseguir ajuda e tirar dúvidas no uso da linguagem. Se um pesquisador em Londres construir um código, um pesquisador aqui no Brasil consegue reproduzir tudo que foi feito se a base de dados estiver disponível; ou, então, replicar o modelo com novos dados primários. A maioria das pesquisas de ponta são desenvolvidas no R ou no Python, conforme pode ser verificado nos estudos recentes publicados na Nature - Le Lan et al (2020), Virtanen et al (2020), Prat et al (2020) e Benítez-Cabelo et al (2020).

Pesquisadores(as) mais atentos(as) e interessados(as) em técnicas como esta que será aqui apresentada podem se perguntar o porquê de utilizarmos a linguagem R e não os *softwares* Iramuteq ® ou Alceste ®. O Iramuteq ® utiliza a linguagem R para executar as suas análises; todavia, funciona na versão clássica do R (versão 3.1.2 de 2014). Ao não utilizar a versão mais atual do R (versão 4.0.1 3 de 2020), o Iramuteq ® se torna obsoleto e defasado em 6 anos no que diz respeito ao processamento de textos. Já o Alceste é um *software* proprietário; ou seja, tem um dono (direitos autorais). E, mais que isso, tanto o Alceste ® quanto o Iramuteq ® não oferecem os *outputs* que apresentamos aqui como recursos, tais como (i) previsão da próxima palavra, (ii) a rede de coocorrência dos termos, (iii) bigramas e (iv) trigramas. Em resumo, embora o Iramuteq ® seja gratuito como a linguagem R, ele é defasado em relação ao que propomos aqui.

A publicação espontânea de pessoas físicas ou jurídicas nas plataformas digitais como o Twitter possibilita um volume grande de informações que podem ser trabalhadas no R. Os dados se encontram presentes em diversas plataformas digitais e são fruto da explosão comunicacional que emerge da *Web2* (FREIRE; FREIRE, 2019b). É interessante notar que a sobrecarga informacional na *web* não se dá apenas para plataformas digitais populares como *Facebook* e Twitter – isso acontece também no processo de compartilhamento da produção científica, conforme apontam Cassotta *et al.* (2017).

Como uma estratégia de manter páginas e websites atualizados, muitas são as organizações que usam de páginas em mídias digitais como *Twitter, Facebook, Instagram* e *WhatsApp* para se comunicarem com clientes, fornecedores e outros *stakeholders*. Fato é que essa interação parece ser mais efetiva que antes - mesmo que apresentem problemas que possam manchar a imagem ou a reputação de uma empresa (CRUZ, 2017). Diversos estudos apontam o uso de dados coletados no ambiente *on-line* e para fazer Ciência em diferentes campos do saber (HOGAN, 2017; GRANELLO; WHEATON, 2011; LEFEVER, DAL; MATTHÍASDÓTTIR, 2006; CANTRELL; LUPINACCI, 2007).

Há de se pontuar na demarcação deste texto que nosso objetivo não é discutir mineração de dados. A mineração de dados (*data mining*) é que cria o *corpus* (nossa base de dados de palavras); e aqui na Análise Quantitativa de Textos a mineração de dados é apenas o primeiro passo pois nosso foco é apresentar *outputs* estatísticos a partir das palavras, tais como a nuvem de palavras, bigramas, trigramas, previsão da próxima palavra, análise de cluster e rede de coocorrência dos termos. Leitores que busquem maior profundidade nas técnicas relacionadas à mineração de dados podem consultar os trabalhos de Wu *et al.* (2008), Hand e Adams (2015) e Wu *et al.* (2014).

Da mesma forma, demarcamos que nossa proposta neste texto não é discutir se devemos adotar Análise de Conteúdo Quantitativa (NEUENDORF; KUMAR, 2016) ou outros métodos qualitativos: aqui apresentamos a Análise Quantitativa de Textos a partir de dados públicos no Twitter – também não se caracterizando como Netnografia. E, especificamente em relação às pesquisas qualitativas na área de Administração Pública e de Empresas, o texto de Cruz e Ross (2018) sinaliza uma importante reflexão para alguns estudos qualitativos que utilizam dados coletados no ambiente virtual e classificam como Netnografia. Os autores realçam o rigor metodológico da Netnografia enquanto Método e criticam estudos que apenas coletam dados nas plataformas digitais e classificam como Netnografia. Desta forma, a técnica que apresentaremos aqui não é (i) um método e (ii) e também não é uma técnica de coleta de dados que permite ser parte de um estudo netnográfico.

A Netnografia é um método qualitativo que implica em uma imersão em uma comunidade virtual e com interação entre pesquisador(a) e a comunidade virtual investigada (KOZINETS, 2010). O que é proposto aqui é exemplificar o uso de uma técnica de coleta de dados (com viés puramente quantitativo) que pode ser utilizada

para complementar a análise de dados em estudos quantitativos e qualitativos; ou, adicionalmente, ajudar a gerar *insights* a partir de um banco de dados existente.

Nesta discussão sobre método e tendo como objetivo a apresentação de uma técnica neste texto, vamos ao Dicionário de Filosofia (JAPIASSÚ; MARCONDES, 1996) para dar robustez ao nosso argumento central de que apresentamos aqui uma técnica e não um método. Assim, temos:

> **Método** – conjunto de procedimentos racionais, baseados em regras, que visam atingir um objetivo determinado. Por exemplo, na Ciência, o estabelecimento e a demonstração de uma verdade científica (p. 181).
>
> **Técnica** – conjunto de regras práticas ou procedimentos adotados em um ofício de modo a se obter os resultados visados. (...) Em sentido derivado sobretudo da ciência moderna, aplicação prática do conhecimento científico teórico a um campo específico da atividade humana (p. 257).

Aqui adotamos a perspectiva de técnica. Todavia, não é simplesmente uma técnica de coleta de dados (como o uso de um grupo focal). É sim uma técnica de coleta de dados por trabalhar com a mineração de dados, mas é principalmente uma técnica de análise de dados ao se apresentar os conceitos de bigramas, trigramas e n-gramas, Rede de Correlação de Palavras e Análise de Cluster. Assim, a Análise Quantitativa de Textos é uma técnica que envolve a coleta de dados qualitativos em uma plataforma digital de conteúdo público para realizar uma análise quantitativa destes conteúdos por meio da apresentação de análise de correlação entre os termos utilizados pelos usuários.

A presença da plataforma digital Twitter é uma realidade no processo de comunicação entre pessoas, entre organizações e entre pessoas e organizações. Criada em 2006, ao permitir o compartilhamento de textos, fotos, vídeos e principalmente o uso de *hashtags* (símbolo #), se tornou uma importante ferramenta de comunicação no planeta, visto que é usada por políticos (AUSSERHOFER; MAIREDER, 2013), empresas (CULOTTA; CUTLER, 2016), ONGs (GUPTA; RIPBERGER; WEHDE, 2016), por celebridades e por anônimos que se tornam celebridades.

Nesse sentido, a análise dos textos postados permite fazer uma avaliação de como se dá esse tipo de interação no Twitter pelas pessoas e organizações diversas na sociedade. No Brasil, por exemplo, política e televisão são tópicos sempre quentes no Twitter (SANTINI *et al.,* 2020). Mas existem também outros temas que merecem destaque – como recentemente o combate ao Coronavírus. Para uma

instituição de ensino pública ou privada, por exemplo, é interessante entender o que alunos e a sociedade escrevem sobre a organização pois assim é possível (re) pensar a imagem e a reputação por meio da comunicação institucional.

Outro exemplo, já na perspectiva da Administração Pública, é entender como se deu a avaliação do Ministério da Saúde e do presidente Jair Bolsonaro pelos usuários do Twitter em meio a crise do Coronavírus. Embora o DataFolha tenha divulgado resultados parciais sobre a avaliação do presidente e dos ministros que estiveram a frente da pasta (DATAFOLHA, 2020), há de se considerar que o comportamento dos usuários no Twitter é diferente daqueles que não usam a plataforma. Assim, subir as *hashtags* (colocar em evidência um assunto nas mídias sociais) é uma estratégia de chamar atenção da mídia para assuntos que os grupos, movimentos sociais e usuários consideram importantes.

Desta forma, o objetivo deste estudo é estruturar e sistematizar fases e o passo a passo da Análise Quantitativa de Textos conduzida na linguagem R como técnica de análise de dados. Especificamente, temos: (a) o uso de dados públicos da conta da Universidade Federal do Estado do Rio de Janeiro (Unirio) no Twitter para apresentar exemplos reais da operacionalização da técnica – não necessitando de uma autorização formal da instituição; e, (b) a discussão sobre a importância de atualizar as técnicas na condução de estudos no campo da Administração Pública e de Empresas com o uso de uma linguagem de programação gratuita como o R. A próxima seção apresenta a operacionalização da Análise Quantitativa de Textos.

