Caminhando para uma Ciência Aberta: Uma abordagem estatística a partir da criação de resumos extrativos manuais e automáticos

Moving towards an Open Science: A statistical approach based on the generation of manual and automatic extractive summaries

Iván Arias, Universidade de Santiago de Compostela, Espanha, https://orcid.org/0000-0003-2673-0899

Margarida Castro, Universidade do Minho, Portugal

Resumo: Nos dias de hoje, a difusão de novos meios facilitou a proliferação de dados científicos, que podem ser divulgados graças a novas técnicas de tratamento da informação. Visa-se, neste artigo, a partir do fluxo de trabalho que se estabelece entre dados abertos e ciência de dados, analisar resumos gerados de forma manual e de forma automática em termos estatísticos. Assim, avaliam-se novas possibilidades de tornar o conhecimento científico mais acessível ao caminharmos para uma democracia de dados. Desta forma, tomando um corpus de textos resumidos como ponto de partida, realizar-se-ão análises quantitativas com recurso a fundamentos teóricos que permitirão retirar conclusões relativamente à viabilidade da automatização para atingirmos uma ciência aberta.

Palavras-chave: automatização; ciência aberta; ciência de dados; corpus; democratização de dados; resumos extrativos.

Abstract: Nowadays, the diffusion of new media has facilitated the proliferation of scientific data, which can be disseminated thanks to new information processing techniques. This article aims, based on the workflow established between open data and data science, to analyse manually and automatically generated summaries in statistical terms. As a result, we evaluate new possibilities of making scientific knowledge more accessible as we move towards a data democratization. Taking a corpus of abstracted texts as a starting point, quantitative analysis will thus be carried out using theoretical foundations that will allow us to draw conclusions about the feasibility of automation to achieve an open science.

Keywords: automation; open science; data science; corpus; data democratization; extractive summaries.

1. Introdução

Na atualidade, o surgimento constante de novos recursos e a otimização daqueles meios de que já dispúnhamos tem feito com que existam cada vez mais intersecções e conexões entre diversas áreas de estudo. É daí que surge a inter-relação entre as humanidades e a informática, que dá lugar à aparição das Humanidades Digitais, convertendo-se numa disciplina dentro da qual se pretende analisar a nossa herança linguística e cultural através da aplicação de técnicas informáticas (cf. Svensson, 2012, p. 6).

A presente pesquisa inscreve-se, portanto, nesta intersecção, estabelecendo-se como objetivo principal alcançar uma análise estatística que nos permita ver e avaliar as semelhanças e disparidades que existem entre textos gerados de forma automática e manual e que foram inicialmente redigidos por divulgadores de ciência. Assim, ter-se-á sempre em vista a necessidade de tornar os dados científicos acessíveis para o público geral e para leigos na matéria, dado que se pretende alcançar um estágio de democratização de dados.

De um ponto de vista mais prático, tomar-se-á como ponto de partida para o presente estudo um corpus desenhado *ad hoc*. Note-se, desde já, que um corpus é um conjunto de textos digitalizados que são reunidos e recolhidos considerando um propósito específico para depois operar com eles (cf. Atkins et al., 1992, p. 1). Para a construção do corpus de trabalho, foi essencial o recurso à plataforma digital *ScienceX*[1], pois, a partir deste repositório, elaborou-se o corpus com diferentes textos de teor científico escritos em inglês.

Atkins et al. (1992) introduzem, adicionalmente, a noção de subcorpus para aludir a um subconjunto formado por textos de um corpus e que servem, principalmente, para analisar estatisticamente determinadas questões derivadas de uma seleção dinâmica a partir dos dados que constituem o corpus original. Assim, para o intuito desta pesquisa, escolheram-se três subconjuntos, isto é, três subcorpora, que serão o fundamento da análise estatística seguida. Destacam-se três categorias:

1. Resumo vulgarizado (doravante, RV): corresponde ao artigo de divulgação publicado no repositório. Não se trata de um artigo científico *per se*, mas sim de um texto redigido com o propósito de ser divulgado para atrair leitores. Geralmente, vem acompanhado de imagens e o texto costuma ser significativamente simples quando comparado com o próprio artigo científico ao qual se refere.
2. Resumo extrativo vulgarizado (doravante, ResExtV): diz respeito ao resumo extrativo realizado de forma manual ao conjugar dez frases que melhor representam o conteúdo do artigo científico de partida.
3. Resumo extrativo automático (doravante, RExtAut): o resumo extrativo automático é criado de forma automática através do recurso à linguagem de programação Python, sendo de igual modo composto por dez frases, para evitar ulteriores adaptações nalgumas fases da pesquisa.

Uma vez delimitado o corpus de trabalho, deve-se considerar que termos como dados abertos, ciência aberta ou ciência de dados são fulcrais para entendermos a análise que se persegue com o presente estudo. Debruçar-nos-emos sobre estas noções, assim como sobre a relação que existe entre elas, na secção 2 do presente artigo. Em suma, presentemente, parece conveniente listarmos os objetivos perseguidos com esta investigação:

1. Tornar o conhecimento científico de que dispomos mais acessível e contribuir para a consolidação da democratização da informação (Mirowski, 2018, p. 176). Igualmente, dar-se-á um contributo para o estudo de técnicas de simplificação de texto científico.
2. Adotar uma metodologia para a compilação e, consequentemente, para a análise de textos que constituem um subcorpus de trabalho determinado por fins específicos.
3. Explorar ferramentas informáticas que permitam analisar os corpora de forma quantitativa e qualitativa e que possibilitem, em primeiro lugar, elaborar resumos de forma automática.
4. Efetuar uma análise estatística a partir de uma série de perguntas de investigação que possibilite uma extração posterior de conclusões objetivas.

Tendo estes objetivos como base (especialmente, o objetivo número 4), dedicar-nos-emos a apresentar os passos que foram seguidos no âmbito desta pesquisa e que levaram à retirada de conclusões que possam eventualmente contribuir para o estabelecimento de conhecimento científico entre a população global. Aliás, explicar-se-ão técnicas de tratamento de dados que são essenciais para as Humanidades Digitais, pela sua capacidade para a avaliação de informação.

2. Conceitos básicos para entendermos a Ciência Aberta

A possibilidade de acesso à Internet a nível mundial provocou uma enorme proliferação de dados, que hoje já são amiúde difíceis de manipular devido à quantidade de informação com a qual contamos. Por sua vez, a digitalização e a digitação de conteúdo que existia apenas em suportes analógicos explica, ao menos parcialmente, a sobrecarga de informação atual (cf. Batarseh & Yang, 2020). É neste contexto de propagação de ciência que se insere o presente estudo, já que se pretendem aplicar diferentes métodos computacionais e estatísticos para tratamento de dados.

Primeiramente, deve-se esclarecer, de uma perspetiva teórica, o que são os dados abertos e os macrodados. A passagem da Web 1.0. para a Web 2.0., assim como os avanços na ciência computacional, levaram a um aumento notável das possibilidades de expansão dos sistemas de informação (cf. van der Aalst, 2016, p. p. 3). Este crescimento exponencial é o que comumente se designa como macrodados.

O estabelecimento da Web 2.0 revelou-se fulcral para a consolidação dos dados abertos e dos macrodados, pois permitiu o estabelecimento de ligações entre pessoas graças às redes sociais, para além das ligações que existiam entre os ficheiros disponibilizados online (cf. Newman et al., 2016, pp. 591-592). No entanto, como apontam estes autores, é o atual desenvolvimento da Web 3.0 que possibilita, na realidade, o armazenamento de macrodados, através de sistemas computacionais que permitem guardar enormes quantidades de ficheiros na nuvem.