A Operacionalização da Análise Quantitativa de Textos

Para exemplificar o uso desta técnica neste trabalho, fizemos um primeiro recorte para a região metropolitana do Rio de Janeiro e quatro instituições federais de ensino superior poderiam ser consideradas: Universidade Federal do Rio de Janeiro (UFRJ), Universidade do Estado do Rio de Janeiro (UERJ), Universidade Federal do Estado do Rio de Janeiro (Unirio) e Universidade Federal Fluminense (UFF). Num total de mais de 34 mil publicações até Dezembro de 2019, trabalhamos com a amostra intencional (*purposive sample*) e escolhemos a Unirio (conta @comunicaUNIRIO), com 4.768 publicações, por termos familiaridade com a instituição – o facilita a interpretação dos resultados na aplicação da técnica.

FASE 1 – COLETA DE DADOS VIA R-TWEET

Existem duas maneiras de capturar dados da Internet: (i) raspagem de dados e via API. Na raspagem de dados cria-se uma rotina para capturar dados da Internet. Só que muitas vezes é ilegal essa raspagem de dados porque trata-se de uma clonagem de dados que pode ser comercializada (e existe mercado para isso). Já o API é um conjunto de rotinas e padrões de programação que permite acessar um aplicativo/plataforma. O R-Tweet é um sistema externo que consulta dados na plataforma Twitter por meio de uma integração via API. Como é um acesso permitido pelo Twitter, não há ilegalidade ou conflito ético do pesquisador (os dados existem, podem ser consultados e utilizados).

O R-Tweet é uma biblioteca do R para acessar a API do *Twitter* (KEARNEY, 2019) e permite baixar até 3.200 tweets recentes de uma determinada conta. Por exemplo, a captura das mais de três mil postagens da Unirio foi realizada por meio deste pacote. Para Kearney (2019), o twitteR se diferencia do R-Tweet em função da possibilidade de buscar usuários, palavras-chaves e pode capturar status (como do *Facebook* na atualização de status).

Nesta coleta de dados inicial nos concentramos somente nas publicações principais. Ou seja, foram excluídos todos os *retweets* e as repostas (*replies*) dadas pelos seguidores da conta ou de usuários do *Twitter* a uma postagem. Para isso, usamos também o pacote *twitteR* (GENTRY, 2015) – um pacote para capturar o *corpus* no Twitter e trabalhar na linguagem R (por isso a letra R maiúscula). Esses textos foram salvos no formato de texto (.txt) e podem ser baixados aqui[1]. Em seguida, foram utilizadas as informações do *Twitter* da Unirio. A coleta de dados ocorreu no dia 03 de Janeiro de 2020, sendo que a captura do *corpus* para a conta @comunicaUNIRIO totalizou 3.140 postagens a partir do dia 14 de outubro de 2015.

FASE 2 – LIMPEZA DE TEXTO

Aqui podemos verificar que o *Twitter* da Unirio usou 8.984 palavras diferentes nas últimas 3.140 postagens entre 2015 e 2019. Os dados foram capturados no início de 2020 e levantamos todas as postagens no período entre 10/2015 e 12/2019. Todavia, conforme pode ser identificado na Figura 1, os comentários aparecem com emojis, preposições e conectores textuais que devem ser eliminados para que

1 https://bit.ly/3jCCndJ

de fato se consiga realizar uma associação entre as palavras. Se não ocorre essa limpeza, estas palavras irão aparecer nas fases seguintes nas primeiras posições dos *rankings* e podem negligenciar outras palavras ou expressões realmente importantes (RAULJI; SAINI, 2016; SCHOFIELD; MAGNUSSON; MIMNO, 2017).

Figura 1 Exemplo do banco de dados – 2015/2019

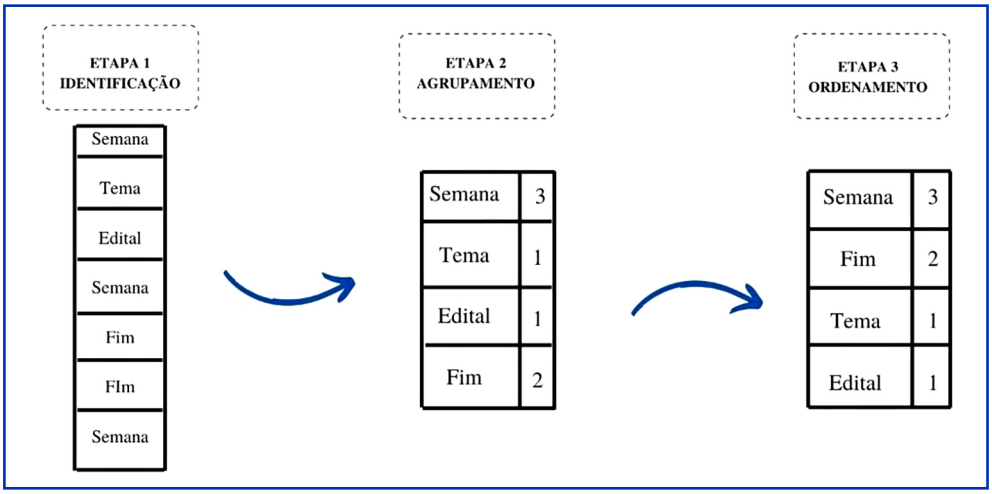
	Date	Tweet
1	2019-12-16 13:53:00	@sleepinthdirt Vem comigo!
2	2019-12-16 13:43:00	Hoje tem filézin de frango com purê de batata. Amanhã tem carne ensopada ou almôndeg...
3	2019-12-13 19:49:00	@PAlinhares @uff_br UNIRIO, PÔ!!!!!
4	2019-12-13 10:09:00	#pracegover Imagem com fundo azul comemorativa do Dia 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...

Fonte: dados da pesquisa.

Após baixar e carregar os dados, foi realizada uma rotina para limpeza dos dados. Isso acontece porque temos muitas palavras com pouco significado como, por exemplo, “de”, “da”, “que” pois consideramos estas palavras de conexão. Em função disso, foi necessário adotarmos o procedimento para deletá-las. Assim, foram mantidos os substantivos, verbos e adjetivos; sendo eliminadas os pronomes, artigos, numerais, preposições, conjunções, interjeições e advérbios. Adicionalmente, também temos palavras ou expressões da *web* com nenhum significado e que são resultantes dos *hiperlinks* do *Twitter*. Esta categoria foi nomeada como linguagem computacional e palavras como “https”, “http”, “www” e “#” também foram excluídas.

Para operacionalizar essa limpeza de dados, precisamos colocar todos os textos no formato Tidy (WICKHAM, 2014) - um banco de dados com cada palavra em uma linha. Após essa etapa, precisamos verificar a frequência de cada palavra. Para demonstrar este formato, dividimos esse método em 3 etapas: Etapa 1 – Identificação das palavras; Etapa 2 – Agrupamento das palavras iguais; Etapa 3 – contagem de palavras iguais e ordenamento em um *ranking*. A Figura 2 apresenta um banco de dados no formato Tidy.

Figura 2 Etapas de inserção dos dados no formato Tidy



Fonte: elaboração dos autores.

A Tabela 1 apresenta as 10 palavras com maior frequência antes e depois deste procedimento de limpeza. Ou seja, destas 10 palavras apenas "Dia" poderia ser alguma palavra que trouxesse algum significado – poderia ser o "prato do dia no bandejão" ou "dia de" algum evento na instituição. Aqui, por meio destes resultados, destacamos a importância de realizar essa limpeza de palavras sem significado. Como é visualizado, outras palavras são identificadas após essa limpeza e a lista de palavras banidas pode ser visualizada aqui neste link[2].

2 https://bit.ly/3kGo7BZ

Tabela 1 As 10 palavras com maior frequência no Twitter da Unirio entre 2015 e 2019 antes e depois da limpeza de palavras e expressões sem relevância

Antes da Limpeza		Depois da Limpeza	
Palavra	**Frequência**	**Palavra**	**Frequência**
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
Dia	985	semana	150
Do	880	tema	147
Unirio	819	edital	138

Fonte: Coleta de dados da pesquisa para a conta @comunicaUNIRIO

Percebemos aqui, após banir algumas palavras e expressões, que a palavra Unirio saltou da 10ª posição para a 1ª posição no *ranking*. Analisando inicialmente os termos que aparecem na Tabela 1 após banir alguns termos, podemos perceber que eles estão associados ao contexto de uma universidade por apresentar palavras como escola, palestra, edital, saiba e tema. Assim, o uso das palavras "confira", "saiba", "inscrições", "palestra" e "edital" também sugere que o *Twitter* oficial da Unirio é utilizado para uma agenda de divulgação da instituição. Possivelmente, o grande uso (150 vezes) da palavra "semana" tem a ver com a Semana de Integração Acadêmica - SIA (um grande evento da universidade). Interessante notar que as palavras associadas a Jornada de Iniciação Científica - JIC (outro grande evento da instituição) não estão entre as dez primeiras. A nuvem de palavras é apresentada na Figura 3.

Figura 3 Nuvens de palavras após a limpeza dos dados

Fonte: Coleta de dados da pesquisa para a conta @comunicaUNIRIO

INICIANDO AS ANÁLISES ESTATÍSTICAS

O R é uma linguagem utilizada por estatísticos de todo o mundo para apresentar modelos simples ou sofisticados por meio de suas bibliotecas (library) – e aqui neste artigo metodológico utilizamos duas relevantes bibliotecas, são elas: Quanteda e Tidetext. Logo, é utilizado por pesquisadores quantitativos e que fazem pesquisa quantitativa com robustez teórica e metodológica, por meio de outras bibliotecas como *OpenNLP, Rweka, RcmdrPlugin.temis, tm, languageR,, koRpus, RKEA, lsa e maxent*. Dessa maneira, a análise de textos aqui apresentada é quantitativa (e não qualitativa). Os resultados que são apresentados são fruto de métodos estatísticos programados no R.

Bigramas

Os n-gramas são os elementos adjacentes em uma sequência de palavras mineradas em um texto e são reconhecidos e gerados estatisticamente. Assim, eles

podem ser classificados pela sua quantidade de combinações, sendo um bigrama para duas combinações, um trigrama para três combinações, um tetragrama para quatro combinações e assim sucessivamente até se chegar aos poligramas.

Um bigrama é uma sequência de **dois elementos adjacentes** de uma sequência de símbolos (*tokens*). Um Bigrama é um n-grama para n = 2. Com um bigrama procuramos responder a seguinte pergunta: 'Que palavras ficam mais vezes juntas?'. Essa estratégia pode ser utilizada para qualquer banco de dados ou conta oficial do Twitter ou de outra plataforma que tenham dados públicos. Para construí-lo, é necessário usar a função unnest_tokens do pacote tidytext (Silge, Robinson,2016). Essa função divide uma tabela em um token por linha. O Anexo 1 (Tutorial para Replicação da Técnica Análise Quantitativa de Textos) apresenta todos os códigos utilizados neste artigo. Com ele, é possível passar por todas as etapas para replicar o passo a passo desta técnica.

Note que Unirio foi a primeira palavra na nuvem de palavras (Figura 3), mas ela não aparece nos bigramas e trigramas porque a Figura 3 é a representação visual da frequência simples de cada palavra. Já bigramas e trigramas buscam associações de palavras e não somente a frequência simples de palavras.

No caso da comunicação institucional da Unirio, buscamos verificar a associação de palavras e a frequência dessa associação. Construímos todos os bigramas possíveis a partir das 3.140 postagens da Unirio. Depois de criar os bigramas, precisamos verificar a frequência de cada um. Após essa etapa, ordenamos pela frequência, os dez bigramas mais frequentes são apresentados a seguir na Tabela 2. Note que o bigrama 'pós graduação' pode ser analisado apenas como uma expressão de um termo separado pelo hífen. Assim, ao realizar a limpeza do texto, se retiramos o hífen de pós-graduação ou de mesa-redonda (que sugere o debate e não o formato físico da mesa), um resultado diferente pode ser apresentado e isso sugere uma limitação da técnica.

Tabela 2 Os dez primeiros bigramas dos *tweets* da Unirio 2019

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

Fonte: Coleta de dados da pesquisa para a conta @comunicaUNIRIO

Interessante notar que a divulgação do restaurante escola é a atividade mais comum da Unirio. Depois da apresentação do Cardápio, podemos ver uma universidade preocupada com a agenda de pesquisa com uma grande frequência das palavras “pós-graduação”, “mestrado profissional”, “mesa redonda” e “iniciação científica”.

Em seguida, percebe-se uma agenda de divulgação dos eventos da Unirio com as palavras “fique ligado”,”inscrições abertas”,”Auditório Vera”, “aula inaugural”, “série Unirio”. Finalmente, podemos ver também uma agenda cultural com as palavras “quintas culturais”,”Unirio musical” e “artes cênicas”.

O mesmo procedimento apresentado acima é utilizado para a construção de Trigramas e outros poligramas e pode também ser replicado por meio do Anexo 1. Assim, a programação é a mesma e uma palavra vai se juntando a outras por meio da identificação e seleção dos n-gramas via estatística no R. A Tabela 3

apresenta um recorte apenas para o ano de 2019 para os trigramas encontrados a partir da conta da Unirio.

Tabela 3 Os dez primeiros trigramas dos *tweets* da Unirio 2019

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

Fonte: Coleta de dados da pesquisa para a conta @comunicaUNIRIO

Conhecer o contexto ou a teoria é importante para analisar o n-gramas que surgem dos resultados apontados pela técnica Análise Quantitativa de Textos. Por exemplo, um leitor distante da realidade da Unirio pode possivelmente não entender o primeiro trigrama “Sala Villa Lobos” e outros como “Série Unirio Musical”. Aqui, um dos autores sendo parte da instituição, podemos supor que tanto a sala quando

a série estejam ligadas ao curso de Bacharelado em Teatro. Assim, poderíamos por meio dessa interpretação supor que exista um grupo de trigramas que esteja relacionado à 'Arte'. É possível ainda supor que em uma universidade com 25 cursos de graduação, talvez os cursos de Teatro, Enfermagem e Medicina sejam aqueles com maior destaque na comunicação oficial da instituição.

Quando nos referimos à Medicina e Enfermagem como cursos com destaque na comunicação oficial da Unirio pelo Twitter, consideramos os trigramas "infecção HIV AIDS ", "Enfermagem Alfredo Pinto", "Alfredo Pinto - EEAP" (nome da escola de enfermagem) e "Hospital Universitário Gaffrée" – e todos eles relacionados à área de Saúde.

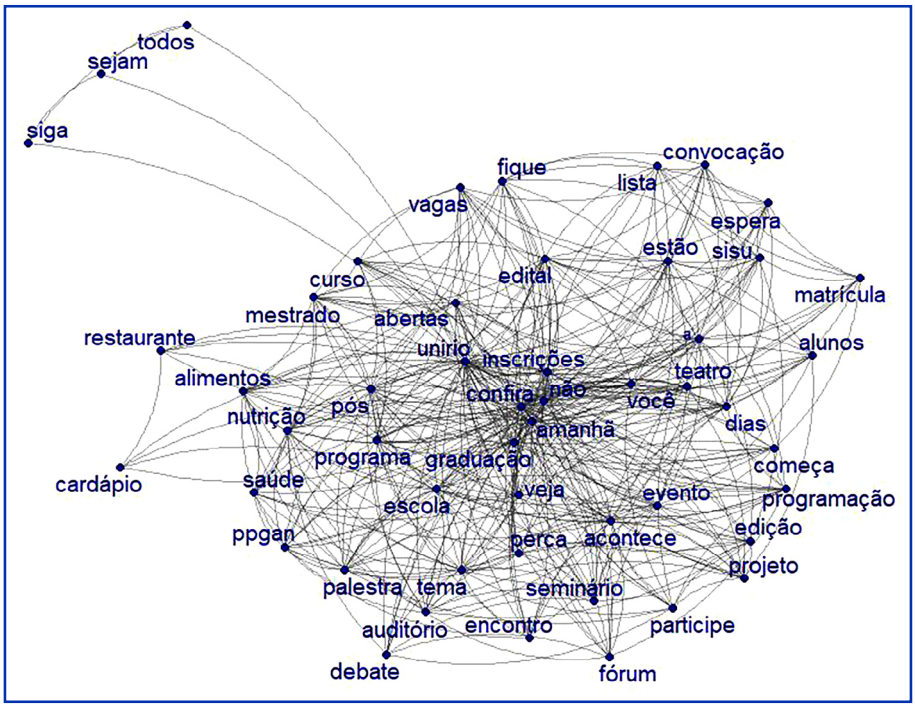
Um terceiro grupo que podemos sugerir a partir de uma análise qualitativa destes trigramas que emergem neste caso específico é 'Agenda de Divulgação'. O auditório que forma o trigrama 'Auditório Vera Janácoulos' evidenciam essa análise, bem como os trigrama "vai rolar palestra" e "sala Villa Lobos". E considerando que o Twitter é mais utilizado por jovens (RUMMO *et al.,* 2020) do que outras gerações anteriores, faz sentido entender esta conta oficial da Unirio com destaques para o que vai acontecer em relação à Arte e Saúde para a graduação.