Na atual era digital, em que já contamos com muitas publicações em apenas versão digital, os dados abertos tornam-se fundamentais: “By open data in science we mean data that are freely available on the public internet permitting any user to download, copy, analyze, re-process, or use these for any other purpose [...]” (Masuzzo, 2017, p, 3). Assim, neste caso, iremos colocar o nosso foco nos dados abertos do âmbito científico, embora não possamos negar que também existe uma importante demanda por dados abertos provenientes de governos, por exemplo. Consequentemente, afirmamos que os macrodados podem ser, da mesma forma, dados abertos *per se*, mas cujo tamanho impede uma análise com meios tradicionais, tornando-se necessário o recurso a novas tecnologias (cf. Provost & Fawcett, 2013, p. 54).

Aliás, devemos esclarecer que é através do acesso aberto que é possível garantir a acessibilidade à maior parte de publicações académicas, pois, tal como explica Masuzzo (2017, p. 2), para se atingir o objetivo da ciência aberta, temos que primeiramente consolidar vários pilares fundamentais: entre eles, destacam-se os dados abertos *supra* mencionados, a criação e implementação de software de código aberto, a consecução de maior participação na ciência por parte do público em geral e a consolidação da revisão por pares em formato aberto. De facto, já é possível salientar diferentes organizações e associações que trabalham em prol de alcançarem uma ciência aberta, pois, de acordo com Mirowski (2018, p. 172), a Comissão Europeia já tinha asseverado que a maioria dos artigos deviam ser publicados em acesso aberto para toda a comunidade, o que se tem verificadox nos últimos anos com a aparição de novos repositórios online (vejam-se os repositórios Elsevier ou Google Scholar, entre outros).

A implementação de software e ferramentas computacionais de código aberto *supra* referidas faz parte da metodologia característica da ciência de dados, pois é frequente a utilização destas ferramentas por cientistas de dados ao longo da sua pesquisa. Davenport e Patil (2012) assinalam que os cientistas de dados fazem descobertas significativas enquanto se encontram mergulhados nos dados que tratam. Isto deve-se ao facto de terem atualmente enormes quantidades de dados que conseguem manipular com ferramentas cada vez mais potentes, o que provoca não só análises *ad hoc*, mas também uma conversa ilimitada com a informação que nos permite retirar novas conclusões de forma constante. Todavia, antes de nos ocupar com alguns pormenores da ciência de dados, cumpre chamarmos a atenção para uma definição exaustiva da mesma, segundo van der Aalst (2016, p. 10):

Data science is an interdisciplinary field aiming to turn data into real value. Data may be structured or unstructured, big or small, static or streaming. Value may be provided in the form of predictions, automated decisions, models learned from data, or any type of data visualization delivering insights. Data science includes data extraction, data preparation, data exploration, data transformation, storage and retrieval, computing infrastructures, various types of mining and learning, presentations of explanations and predictions, and the exploitation of results taking into account ethical, social, legal, and business aspects.

A partir desta definição, pode-se deduzir que a área de trabalho da ciência de dados compreende mais tarefas do que a área de trabalho normalmente delimitada pela estatística ou pela prospeção de dados, embora os métodos destas duas disciplinas também sejam comumente empregados no âmbito da ciência de dados. Antes de mais, pode-se definir o conceito de prospreção de dados, de acordo com a Comissão Europeia(2016, p. 51) como sendo "a set of methods that use to automatically search, filter, and interpret large amounts of digital and online content". No que diz respeito à preparação e extração dos nossos dados para ulterior análise, e não fugindo do corpus em que se baseia este estudo, ocupar-nos-emos em seguida com algumas das técnicas computacionais que contribuíram para a constituição do nosso conjunto de dados.

Primeiramente, e numa fase preparatória de dados, tornou-se essencial o recurso à plataforma *ScienceX* para a extração de diferentes textos que acabaram por compor o nosso corpus. Embora não fossem aplicadas técnicas de *web scrapingstrictosensu*, devemos esclarecer que este conceito é definido como recurso essencial para a extração de dados da web (Zhao, 2017). No nosso caso, a extração foi realizada manualmente, pois a plataforma permite transferir ficheiros de forma fácil e intuitiva.

Quanto à criação do conjunto de dados estatísticos *per se*, e considerando os três subcorpora elencados na secção *supra*, recorremos a algumas bibliotecas de Python para alcançar os nossos objetivos (cf. Layton, 2015). Determinadas bibliotecas de Python, entre as que devemos destacar NLTK ou Textstat, permitem, para além do pré-processamento, realizar medidas estatísticas e preditivas em textos, relativamente, por exemplo, à complexidade ou à legibilidade do mesmo. Esclarece-se, consequentemente, que através de aplicações computacionais, se elaborou um conjunto de dados que será apresentado de forma mais pormenorizada na secção 4 do presente artigo.

Em suma, demonstra-se que todas as ferramentas computacionais de exploração e análise de dados operam com a informação existente (macrodados, dados abertos) para fornecer à comunidade científica e ao público geral a possibilidade de recuperarem ficheiros de teor científico em aberto (que fazem parte, portanto, da ciência aberta). Tem-se vindo a ratificar que a ciência aberta é o objetivo que devemos considerar na aplicação de métodos de ciência de dados nos quais tomamos dados abertos como ponto de partida. É frequente a divulgação de ciência aberta através de websites cuja abordagem visa tornar o conhecimento científico mais acessível para todas as pessoas. Consoante o exposto por Mirowski (2018, p. 193), estas iniciativas devem adaptar o conhecimento científico para a população global, provocando um comportamento mais reativo e recetivo por parte dos leitores, independentemente de serem especialistas ou leigos na matéria.

3. O que são os resumos extrativos?

Sendo que dois dos três subcorpora com os quais trabalharemos estão constituídos por resumos extrativos, parece relevante debruçarmo-nos sobre a seleção e constituição deste tipo de resumos. Antes de aprofundar o assunto, parece essencial definirmos o que se entende por "resumir". De acordo com o dicionário da Infopédia, resumir é "dizer em poucas palavras o que se disse ou escreveu mais extensivamente; sintetizar; abreviar; condensar". A partir desta definição podemos desde já deduzir qual é o objetivo dos dois subcorpora, sendo que se tenciona elaborar resumos que permitam comunicar o conteúdo de artigos científicos de forma sucinta.

Todavia, tomando em consideração a delimitação dos subcorpora, parece fulcral esclarecermos ainda o que entendemos por resumos extrativos. Conforme o defendido por Lloret (2021, p. 88):

Si se sigue una estrategia extractiva, se seleccionarán y extraerán, literalmente, las frases más importantes del documento sin realizar ninguna modificación sobre ellas, lo que sería equivalente al proceso de subrayado de la información más relevante del documento […] de origen.

Vê-se, desta forma, como os resumos extrativos apenas pretendem extrair frases do texto original para relatar e reproduzir os aspetos do conteúdo que podem ser considerados mais relevantes. Aliás, todos os resumos selecionados apresentam outra característica importante, como assinalado por Lloret (2021, p. 88), nomeadamente o facto de todos terem sido formados a partir de um só artigo científico original. Trata-se, consequentemente, de resumos extrativos mono-textuais.

No que diz respeito aos resumos extrativos manuais ou vulgarizados, devemos salientar que foram criados após uma leitura detalhada dos textos, o que possibilitou a extração das frases

que se destacam como sendo as mais significativas e também, em certo ponto, devido à necessidade de alcançarmos uma transmissão bem-sucedida da mensagem. Para clarificar a elaboração comum destes resumos manuais, devemos recorrer à linguística textual tradicional. De acordo com van Dijk (1979), do ponto de vista da cognição e da semântica intrínseca à textualidade, para resumir qualquer artigo devemos seguir diferentes passos: primeiro, tentamos perceber as microestruturas e proposições internas do texto, para depois podermos identificar uma macroestrutura em termos categóricos (secções, por exemplo) e, por último, acabamos por ser capazes de exprimir essa macroestrutura com coerência de modo a transmitirmos as mensagens fundamentais do texto selecionado.