Rede de Coocorrência de Palavras

Apresentamos, a seguir, uma ferramenta para gerar automaticamente um resumo visual de dados de texto não estruturado. Ao contrário de ferramentas como nuvens de palavras, nesta visualização buscamos ver as estruturas de relacionamentos entre palavras – enquanto na nuvem de palavras a maior palavra representa a maior frequência, na Rede de Coocorrência de Palavras essas relações são determinadas por uma abordagem da análise de coocorrências.

Assim, o algoritmo aplica uma função decrescente à distância entre os pares de palavras encontrados, de modo que as palavras que ocorram regularmente próximas uma da outra tenham uma pontuação alta (quanto menor a distância, maior o valor); mas mesmo palavras que ocorram a alguma distância farão uma pequena contribuição para a coocorrência geral. Para traçar uma rede de coocorrências de palavras utilizamos a função textplot_network do Pacote Quanteda (BENOIT *et al*., 2018). O passo 3 do tutorial apresenta a programação realizada e o Gráfico 1 apresenta os resultados do caso da Unirio neste texto.

Gráfico 1 Correlação de Palavras (Coocorrências)



Para o Gráfico 1, visualizamos que as palavras “siga”, “sejam” e “todos” estão distantes das demais. Isso se justifica porque ao analisar as postagens da Unirio verificamos existir um preâmbulo nas mensagens, é ele:“sejam todos bem vindos”; terminando as mensagens com a frase “siga o @Unirio” – em referência à conta da instituição no Twitter. Logo, é por isso que essas três palavras estão associadas e elas têm co-ocorrencia (aparecem sempre próximas – quando uma aparece, possivelmente as outras também aparecem). Nesta rede, é possível visualizar que estas três palavras estão associadas à palavra Unirio – que se associa a todas as demais palavras; o que evidencia que a palavra Unirio é a principal palavra nesta rede de palavras (o que corrobora os achados da nuvem de palavras).

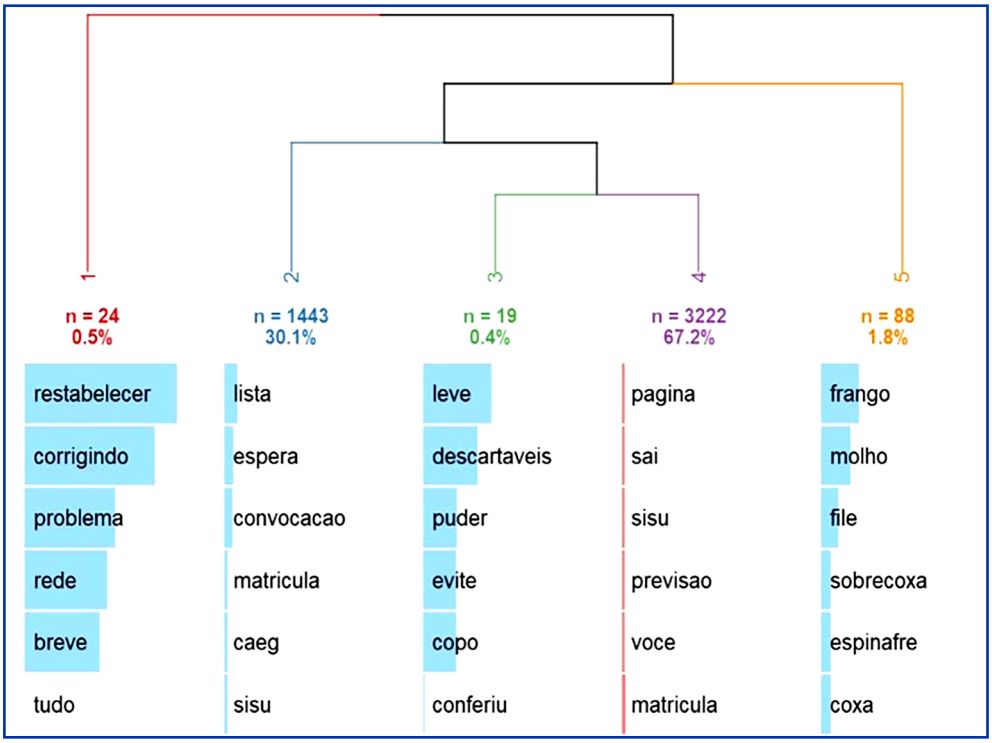
Essa foi uma primeira análise possível no Gráfico 1 e ela pode ser estendida para outras palavras que aparecerem em um gráfico como este que é fruto da técnica aqui apresentada. Por exemplo, a palavra “restaurante” está relacionada às palavras “cardápio”, “alimento” e “nutrição”. Da mesma forma, a palavra aluno

está associada à matrícula – que por sua vez está associada às palavras Sisu, lista, espera e convocação. Em relação a trabalhos empíricos, a teoria escolhida ou o conhecimento do campo poderá ajudar a explicar a coocorrência de palavras.

Análise de Cluster

A análise de *cluster* é uma técnica estatística usada para classificar elementos em grupos, de forma que elementos dentro de um mesmo conglomerado (*cluster*) sejam parecidos, e os elementos em diferentes *clusters* sejam distintos entre si (MYERS; SIROIS, 2006). Para definir a semelhança – ou diferença – entre os elementos é utilizada uma função de distância. Nesta abordagem foi utilizada a distancia euclidiana. Além disso, cumpre registrar que adotamos o método hierárquico. O resultado é apresentado no gráfico 04 abaixo e sua operacionalização no Anexo 1.

Gráfico 2 Dendograma com os resultados da Análise de *Cluster* para a conta da Unirio no Twitter



A partir do Gráfico 2 podemos realizar a interpretação dos grupos que foram identificados na Análise de Cluster. Cada grupo apresenta a sua frequência (n) e o seu percentual dentro do banco de dados. Perceba que as cinco primeiras palavras de cada grupo contribuem significativamente para podemos nomear cada grupo. Assim, tem-se:

- **Grupo 1 – Problemas:** Tem palavras como 'restabelecer', 'corrigindo' e 'problema'. Isso indica que o *Twiitter* da Unirio informa problemas que acontece dentro do espaço da Universidade. Por exemplo, falta de internet, falta de água, falta de energia.
- **Grupo 2 – Convocação de Novos Alunos:** Essa categoria está relacionada à convocação de novos estudantes para ingressarem na Unirio, evidenciando aspectos relacionados à matrícula, lista de espera e informações gerais sobre a convocação pelo Sisu.
- **Grupo 3 – Seja Sustentável no Bandejão:** Nesta categoria se destaca a grande quantidade de informações referentes à uma postura mais sustentável dos usuários do restaurante universitário, uma vez que se pede para levar seus copos e evitar a produção de lixo por meio dos copos descartáveis.
- **Grupo 4 - ENEM:** Essa categoria tem as palavras que estão relacionadas ao ENEM, desde matrícula como lista de espera, convocação e edital. A peculiaridade das palavras aqui não ajuda entender de fato qual é a principal característica que diferencia o grupo 2 deste grupo e uma análise de sentimentos via R poderia ser aplicada para verificar se a diferença entre os grupos se dá por meio dos sentimentos associados às palavras destes dois grupos. Essa possível similaridade nestes dois grupos no caso deste exemplo aqui utilizado evidencia como a teoria pode ajudar a interpretar os resultados da técnica Análise Quantitativa de Textos aqui apresentada.
- **Grupo 5 - Prato do dia:** Esta é a categoria mais bem definida, com a descrição de diversos alimentos que compõem o prato do dia no restaurante universitário da Unirio.

Com base nesta Análise de Cluster aqui conduzida, saímos de uma análise subjetiva do que possivelmente podemos encontrar na conta da Unirio e temos

uma análise racional-estatística sem interferência dos valores pessoais de um pesquisador, por exemplo. Os resultados deste caso aqui apresentado por esta técnica podem inclusive trazer implicações gerenciais que podem ser pensadas pelos gestores da referida instituição em relação à comunicação institucional. A próxima seção apresenta o recurso 'previsão da próxima palavra'.

Previsão da Próxima Palavra

O recurso de autocompletar do Google é um exemplo de previsão da próxima palavra e será apresentado aqui como um dos resultados possíveis para a Análise Quantitativa de Textos. Assim, quando estamos escrevendo uma busca no Google, ele parece adivinhar a próxima palavra. Também podemos criar um modelo para prever isso a partir de uma base de dados. Por exemplo, quando a comunicação da Unirio usa a palavra "Unirio", qual palavra poderia vir a seguir? Assim, a partir da Tabela 4, evidenciamos que é possível prever a próxima palavra que a comunicação oficial da Unirio utilizaria.

Tabela 4 Exemplo para previsão da próxima palavra a partir da palavra 'Unirio'

Palavra 1	Palavra 2	Frequência Absoluta	Frequência Relativa (%)
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

Fonte: elaboração dos autores a partir da base de dados.

Após a criação do bigrama, sequência de dois elementos adjacentes, podemos utiliza-los para prever a próxima palavra. Para isso, é necessário fixar a primeira palavra do bigrama – fizemos isso na Tabela 4. A palavra modal (que mais se repete) após o termo "Unirio" é a "musical". Baseada na frequência relativa, essa palavra é mais provável de aparecer quando o termo Unirio é usado pela conta oficial da instituição.

Para esse método funcionar, é necessário a lematização de todas as palavras (tratamento de equivalência). A lematização é o processo de agrupar as formas flexionadas de uma palavra para que possam ser analisadas como um único item, identificado pelo lema (WACHELKE; WOLTER, 2011) – por exemplo, bela, belo e beleza estariam agrupadas em uma mesma categoria. Em muitos idiomas, as palavras aparecem em várias formas flexionadas. Por exemplo, o verbo 'andar' pode aparecer como 'andar', 'marchar',' percorrer' e 'caminhar'. A forma básica, 'andar', que pode ser encontrada em um dicionário, é chamada de lema da palavra. Assim, todas essas palavras foram substituídas por "andar" (forma básica).

Por se tratar de uma universidade pública, replicamos os bigramas para as palavras 'Educação' e 'Pesquisa' a fim de verificar quais as palavras poderiam estar associadas a esses termos. Os resultados da conta oficial da Unirio podem ser visualizados na Tabela 5.

Tabela 5 Previsão da próxima palavra para os termos 'Educação' e 'Pesquisa'

Palavra 1	Palavra 2	Frequência	Palavra 1	Palavra 2	Frequência
	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

Fonte: elaboração dos autores a partir da base de dados.

Nesta seção evidenciamos os principais *outputs* gerados a partir da Análise Quantitativa de Textos como técnica de pesquisa neste artigo metodológico. Se

o Iramuteq ® traz como *outputs* a nuvem de palavras e a análise de cluster, aqui por esta técnica fomos além ao apresentarmos os bigramas, trigramas, a previsão da próxima palavra e a rede de coocorrência de termos desta técnica de pesquisa quantitativa programada no R. Todos estes outputs de codificação aberta podem ser reproduzidos por meio do tutorial que disponibilizamos adicionalmente.

Considerações Finais

A disponibilidade de dados em diversas plataformas e mídias digitais é uma realidade do novo contexto das pesquisas em Ciências Sociais Aplicadas nas últimas duas décadas. Se antes uma coleta de dados muitas vezes envolvia um planejamento sistemático de contato com fontes e organizações, hoje muitos dados são obtidos por meio de portais públicos ou privados. No Portal da Transparência (Controladoria-Geral da União) é possível obter diversos dados que podem ser utilizados em diferentes pesquisas na Administração Pública; no site Reclame Aqui é possível entender a imagem de uma empresa por meio das reclamações de consumidores; no Twitter é possível analisar inclusive discurso de políticos.

Assim, em um momento em que o Brasil sofre com uma tentativa de sucateamento das universidades públicas e da Ciência (embora para alguns esse sucateamento exista há anos), encontrar novas estratégias e empregar novas técnicas que permitem análises de um grupo maior de observações é uma eficiente tática a ser adotada por pesquisadores(as). Por isso, consideramos relevante ter apresentado aqui a Análise Quantitativa de Textos por meio da linguagem R usando o Twitter como *corpus*.

É importante destacar que o objetivo da Análise Quantitativa de Textos não é substituir qualquer outro método qualitativo como a Análise de Discurso, a Etnografia, a Análise de Conteúdo (que também tem sua abordagem quantitativa); ou substituir outra técnica de coleta de dados (Observação Participante, Grupo Focal ou Observação Não Participante). Nosso objetivo foi evidenciar que esta técnica pode inclusive estar associada às técnicas e métodos qualitativos a fim de tornarem mais robustos os resultados de algumas pesquisas. Sendo uma técnica recente que vem sendo utilizada por estatísticos, para uma ciência predominantemente Positivista como a Administração - de acordo com alguns autores como

Carton e Moricou (2018) – a Análise Quantitativa de Textos parece ser interessante para se estudar um mesmo fenômeno com outra lupa: a da pesquisa quantitativa nas plataformas digitais.

Embora a Análise Quantitativa de Textos tenha sido apresentada aqui por meio da plataforma *Twitter*, ela pode ser utilizada para dados que estejam no ambiente *off-line*. Por exemplo, tem crescido a utilização desta técnica na área Jurídica – sendo chamada de Jurimetria. Trecenti (2015) e Rangel (2014) apontam que é possível pelas sentenças de juízes prever por meio da análise quantitativa de textos como poderão ser suas próximas sentenças – a decisão em si. Alguns outros estudos analisam discursos de presidentes e demais políticos (JOATHAN; ALVES, 2020; EVANS; CORDOVA; SIPOLE, 2014).

As palavras têm poder, mais até que os números (mas isso muitas vezes ficou negligenciado em estudos com viés Positivista). Pelas palavras é possível entender as ideologias, valores e crenças do emissor de uma mensagem. É possível, por exemplo, entender se um cidadão é de Esquerda ou Direita por meio de seu discurso e de suas publicações nas redes sociais virtuais. Por exemplo, pela expressão "lugar de fala" é possível supor que quem a usa seja mais conectado à ideologia de Esquerda do que Direita – e essa suposição se torna robusta quando vamos ao campo da Ciência Política e identificamos a relação entre essa expressão e a esquerda, (como vemos no trabalho de Morais, 2018). Logo, a Análise Quantitativa de Textos pode ser usada em diferentes campos do conhecimento.

Não queremos afirmar que análises subjetivas que emergem de técnicas qualitativas não são relevantes; pelo contrário, entendemos sua importância inclusive nos estudos positivistas. Todavia, advogamos aqui a importância de uma análise complementar por meio da utilização da Análise Quantitativa de Textos – seja para corroborar os achados de uma pesquisa, seja para gerar novos *insights*. O texto é bem mais complexo que o número (você tem que lematizar e entender o sentido da frase – é uma técnica artesanal). Todavia, a estatística deve avançar cada vez mais para a utilização de técnicas como essa aqui apresentada neste artigo metodológico. Na percepção de alguns estatísticos, a Estatística também está se reinventando por meio de técnicas como essa - que desmistificam que é uma Ciência apenas de números.

A discussão de Sales e Saião (2019) sobre a Pequena Ciência *versus* a Grande Ciência nos traz *insights* interessantes em relação às dicotomias existentes na

geração de dados. Se na Grande Ciência há uniformidade na geração de dados, investimento e infraestrutura, na Pequena Ciência (aquela que muitos de nós fazemos), os dados são heterogêneos e raramente arquivados para reuso. Assim, a Pequena Ciência que sofre com investimentos e infraesturura (vide o recente corte de bolsas de iniciação científica nas Humanidades em 2020) tem que se adaptar com as possibilidades existentes. E, nesse contexto, a Análise Quantitativa de Textos por meio de dados públicos (*online* ou *off-line*) é uma estratégia que pode ser adotada para trabalhar com um volume maior de dados (todavia, isso não que dizer que faremos a Grande Ciência).

Inicialmente consideramos como relevante jogar luz sobre a possibilidade de se fazer Ciência ao usar dos dados disponíveis em diferentes plataformas digitais. Embora nossa escolha por uma instituição de ensino como objeto de estudo seja pontual, entendemos que a Análise Quantitativa de Textos pode permitir um olhar multirrefenricial aos seus usuários a partir das teorias utilizadas. Por isso, destacamos aqui algumas futuras pesquisas que podem ser conduzidas por meio desta técnica.

Na perspectiva do consumidor, há possibilidade de analisar os comentários de usuários do *Instagram* sobre produtos divulgados por influenciadores digitais para verificar a relação dessas ações patrocinadas com um possível consumo aspiracional ou características de grupos de referência. Na Administração Pública, por exemplo, é possível buscar compreender uma infinidade de perspectivas como relatórios de gestão pública municipal, revisão de atas de conselhos municipais para entender especificidades e até a coerência entre o discurso de postulantes aos cargos do Executivo e a implementação do que foi planejado para posteriormente respondermos: será que a principal agenda foi implementada?