A componente relativa à estruturação do novo resumo é facilitada, neste caso, pelo facto de lidarmos com artigos científicos, em que amiúde pode ser observada uma divisão em secções muito clara: introdução, objetivos, enquadramento teórico, metodologia, análise de dados, conclusões. Isto facilita o processo de resumo manual, pois, embora não sejamos conscientes, acabamos sempre por recorrer àquelas partes, respetivas a cada uma das secções de que dispõe o artigo, que melhor sintetizam a informação que pretendemos continuar a transmitir com o resumo.

Por sua parte, o resumo extrativo automático baseia-se em algoritmos de frequência que calculam as palavras mais repetidas no texto, para depois outorgar uma nota às diferentes frases que contêm os lexemas selecionados. Esta técnica não foge à tradicional heurística própria da geração automática de resumos, pois a importância das frases é dada pela presença ou ausência de determinadas palavras (cf. Lloret, 2021, p. 90). Desta forma, foram utilizados módulos para pré-processar os textos e conseguir uma maior exaustividade na geração de resumos. Depois, empregou-se o módulo de tokenização na biblioteca NLTK (cf. Bird et al., 2009) para atingirmos uma separação objetiva de todos os lexemas. Considerando a frequência das palavras, todas as frases do texto recebem um valor que nos permite obter um número fixo de frases para o resumo. Neste caso, quer para o resumo manual, quer para o resumo automático, o número de frases fixou-se em 10.

Não obstante, devemos considerar que os resumos extrativos manuais são essenciais, uma vez que são habitualmente utilizados para treinar máquinas para que os computadores possam extrair automaticamente frases e palavras (cf. Cheng & Lapata, 2016). Destarte, estabelece-se um fluxo de trabalho entre estes dois tipos de resumos extrativos com os quais trabalhamos, já que os resumos manuais podem ser empregues como base ou ponto de partida para pôr em funcionamento diferentes interfaces de programação. Ao estabelecer estas duas categorias, somamos 33 textos respetivamente, constituindo-se um subcorpus maior de resumos extrativos em geral formado por 66 textos.

Adicionalmente, é possível comparar a média de frases que ocorrem de forma coincidente entre o resumo extrativo manual e automático. Acabou-se por determinar que apenas uma média de 1,42 frases aparecem simultaneamente nas duas classes. As proposições repetidas estão normalmente relacionadas com mensagens fulcrais para a compreensão geral do tema do artigo. A geração automática mostra, em termos gerais, uma preferência por orações longas e provavelmente com uma maior densidade lexical. A extração manual, por sua parte, presta atenção a aspetos como a coesão textual.

A partir da elaboração destes dois subcorpora, afirma-se que, no âmbito deste projeto, dispomos de três categorias para a análise estatística, o que soma um valor total de 99 textos, a partir dos quais se realizarão cômputos estatísticos e preditivos: (i) 33 resumos extrativos

manuais ou vulgarizados, (ii) 33 resumos extrativos automáticos e (iii) 33 resumos abstrativos[2] feitos manualmente por divulgadores de ciência.

4. Metodologia

4.1. Conjunto de dados[3]

Antes de estabelecermos as perguntas de investigação, debruçar-nos-emos sobre a descrição das diferentes variáveis estatísticas que formam o conjunto de dados de trabalho, conseguidas com metodologia de ciência de dados. Neste caso, graças tanto à interface de programação Python como à biblioteca Textstat, conseguimos extrair dados de carácter quantitativo para os três subcorpora de análise. Destarte, podemos, após uma observação e comparação pormenorizada de dados, concluir qual é a categoria textual que leva a uma leitura mais demorada, qual contém mais termos específicos ou qual tem maior nível de subjetividade. Aliás, o facto de contarmos com dados numéricos facilita não só a análise estatística, mas também a retirada de conclusões objetivas, uma vez que a comparação é feita em termos de análise matemática. As variáveis, consideradas como sendo características comuns a todos os elementos de uma população (neste caso, dos três subcorpora) e que podem ser atribuídas um valor categórico ou numérico, serão apresentadas a seguir. Para além disso, devemos ter em consideração que a população com a qual trabalhamos acaba é descrita através de diferentes valores associados a cada uma das variáveis de que dispomos. É com os valores numéricos atribuídos a cada elemento com respeito às variáveis que realizaremos a análise estatística posterior. As variáveis para a análise estatística são as seguintes:

1. **sent**: esta variável diz respeito ao número de frases de cada texto. Trata-se de uma variável quantitativa discreta. Não obstante, para dois dos subcorpora aqui analisados, deve-se recordar que o número de frases foi previamente fixado em 10.
2. **av_sent_length**: alude à média de palavras por cada frase, isto é, ao tamanho das frases. É uma variável quantitativa contínua.
3. **lex_count**: refere-se ao número total de palavras por texto (*tokens*). É uma variável quantitativa discreta.
4. **unique_word**: corresponde ao número de palavras únicas, isto é, os *types*que aparecem no texto. Do ponto de vista linguístico, a máquina conta aqui os diferentes lexemas ou palavras lexicais. Trata-se de uma variável quantitativa discreta.
5. **ttr**: o TTR refere-se à densidade lexical do texto analisado. É o rácio entre o número total de palavras únicas (*types*) e a contagem total de palavras de um texto (*tokens*). De acordo com Mitchell (2015), o TTR acostuma ser apresentado como rácio ou percentagem, embora possa também aparecer como sendo uma variável quantitativa contínua.
6. **av_letter_word**: corresponde à média de letras (grafemas) de cada palavra. É uma variável quantitativa contínua.
7. **difficult_words**: representa o número de palavras difíceis que aparecem no texto. Trata-se de uma variável quantitativa discreta.
8. **reading_time_secs**: obtém-se o tempo de leitura de um texto em segundos como valor. É uma variável quantitativa contínua.
9. **read_ease**: oferece informação relativa à facilidade de leitura do texto que se introduz como *input*. É uma variável quantitativa contínua em escala intervalar.
10. **polarity**: proporciona informação sobre a análise de sentimentos. É uma variável quantitativa contínua em escala intervalar.

11. **subjectivity**:fornece informação acerca do sentimento existente no texto. Trata-se de uma variável quantitativa contínua em escala intervalar.

4.2. Perguntas de investigação

A delimitação das variáveis, assim como a seleção de três subcorpora, permitir-nos-ão estabelecer algumas perguntas de investigação que aqui se enumeram:

1. Como anteriormente referido, o número de frases para a geração dos resumos quer manuais quer automáticos foi previamente fixado em 10. Não é então relevante medir o seu tamanho. Porém, de modo a considerarmos esta informação posteriormente, é fulcral respondermos à seguinte pergunta: qual é a média de frases que aparecem nos resumos vulgarizados (RV) criados por divulgadores de ciência? Aliás, em termos de tamanho dos textos, podemos analisar e comparar o cumprimento médio das frases presentes nas três categorias com as quais trabalhamos.
2. Qual é a média de palavras totais (tokens) para cada tipo de texto? Qual é a variabilidade das palavras únicas ou lexicais (types) nas três categorias? Qual é o TTR médio dos diferentes tipos de resumo? Em termos de terminologia e linguagem de especialização, em que tipo de texto é que aparece um maior número de termos (palavras "difíceis")?
3. Parece pertinente observar também a correlação existente entre estas "palavras difíceis" e a média de letras por palavra nas diferentes categorias.
4. Para retirarmos conclusões relativas à facilidade de leitura dos diferentes resumos com os quais trabalhamos, pode ser interessante responder à seguinte pergunta: se se criarem categorias para a variável **read_ease**, qual é a distribuição de frequência de facilidade de leitura em cada tipo de texto?
5. De acordo com Hartley (2016), a fórmula Flesch para o cálculo da facilidade de leitura devia ser considerada obsoleta por se basear exclusivamente no tamanho das frases. Para comprovar esta hipótese, tenciona-se pesquisar a correlação entre as duas variáveis que seguem: **av_sent_length** e **read_ease**.
6. Seguindo a pergunta anterior, pretende-se averiguar igualmente a correlação existente entre a facilidade de leitura de textos e o tempo que, de acordo com Python, o leitor demora a lê-los. Há de ser pertinente, para tal objetivo, cruzar essas duas variáveis.
7. Tendo-se revelado que os textos do âmbito científico, seja qual for a área de especialização, contêm um número elevado de termos, podemos presumir que eles mostram tendência para uma polaridade neutra e para um carácter objetivo. No entanto, como será a correlação entre a polaridade e a subjetividade nas categorias selecionadas?