Considerando também que parece cada vez mais que ativistas (políticos, ambientais, sociais e morais) utilizam o *Twitter* para jogarem luz sobre um tema, pessoa ou organização, quais são as coocorrências de palavras em *hashtags* levantados no *Trending Topics* sobre um tema? Em tempos de *fake News* e de disfunções éticas da sociedade, é possível identificar robôs por meio daquilo que algumas contas publicam? Estas e outras questões podem ser respondidas direta ou tangencialmente por meio da aplicação da Análise Quantitativa de Textos. O tutorial construído para exemplificar a possibilidade desta técnica e a resposta a outras perguntas de pesquisa é apresentado ao final.

Anexo 1

TUTORIAL PARA REPLICAÇÃO DA TÉCNICA APRESENTADA

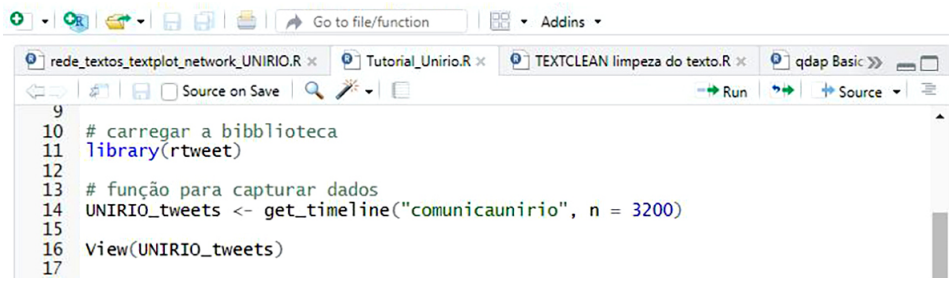
Passo 1: Antes de abrir o R , obter o token de acesso do Twitter - O passo inicial é obter o token de acesso do Twitter pelo site developer.twitter.com. Adquira uma conta de desenvolvedor, navegue para developer.twitter.com/en/apps, e clique em Create a New App e preencha o formulário.

Passo 2: Navegar no Twitter e buscar a conta que você quer baixar os dados - Nesse tutorial, vamos utilizar a conta da UNIRIO no Twitter. O endereço https://twitter.com/comunicaunirio tem a conta da UNIRIO.

Passo 3: Abrir o RStudio e instalar a biblioteca RTwitter - Para a biblioteca RTwitter funcionar, após carregar a biblioteca, você precisa da chave que você conseguiu no passo anterior. Um exemplo é apresentado a seguir:

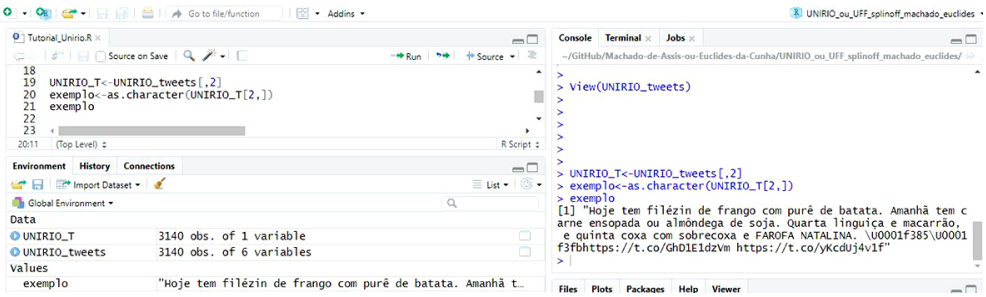
```
## load rtweet
library(rtweet)

## 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)
```

Para baixar os tweets da conta da UNIRIO, precisamos executar o código abaixo no R:

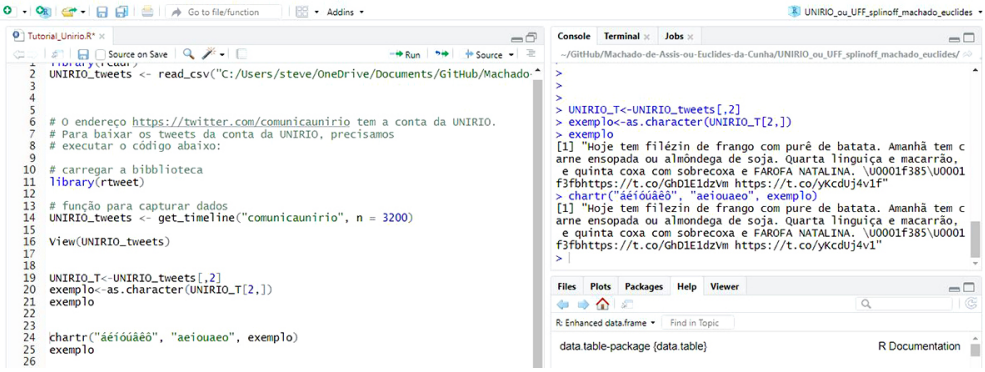


Passo 4: Visualização do Banco de Dados - O comando *View(UNIRIO_tweets)* ***é utilizado para mostrar a base de dados*** (para a gente verificar se a captura de dados funcionou). Esse comando apresentou, anteriormente, a Figura 01 no texto.

Passo 5: Filtrar somente o que a gente precisa - Como nesse momento precisamos apenas dos tweets, criamos um novo objeto somente com os tweets, como no exemplo abaixo.



Passo 6: Limpeza dos dados - Para facilitar os próximos procedimentos, vamos retirar todos os acentos (com exceção do '~')o a partir do comando *chartr*:



Para ver como essa função funciona, apresenta-se um texto de exemplo.

> exemplo

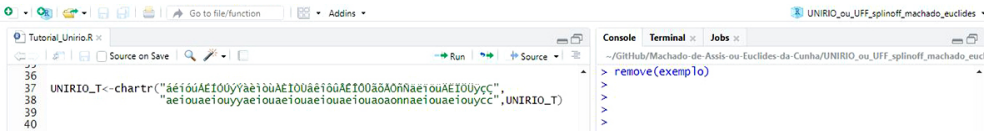
[1] "Hoje tem filézin de frango com purê de batata. Amanhã tem carne ensopada ou almôndega de soja. Quarta linguiça e macarrão, e quinta coxa com

sobrecoxa e FAROFA NATALINA. \U0001f385\U0001f3fbhttps://t.co/GhD1E-1dzVm https://t.co/yKcdUj4v1f"

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

[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/GhD1E-1dzVm https://t.co/yKcdUj4v1"

Essa função retirou o acento. Isto é, transformou "filézin" em "filezin". Agora que sabemos o que esse comando faz, vamos aplicá-lo em toda a base de dados.

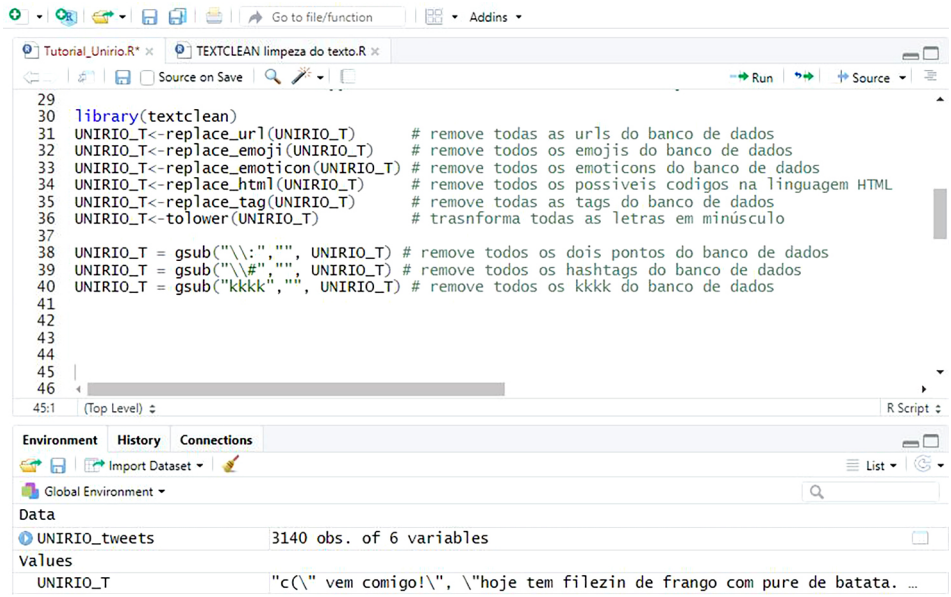


Vamos agora trabalhar na limpeza dos textos. Além dos acentos, precisamos padronizar os textos.