A demarcação destas perguntas de investigação conduzirá a uma análise estatística mais pormenorizada, para a qual devemos dispor de um raciocínio baseado em fundamentos matemáticos para, no final, retirarmos conclusões objetivas que nos permitirão encerrar a pesquisa apresentada.

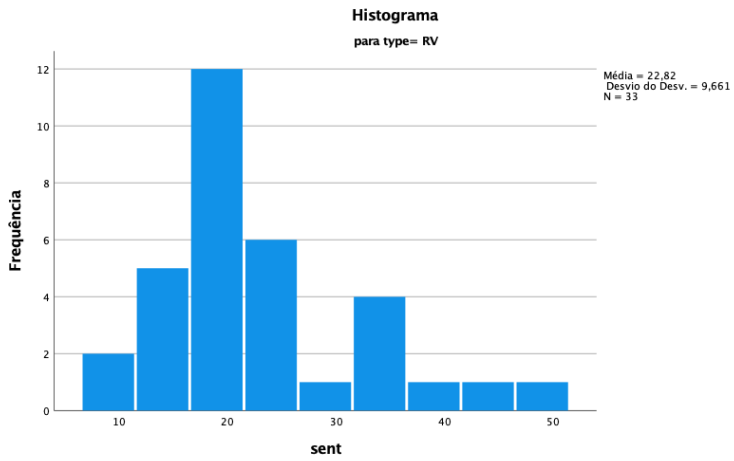
5. Análise de resultados

5.1. Primeira pergunta de investigação

No que diz respeito a esta pergunta de investigação, tenciona-se, primeiramente, analisar o número de frases do RV, pois já sabemos, à partida, que o número de frases para as categorias RExtAut e ResExtV foi definido em 10. O gráfico de barras da figura 1 representa, com base na nossa amostra, o número de frases dos RV, sendo que a média (μ) deles é de 22,82. Deste

modo, no que concerne estatísticas descritivas, podemos afirmar que a mediana, isto é, o valor do meio do conjunto de dados analisado, é 21 e a moda, por sua parte, se aproxima do valor 20. Em relação à nossa amostra, podemos, aliás, determinar que o valor mínimo é de 9 e o máximo é de 51, o que faz com que exista uma amplitude ou uma dispersão de 42 entre o resumo vulgarizado com menos frases e aquele com o maior número. Baseando-se nestes dados, é fornecido um gráfico de barras (Figura 1) no qual podemos apreciar uma clara distribuição assimétrica positiva, existindo uma maior concentração no número de frases mais baixo. Por outras palavras, a maioria dos resumos vulgarizados são mais curtos (cerca de 20 linhas). Contamos, portanto, com um enviesamento à direita do gráfico, aparecendo uma cauda alongada no outro lado, embora a classe 35 tenha um maior número que as outras mais próximas.

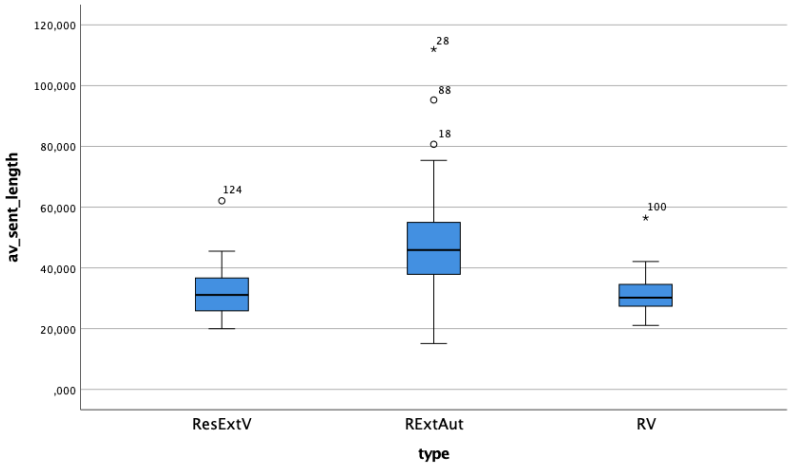
Figura 1 – Gráfico de barras para a variável sent na categoria RV.



Como já foi referido previamente, o tamanho das categorias textuais com as quais trabalhamos não podia ser analisado a partir do seu número de frases, pois só no caso do RV é que este valor podia variar, tal e como se pode observar no gráfico de barras da Figura 1. Por consequência, decidiu-se comparar os valores a respeito da média do tamanho das frases, o que se revelou essencial para respondermos à primeira questão de investigação. Tal como se pode apreciar no *boxplot* da Figura 2, que utilizamos para comparar duas variáveis quantitativas em base à sua distribuição, o RExtAut é a categoria com a mediana mais elevada (45,9), para além de apresentar uma maior variabilidade (visível na amplitude interquartil). Os resumos vulgarizados ou realizados por pessoas, pelo contrário, contam com uma mediana inferior e mais semelhante entre si (31,1 no caso do ResExtV e 30,2 no RV). No que diz respeito à sua μ, ela difere bastante entre os resumos vulgarizados (32,8 na categoria ResExtV e 31,5 no RV) e o automático (42,7).

Aliás, a distância ou amplitude que existe entre o máximo (112) e o mínimo (15,1) no resumo automático evidencia que os critérios ou algoritmos nos quais o Python se baseia para a geração automática são heterogéneos. Igualmente, também é esta categoria que conta com três *outliers*, um deles (assinalado com o número 28) severo. Neste sentido, podemos afirmar que os *outliers* para as três categorias não se referem em nenhum caso ao mesmo número de artigo. Se essa correspondência acontecesse, poderia ser devido à presença, por exemplo, de frases muito longas num artigo determinado. Não obstante, não podemos retirar conclusões relativamente a esse comportamento, uma vez que os *outliers*correspondem a diferentes textos originais.

Figura 2 – Boxplot para a variável av_sent_lenght nos três subcorpora.

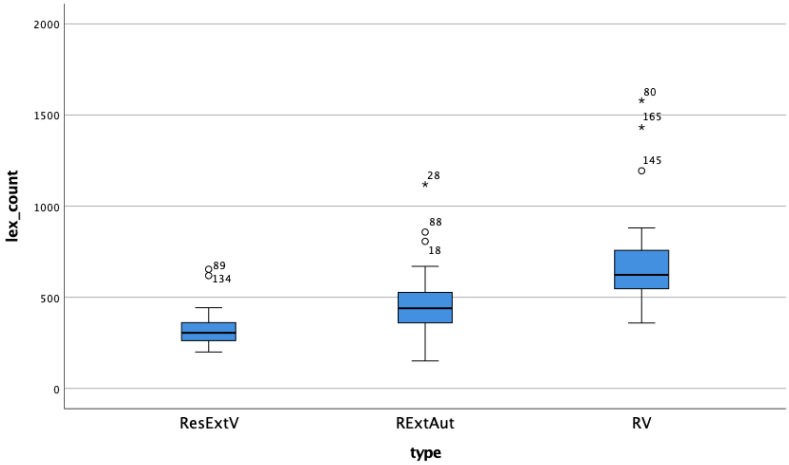


Pode-se, porém, encerrar esta primeira questão de investigação ao verificar-se que os resumos gerados de forma automática apresentam uma maior disparidade relativamente à média do tamanho de frases. Por outras palavras, o Python, com base nos dados estatísticos, parece preferir orações maiores por concentrarem talvez um maior número de termos ou "palavras difíceis", o que lhes acaba por outorgar uma melhor nota na interface de programação a essas frases escolhidas para a geração final do RExtAut. Por sua vez, o RV é a categoria com menor amplitude interquartil (7,5), o que se traduz numa predisposição à seleção de frases mais curtas para transmitir mensagens científicas por parte dos divulgadores de ciência encarregues pela elaboração destes resumos. Em suma, pode-se constatar que, no que concerne à média de frases por classe de resumo, as categorias geradas manualmente apresentam um menor número de palavras por frase porque tomam em consideração a necessidade de transmitir conteúdo científico de forma eficaz, enquanto esta dimensão não é sopesada pelo Python.

5.2. Segunda pergunta de investigação

Nesta secção, debruçar-nos-emos sobre três questões fundamentais: o número de *tokens*, o número de *types* e o número de "palavras difíceis" para cada tipo de artigo considerado. Primeiramente, tentaremos responder à pergunta da variabilidade que existe, dentro das três categorias, no que diz respeito ao número total de palavras gráficas, isto é, aos *tokens*. Como esperado pelo número de frases mais elevado do RV, podemos observar na Figura 3 que o resumo vulgarizado apresenta uma mediana superior à das outras categorias (623 frente à mediana de 440 do RExtAut e 305 do ResExtV). Assim, o RV mostra uma maior amplitude interquartil, pois a distância entre o primeiro e o terceiro quartil é de 218.

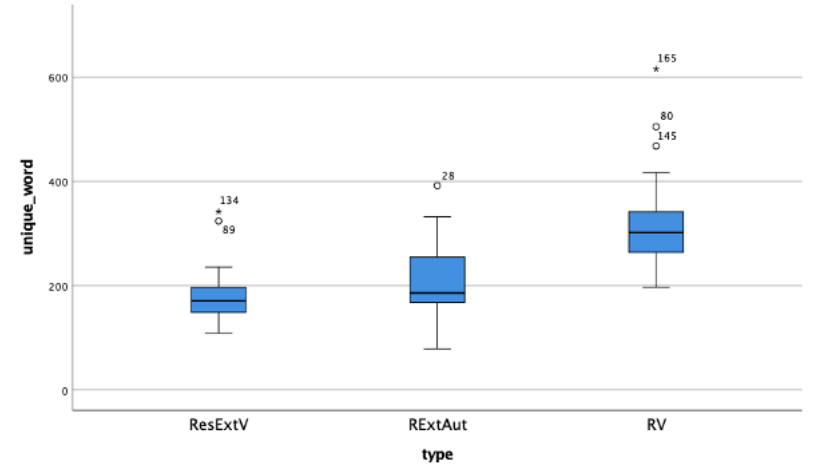
Figura 3 – Boxplot para a variável lex_count nos três subcorpora.



Para além de vermos que o ResExtV conta com a menor variabilidade (a sua amplitude interquartil é de apenas 105), apreciamos também que apresenta menos *outliers* e menor distância entre o máximo (210) e o mínimo (38), sendo que isto corrobora ainda a hipótese da primeira pergunta de investigação, em que se tencionava verificar a preferência da geração humana por frases mais curtas e com menor número de palavras. Do mesmo modo, devemos afirmar que os *outliers* 89 e 88 (no ResExtV e no RExtAut, respetivamente) se referem ao artigo 18, o que se pode traduzir numa dificuldade acrescentada para esse caso concreto.

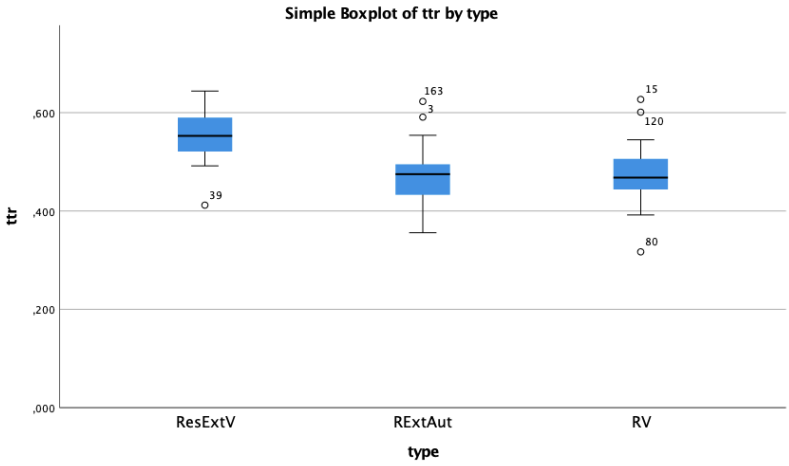
Por sua parte, no que concerne ao número de *types* ou palavras lexicais, como é possível observar no *boxplot* da Figura 4, a amplitude interquartil é maior no RExtAut, embora a mediana (186) seja bastante inferior àquela do RV (302). Desse facto podemos deduzir que a abstração manual (isto é, o RV) tende a concentrar-se mais em palavras lexicais com significado referencial (ou seja, verbos, adjetivos, substantivos e advérbios), enquanto que a geração automática pode ainda incorporar aspetos como interjeições, números ou citações que não contribuem para o cômputo das palavras únicas (*types*). Assim, conclui-se que a maioria dos *outliers* assinalados coincidem ainda com os do gráfico da figura 3. Não obstante, estas conclusões podem ser mais acertadas se analisarmos o *boxplot* referente ao TTR (Figura 5).

Figura 4 – Boxplot para a variável unique_word nos três subcorpora.



Se fixarmos a nossa atenção no cálculo do TTR para cada uma das categorias com que trabalhamos, podemos ver que a mediana (0,55) e a μ (0,58) no ResExtV são claramente superiores a estas medidas para as outras duas categorias, o que implica uma maior densidade lexical, isto é, uma maior concentração de palavras únicas para as palavras totais. Para o TTR, quanto mais próximo for de 1, maior densidade lexical mostra o texto. A amplitude interquartil é superior no RExtAut, categoria que apresenta uma maior variabilidade, o que também pode ser provocado pela presença de elementos que causam ruído no cálculo automático, como referências a figuras ou pontuação irregular que computam como sendo *tokens*.

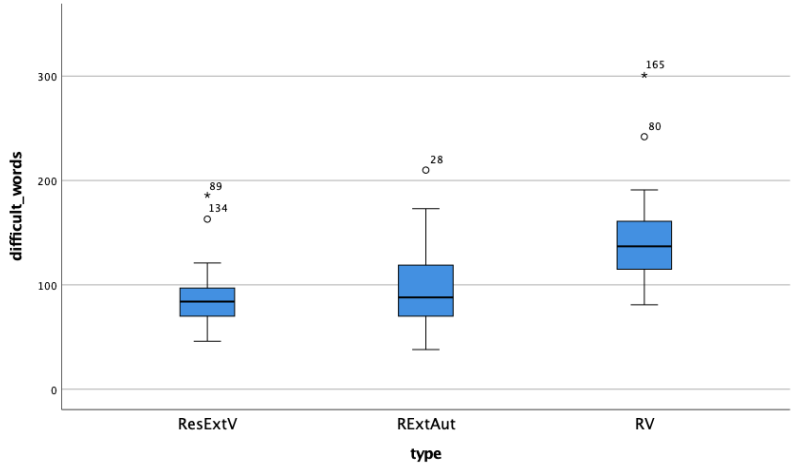
Figura 5 – Boxplot para a variável ttr nos três subcorpora.