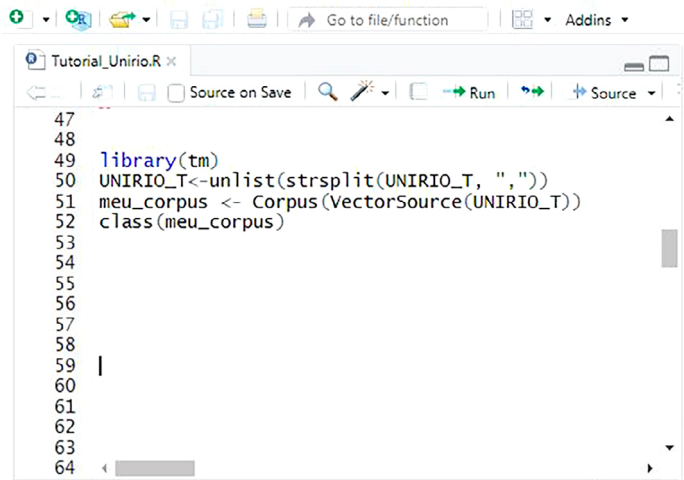
Passo 7: Padronização dos dados -Vamos usar a biblioteca *textclean*. Precisamos carregar a biblioteca com o comando "*library(textclean)".* Após essa etapa vamos retirar alguns elementos do Tweeter que não são texto.

- Para remover todas as urls do banco de dados e demais informações, para cada uma delas, utilizam-se os comandos abaixo – que serão aqui separados por ponto e vírgula que não entra na programação:

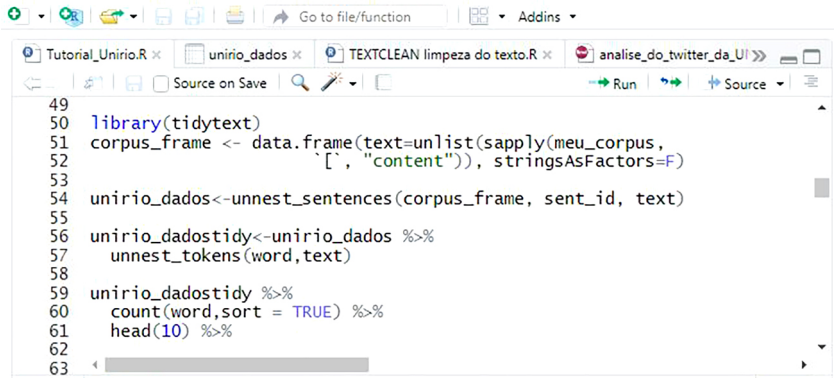
UNIRIO_T<-replace_url(UNIRIO_T); UNIRIO_T<-replace_emoji(UNIRIO_T); UNIRIO_T<-replace_emoticon(UNIRIO_T); UNIRIO_T<-replace_html(UNIRIO_T); UNIRIO_T<-replace_tag(UNIRIO_T); UNIRIO_T<-tolower(UNIRIO_T); UNIRIO_T = gsub("\\:","", UNIRIO_T); UNIRIO_T = gsub("\\#","", UNIRIO_T).



Passo 8: Transformar o banco de dados em um Corpus - Após esse processo, precisamos transformar esse objeto em um Corpus. Em Linguística de Corpus, estas bases de dados textuais são objetos de pesquisa chamadas de Corpus. Corpora é o plural de corpus – conjunto de dados linguísticos pertencentes ao uso oral ou escrito da língua e que podem ser processados por computador (IBPAD, 2020).



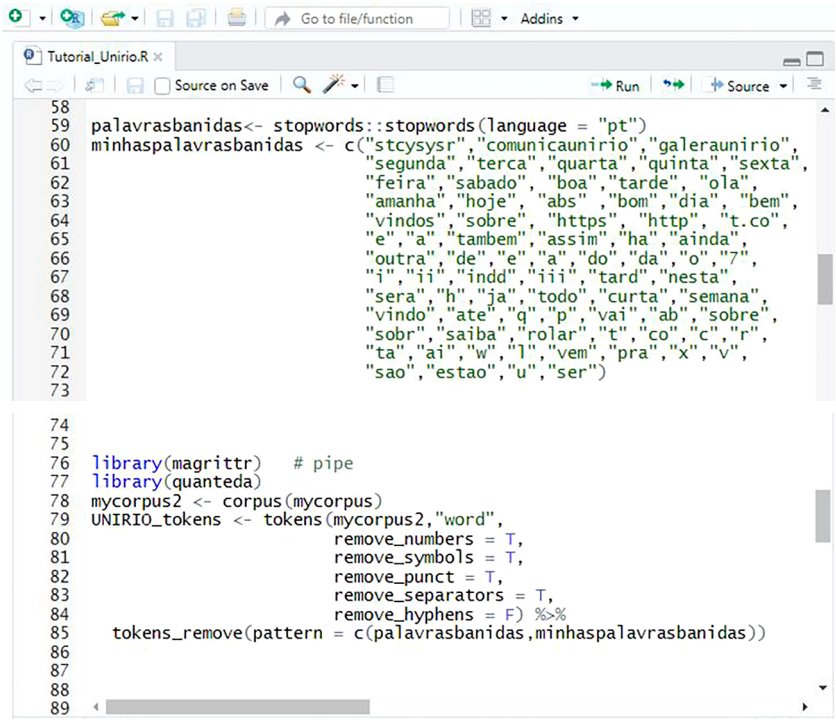
Vamos usar a biblioteca *tidytext* para dividir o banco de dados em frases (sentences), e depois em palavras (words). Um tweet é um texto com até 280 caracteres. Vamos quebra-lo em tamanhos menores. Desse modo, podemos criar banco de dados de frases e um banco de dados de palavras. As funções utilizadas para isso são *unnest_sentences* e *unnest_tokens*. O banco de dados de palavras será muito útil para a criação da tabela de palavras mais frequentes. Com a função *count()* podemos contar os objetos. O comando sort=TRUE diz que queremos ordernar essa contagem. Finalmente, o comando head() mostra as dez palavras mais frequentes – que foi apresentada previamente na Tabela 1.



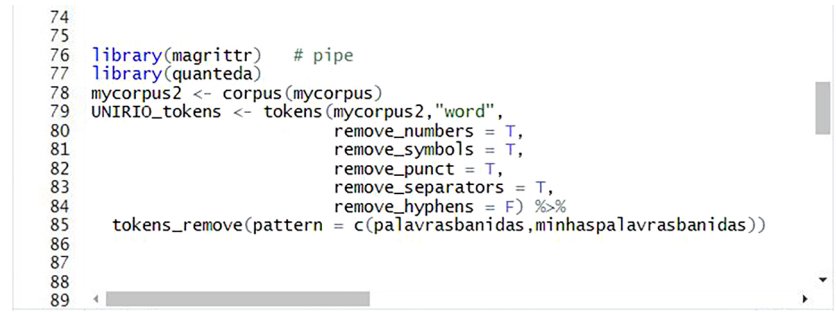
Passo 9: Remover as stopwords - O próximo passo é remover as stopwords. Uma das principais formas de pré-processamento de corpus é filtrar dados inúteis. No processamento de linguagem natural, palavras inúteis (dados) são chamadas de stopwords. Stopwords é uma palavra comumente usada (como “o”,”a”, “um”, “uma”) que um mecanismo de pesquisa foi programado para ignorar, tanto ao indexar entradas para busca quanto ao recuperá-las. como resultado de uma consulta de pesquisa. Nesse tutorial, temos dois grupos de palavras “stopwords”. O primeiro são aquelas palavras que já foram definidas pela comunidade científica. O segundo grupos foram as palavras que definimos como aquelas que agregam pouco valor (exemplos: “bom dia”,”boa tarde”, “quarta”,”quinta”,”sexta”).

Para remover as *stopwords* vamos utilizar a função tokens_remove() do pacote Quanteda. Um exemplo de como remover essas palavras é apresentado a seguir.

Neste exemplo, vamos remover as palavras que já são definidas como stopwords e aquelas que definimos.