Assim, os *outliers* da Figura 5 também não se referem ao mesmo artigo, o que não permite estabelecer uma alusão direta a um texto determinado. Além disso, pela primeira vez contamos neste *boxplot* com *outliers* inferiores ao valor mínimo, tal é o caso do *outlier* 39 no ResExtV e do 80 no RV, que têm um TTR muito pouco significativo.

Por último, centraremos a nossa atenção na contagem das palavras difíceis para estas três categorias. Como se pode contemplar na Figura 6, o ResExtV e o RExtAut apresentam uma mediana inferior (84 e 88 respetivamente) à do RV (137), embora a variabilidade ou amplitude entre mínimo e máximo seja maior no RExtAut (172). O ResExtV conta com menor variabilidade, o que se poderia associar a outros parâmetros valorados anteriormente de forma a concluir que essa menor dispersão pode estar relacionada com o facto de as frases serem mais curtas ou com a menor presença de palavras totais e únicas. Por sua vez, os *outliers* coincidem quase precisamente com aqueles que apresentava o *boxplot* das palavras únicas (Figura 4), o que evidencia que pode existir algum caso especial para esses artigos. Com esta análise, conseguimos encerrar, portanto, a segunda questão de investigação, ao termos oferecido respostas objetivas e baseadas em dados estatísticos às perguntas *supra* formuladas.

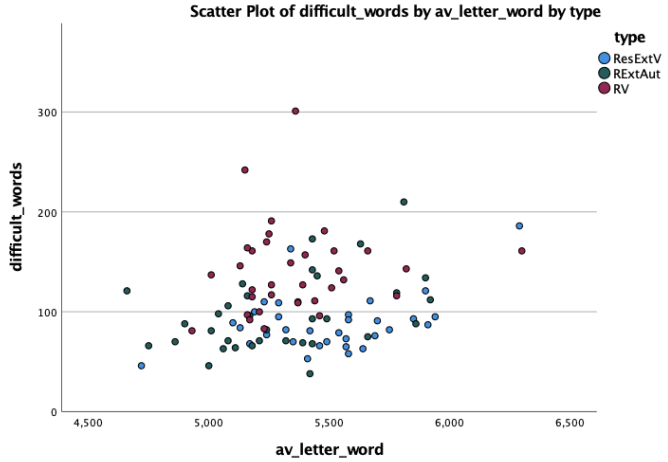
Figura 6 – Boxplot para a variável difficult_words nos três subcorpora



5.3. Terceira pergunta de investigação

No que diz respeito à terceira questão de investigação, tenciona-se pesquisar qual é a relação que existe entre as "palavras difíceis" e a média de letras por palavra, bem como perceber qual é a sua intensidade. Para isso, decidiu-se criar um *scatter plot* que nos permitirá observar a correlação entre estas duas variáveis. Na Figura 7, neste caso, é possível observar que existe uma correlação muito fraca entre os valores das duas variáveis, já que os valores estão bastante dispersos e há um número considerável de *outliers*, especialmente em resumos vulgarizados. Ao calcular o coeficiente de correlação de Pearson (ρ), que corresponde a 0,189, percebe-se que existe de facto uma correlação positiva desprezível entre as duas variáveis, sendo que um resultado tão próximo de 0 implica quase a inexistência de dependência linear entre as "palavras difíceis" e a média de letras por palavra.

Figura 7 – Scatter plot para a correlação entre as variáveis difficult_words e av_letter_word nos três subcorpora



5.4. Quarta pergunta de investigação

Nesta secção pretende-se responder à quarta questão de investigação, que centrava o seu foco na variável **read_ease_cat**, que foi criada *ad hoc* de modo a conseguirmos uma melhor visualização dos dados estatísticos. Assim, como *supra*mencionado, mantemos sete categorias que nos permitem ver a facilidade/dificuldade dos textos ou das categorias textuais com as quais trabalhamos. Para isso, lançamos mão de tabelas cruzadas que nos permitem analisar a distribuição percentual da facilidade de leitura para cada uma das nossas categorias.

A Figura 8 mostra o crosstabs entre a variável da facilidade de leitura e o tipo de texto. Vemos, claramente, que 100% dos artigos classificados como "bastante fácil" correspondem à categoria do RV, enquanto 100% dos textos catalogados como "normal" se referem ao ResExtV. Por sua parte, todos os RExtAut são atribuídos às categorias "difícil" ou "muito confuso", o que se implica uma maior dificuldade e menor legibilidade para os resumos elaborados de forma automática com o Python.

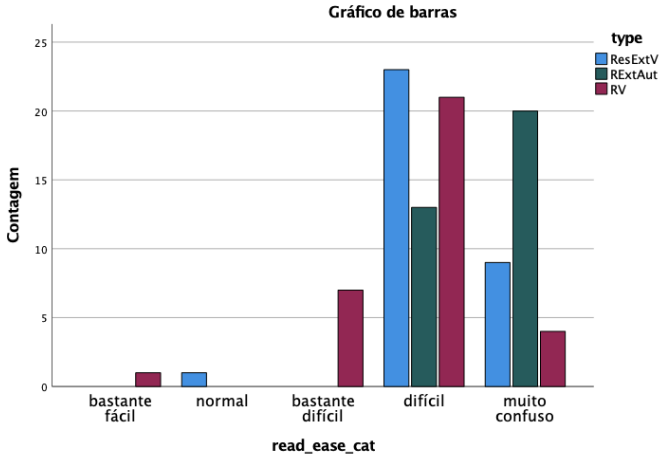
Figura 8 – Tabulação cruzada read_ease_cat (categórica) nos três subcorpora.

Tabulação cruzada read_ease_cat * type

			type			
			ResExtV	RExtAut	RV	Total
read_ease_cat	bastante fácil	Contagem	0	0	1	1
		% em read_ease_cat	0.0%	0.0%	100.0%	100.0%
		% em type	0.0%	0.0%	3.0%	1.0%
		% do Total	0.0%	0.0%	1.0%	1.0%
	normal	Contagem	1	0	0	1
		% em read_ease_cat	100.0%	0.0%	0.0%	100.0%
		% em type	3.0%	0.0%	0.0%	1.0%
		% do Total	1.0%	0.0%	0.0%	1.0%
	bastante difícil	Contagem	0	0	7	7
		% em read_ease_cat	0.0%	0.0%	100.0%	100.0%
		% em type	0.0%	0.0%	21.2%	7.1%
		% do Total	0.0%	0.0%	7.1%	7.1%
	difícil	Contagem	23	13	21	57
		% em read_ease_cat	40.4%	22.8%	36.8%	100.0%
		% em type	69.7%	39.4%	63.6%	57.6%
		% do Total	23.2%	13.1%	21.2%	57.6%
	muito confuso	Contagem	9	20	4	33
		% em read_ease_cat	27.3%	60.6%	12.1%	100.0%
		% em type	27.3%	60.6%	12.1%	33.3%
		% do Total	9.1%	20.2%	4.0%	33.3%
Total		Contagem	33	33	33	99
		% em read_ease_cat	33.3%	33.3%	33.3%	100.0%
		% em type	100.0%	100.0%	100.0%	100.0%
		% do Total	33.3%	33.3%	33.3%	100.0%

Para oferecermos uma representação mais visual dos dados, apresentamos esta relação com o gráfico de barras da Figura 9.

Figura 9 – Gráfico de barras para a variável read_ease_cat (categórica) nos três subcorpora.