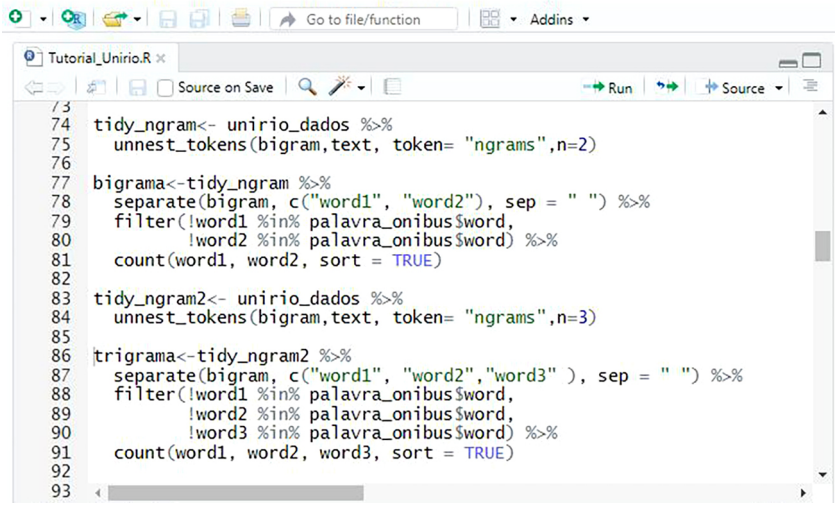


Passo 10: Criação da nuvem de palavras - Como já temos um banco de dados com as palavras, já podemos criar uma nuvem de palavras (Figura 3 do texto). Para a criação da nuvem, precisamos construir o seguinte código:



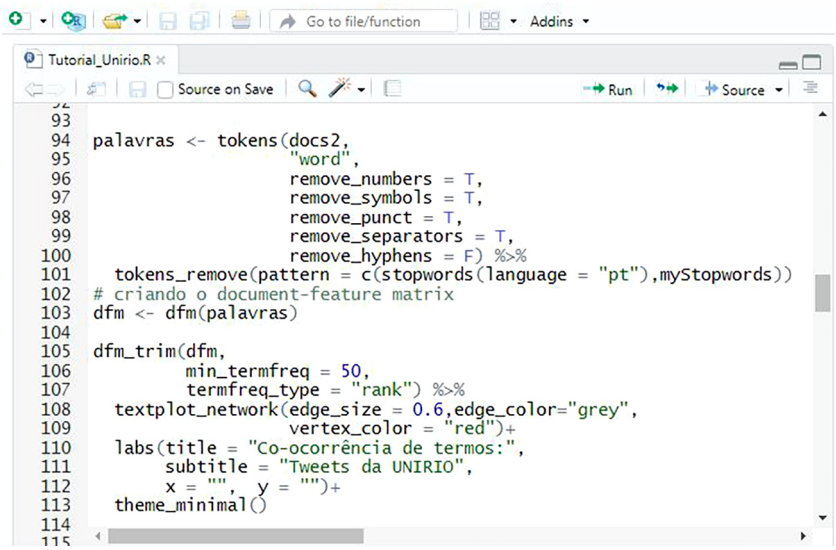
Note que: **Wordcloud** é a função para gerar a nuvem de palavras; **Words** é onde você precisa definir o objeto com as palavras; **Freq** é o parâmetro para definir o tamanho das palavras (aqui será a frequência); **Max.words = 100** diz que queremos as 100 palavras mais frequentes.

Passo 11: Criação dos bigramas e trigramas - Para gerar os bigramas e os trigramas (Tabelas 02 e 03), precisamos utilizar a função unnest_tokens(). Essa função "quebra" as frases em sequências consecutivas de palavras, chamadas n-gramas. Se você definir n=2, você escolhe o bigrama, se definir n=3, será um trigrama. O código para gerar os n-gramas é apresentado a seguir:

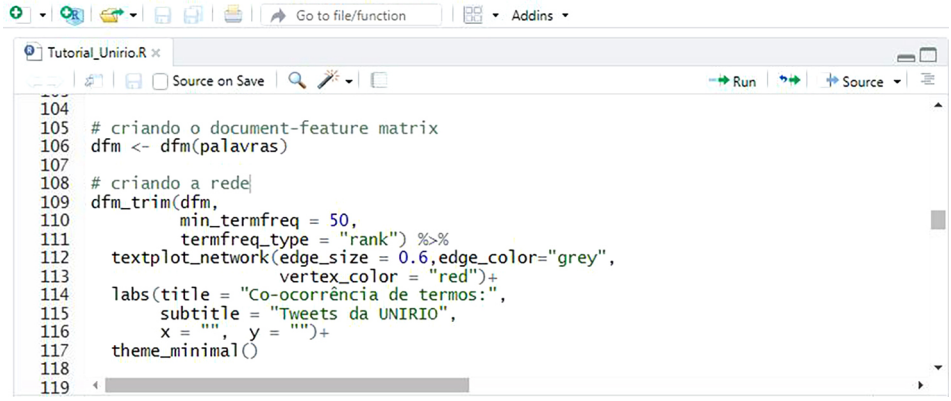



```
tidy_ngram<- unirio_dados %>%
  unnest_tokens(bigram,text, token= "ngrams",n=2)

bigrama<-tidy_ngram %>%
  separate(bigram, c("word1", "word2"), sep = " ") %>%
  filter(!word1 %in% palavra_onibus$word,
         !word2 %in% palavra_onibus$word) %>%
  count(word1, word2, sort = TRUE)

tidy_ngram2<- unirio_dados %>%
  unnest_tokens(bigram,text, token= "ngrams",n=3)

trigrama<-tidy_ngram2 %>%
  separate(bigram, c("word1", "word2","word3" ), sep = " ") %>%
  filter(!word1 %in% palavra_onibus$word,
         !word2 %in% palavra_onibus$word,
         !word3 %in% palavra_onibus$word) %>%
  count(word1, word2, word3, sort = TRUE)
```

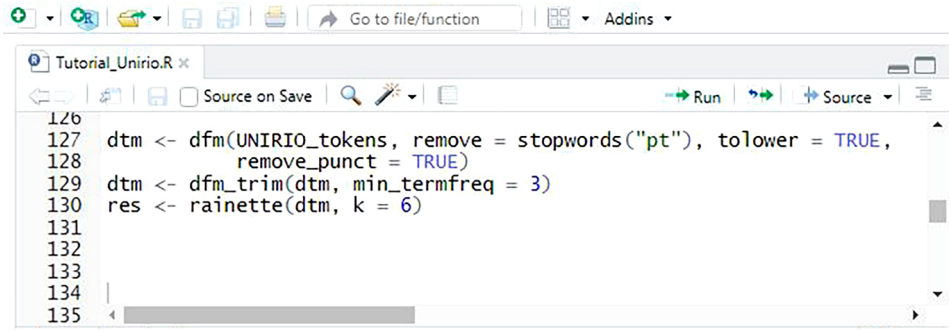
Passo 12: Criação do DFM (document-feature matrix) - Antes de criar a rede de coocorrências e o *cluster*, precisamos criar o *document-feature matrix* - DFM. Para executar análises estatísticas, precisamos extrair uma matriz que associa valores para determinados recursos a cada documento. Usamos a função 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:



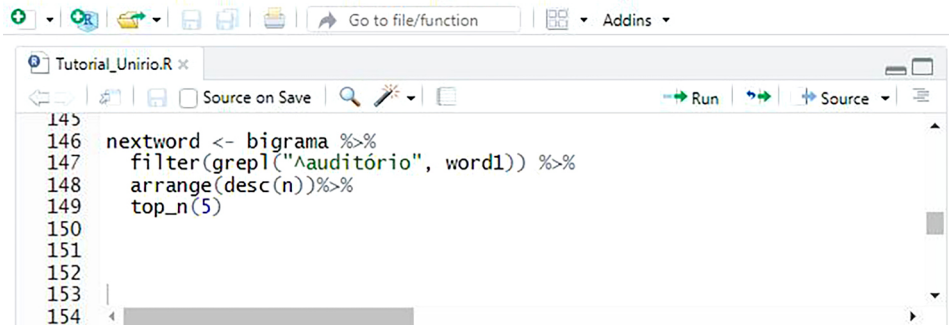
Passo 13: Criação da rede de co-ocorrencia de palavras - Para a criação da rede de co-ocorrencia (Gráfico 1) utilizamos a função textplot_network() da biblioteca Quanteda. Aqui podemos ver as possibilidades de mudar o título, cores e linhas.



Passo 14: Criação dos clusters - Precisamos aqui de um DFM e da função rainette da biblioteca Rainette. No exemplo abaixo, estamos criando clusteres no dfm e k=6 indica que queremos 6 clusters. O Gráfico 2 apresentou os resultados desta codificação.



Passo 15: Criação da Próxima Palavra - A previsão da próxima palavra é baseada no Bigrama e teve seus resultados apresentados nas tabelas 04 e 05. Essa função não está em nenhum pacote. Será necessário criá-la. Para criar a Próximas palavras, precisamos do seguinte código:



Note que **filter** é a função para selecionar as palavras que começam com auditório; **grepl** é função para encontrar o *matching, isto é, as coincidências;* **Top_n(5)** é para mostrar as cinco palavras com maior associação com auditório; **^** é uma expressão regular. Uma ancora, isto é, as palavras que começam por auditório; **Essa função** selecionar as palavras que começam com auditório, e mostram a próxima palavra do bigrama.

Referências

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-GII, 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.