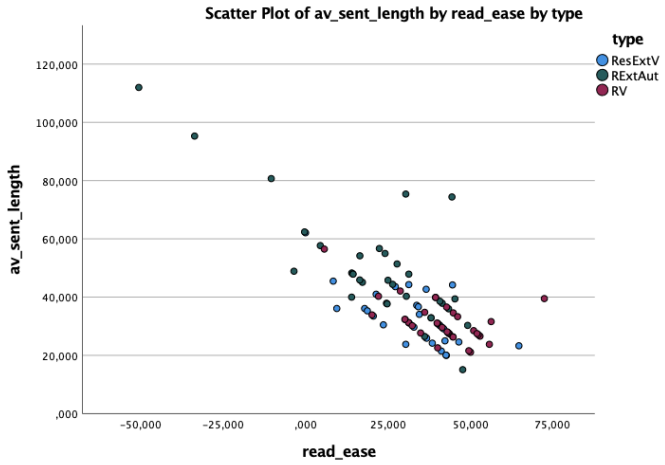


Do gráfico acima, podemos chegar à conclusão de que a maioria dos ResExtV são classificados como "difícil", ainda que existe uma parte deles classificada como "muito confuso". Esta tendência acrescenta-se ao olharmos para a categoria RExtAut, em que a maior parte (20) deles estão associados à classe "muito confuso" e nenhum é classificado como "bastante fácil" ou "normal". São os resumos vulgarizados, feitos por comunicadores de ciência, que mostram uma maior variabilidade neste sentido, pois 1 é "bastante fácil", alguns (6) são "bastante difíceis" e a maioria corresponde à categoria "difícil". Em suma, embora o RV seja uma categoria que deveria apresentar maior legibilidade tendo em conta que foi criado por divulgadores científicos, a comunicação de ciência ainda mostra impedimentos relativamente à simplificação textual. Este facto é evidenciado pela inexistência na análise destas categorias das classes "fácil" e "muito fácil". Destarte, e como antevisto, o ResExtAut é o mais difícil de compreender e, consequentemente, o que mais dificulta uma transmissão eficaz da mensagem científica.

5.5. Quinta pergunta de investigação

Como já se referiu anteriormente, de acordo com Hartley (2016), a fórmula Flesch para calcular a facilidade de leitura de um texto descarta quaisquer aspetos semânticos e só considera o tamanho das diferentes frases, assim como do texto no conjunto. Vamos comprovar agora, portanto, com dados estatísticos e matemáticos, qual é a correlação existente entre estas duas variáveis (a **av_sent_length** e a **read_ease**). Para tal fim, criámos um *scatter plot*, que nos permitirá ver se existe alguma correlação entre as duas variáveis.

Figura 10 – Scatter plot para a correlação entre as variáveis av_sent_lenght e read_ease nos três subcorpora.



Observando o gráfico da Figura 10, deduzimos que existe uma relação bastante significativa entre estas duas variáveis, já que quando a **read_ease** aumenta, os valores atribuídos à **av_sent_length** claramente diminuem. A partir daí podemos concluir que quanto maior facilidade de leitura, menor a média de tamanho frásico. Este comportamento mantém-se, aliás, nas três categorias selecionadas, embora no caso do RExtAut contemos com vários *outliers*, representados como pontos separados à esquerda. Como já foi afirmado, o valor de uma variável decresce (no eixo vertical ou dos Y) com o aumento do valor da outra variável (no eixo horizontal ou dos X), o que significa uma correlação negativa entre elas. Aliás, o resultado do cálculo da ρ de Pearson é -0,771, o que implica uma correlação negativa elevada, por estar muito próxima a -1. Desta análise, podemos concluir como resposta à quinta pergunta de investigação que as duas variáveis escolhidas apresentam uma grande correlação ou dependência entre si, sendo que neste caso quando uma aumenta, a outra diminui (correlação negativa).

5.6. Sexta pergunta de investigação

A questão que se segue leva-nos a relacionar os segundos necessários para ler os artigos com a sua facilidade de leitura. Para retirar conclusões, começámos por fazer um gráfico com as três categorias (ResExtV, RExtAut, RV). No entanto, visto que se apreciou uma correlação mais forte no caso dos ResExtV e RExtAut, decidimos separá-las para retirar considerações mais relevantes. No caso do primeiro gráfico em análise (Figura 11), assiste-se a uma correlação negativa relativamente próxima de -1, mais especificamente com um valor de -0,726, o que indica uma correlação forte. Por essa razão, é possível dizer que quanto menos tempo for necessário para ler os resumos extrativos, mais fácil será lê-los. Por outro lado, quando tratamos apenas os resumos vulgarizados (Figura 12), percebe-se que existe uma correlação agora positiva, nomeadamente de 0,378, mas significativamente mais fraca, pelo que não se pode chegar à mesma conclusão.

Figura 11 – Scatter plot para a correlação entre as variáveis reading_time_secs e read_ease nos resumos extrativos.

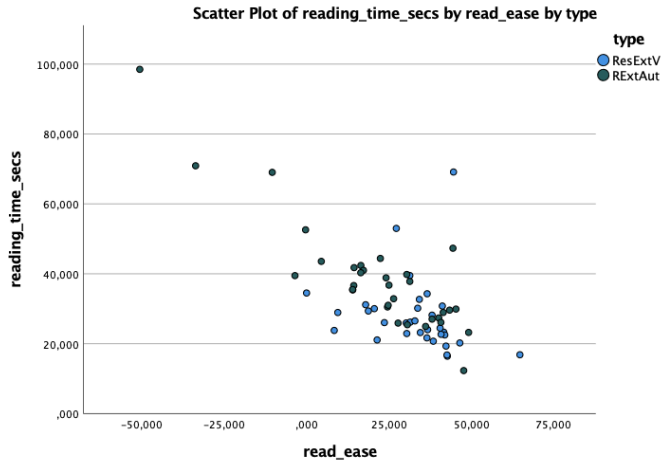
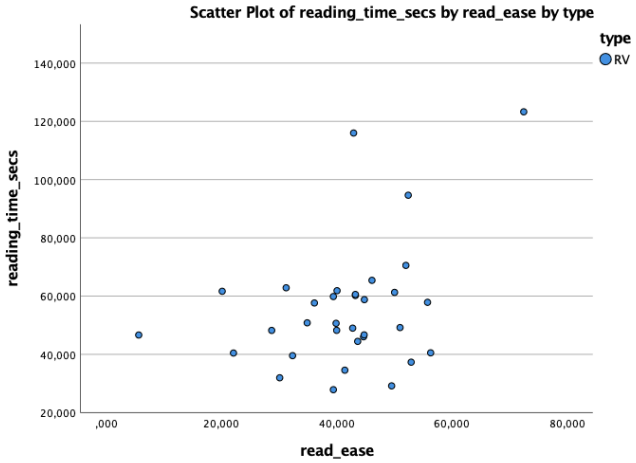


Figura 12 – Scatter plot para a correlação entre as variáveis reading_time_secs e read_ease no RV.



5.7. Sétima pergunta de investigação

Para a análise desta questão, partimos da hipótese de que os textos de especialização científica devem apresentar carácter objetivo e uma polaridade com tendência à neutralidade, pelo facto de não terem uma forte componente argumentativa e pela ausência de asseverações parciais. De modo a alcançarmos uma resposta à pergunta formulada, decidimos utilizar as variáveis categóricas geradas para a polaridade e a subjetividade e realizar uma tabela cruzada para ver como se distribuem as frequências.

Figura 13 – Tabulação cruzada para as variáveis relacionadas com a polaridade e a subjetividade categóricas.

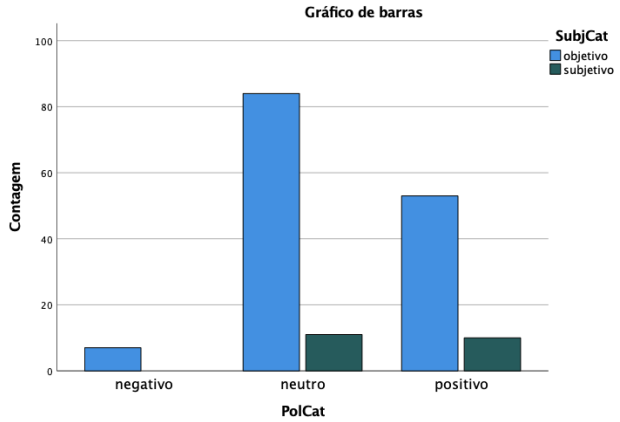
Tabulação cruzada PolCat * SubjCat

			SubjCat		
			objetivo	subjetivo	Total
PolCat	negativo	Contagem	7	0	7
		% em PolCat	100.0%	0.0%	100.0%
		% em SubjCat	4.9%	0.0%	4.2%
		% do Total	4.2%	0.0%	4.2%
	neutro	Contagem	84	11	95
		% em PolCat	88.4%	11.6%	100.0%
		% em SubjCat	58.3%	52.4%	57.6%
		% do Total	50.9%	6.7%	57.6%
	positivo	Contagem	53	10	63
		% em PolCat	84.1%	15.9%	100.0%
		% em SubjCat	36.8%	47.6%	38.2%
		% do Total	32.1%	6.1%	38.2%
Total		Contagem	144	21	165
		% em PolCat	87.3%	12.7%	100.0%
		% em SubjCat	100.0%	100.0%	100.0%
		% do Total	87.3%	12.7%	100.0%

Como é visível na Figura 13, na polaridade negativa, todos os textos são objetivos, e daí também é possível deduzir que apenas 7 do total (representação de 4,2% no conjunto) têm polaridade negativa. Os outros textos apresentam, então, uma polaridade neutra ou positiva, sendo que a maioria deles se concentra na polaridade neutra (57,6%). No caso dos neutros, 88,4% são objetivos e somente 11,6% são subjetivos. Para os textos com polaridade positiva, a maioria continua a ter pouca subjetividade, apesar de 10 deles apresentarem um carácter mais subjetivo.

Em seguida, para visualizarmos esta distribuição, considerámos pertinente desenhar um gráfico de barras que corresponda ao comportamento descrito na Figura 14. Desta forma, na Figura 14 podemos observar que quando falamos em subjetividade, a distribuição entre textos neutros e positivos é relativamente idêntica. No entanto, devemos tomar em consideração que o carácter subjetivo não surge em nenhum caso com uma polaridade negativa. Igualmente, os textos com polaridade neutra e carácter objetivo acabam por ser os que contam com um maior número de itens, pois como podia ser previsto, a maior parte de artigos de especialização científica apresentam um comportamento objetivo e neutro. Consequentemente, temos neste caso um gráfico bastante elucidativo que nos permite compreender melhor a relação entre estas duas variáveis qualitativas, para além de permitir corroborar a nossa hipótese de partida.

Figura 14 – Gráfico de barras para a relação das variáveis construídas com a polaridade e a subjetividade categóricas.



6. Conclusão

Depois da análise estatística realizada na secção anterior, é possível refletir, de um ponto de vista objetivo e centrado em conceitos matemáticos, acerca dos resultados obtidos ao longo do desenvolvimento do trabalho. Primeiramente, iremos rever os objetivos apresentados na fase inicial de conceção do projeto. Neste sentido, afirma-se que o presente trabalho representa um contributo bem-sucedido e essencial para a problemática de simplificação científica e geração de resumos através de diferentes métodos. Aliás, assumiu-se uma atitude crítica perante a compilação de subcorpora a partir de diferentes tipologias textuais.

No que diz respeito à metodologia empregue, explorámos diferentes caminhos de ciência de dados através de variados métodos tanto manuais como automáticos, onde se destaca o recurso à interface de programação Python e a algumas bibliotecas disponíveis. O facto de aplicarmos uma metodologia como esta permitiu a realização, *a posteriori*, duma análise estatística, cujas conclusões mais relevantes serão apresentadas para concluir este estudo:

1. Quanto ao tamanho dos textos com os quais trabalhámos, deduzimos que os resumos extrativos automáticos são os que apresentam frases mais extensas e também maior variabilidade no que concerne a este parâmetro. Como consequência, verifica-se que o Python tem tendência para extrair frases com maior cumprimento, enquanto a extração e/ou abstração manual não revela esse comportamento, preferindo-se então frases mais curtas para a transmissão da mensagem.
2. Como previsto, o RV conta com um número mais elevado de *tokens*, devido ao cumprimento quer das frases quer do texto em geral. Relativamente às palavras únicas ou lexicais, e como consequência do concluído anteriormente, também é a categoria RV a que apresenta mais *types*. Não obstante, quando estabelecemos os valores para a variável TTR, são os resumos extrativos manuais que têm um TTR mais próximo de 1, o que se traduz em maior densidade lexical. Para finalizar, é o RV a classe com um número mais significativo de palavras difíceis, que podem referir-se ocasionalmente a lexemas bastante comuns na língua geral.
3. O raciocínio associado à terceira questão de investigação permitiu-nos afirmar que a correlação entre as palavras difíceis e a média de letras por palavra é significativamente próxima de 0, o que implica a sua quase total inexistência.

4. Ao categorizar a facilidade de leitura em cada tipo de texto, observámos que na sua grande maioria os artigos de especialização científica se destacam pela sua dificuldade e baixa legibilidade. Este comportamento é especialmente relevante para a categoria dos RExtAut, onde todos os textos são considerados "difíceis" ou "muito confusos".
5. Do cruzamento da variável **av_sent_length** com a variável **read_ease**, conclui-se que a fórmula Flesch para a facilidade de leitura está fortemente correlacionada com a média do tamanho das frases. Como se tinha conjeturado, a facilidade de leitura é um parâmetro quantitativo que considera principalmente o tamanho das orações, não atendendo ao conteúdo semântico.
6. Da sexta pergunta de investigação deduz-se que, para as categorias ResExtV e RExtAut, quanto maior for a facilidade de leitura dos textos, menos tempo é necessário para a sua leitura. Não obstante, esta tendência é mais aleatória e não se pôde verificar relativamente aos resumos vulgarizados feitos por divulgadores de ciência.
7. Como antevisto, a maior parte dos textos que analisámos são objetivos e têm uma polaridade neutra, já que se trata de artigos científicos. Aliás, há simultaneamente um número significativo de textos com polaridade positiva e que são objetivos. Devemos ainda salientar que nenhum dos textos considerados subjetivos têm polaridade negativa.

Em suma, esta pesquisa possibilitou a análise estatística alguns dados matemáticos processados com Python de forma a retirar conclusões sobre as categorias de resumos extrativos manuais e automáticos, ao compará-los também com um resumo abstrativo redigido por divulgadores de ciência. Para o futuro, esperamos que a metodologia de ciência de dados possa ser adaptada para levar a cabo trabalhos com corpus em português, uma vez que algumas funções não se encontram atualmente disponíveis, e o fase de pré-processamento textual pode ser influenciada pelas mesmas.

Notas de fim

(1) A plataforma ScienceX (disponível em https://sciencex.com) define-se como uma rede de variadas páginas web relacionadas com a atividade científica e tecnológica. Trata-se, em síntese, de um repositório científico onde se publicam artigos e descobertas científicas de forma regular.

(2) Falamos em resumos abstrativos porque, de acordo com Lloret (2021, p. 89), os elementos textuais do artigo original foram não só selecionados através de diferentes técnicas, mas também alterados do ponto de vista linguístico. A informação é, em suma, transmitida de outra forma, pois neste caso são divulgadores de ciência que fizeram os resumos vulgarizados.

(3) O conjunto de dados pode ser disponibilizado mediante pedido aos autores deste artigo.

Referências

Atkins, S., Clear, J., & Ostler, N. (1992). Corpus Design Criteria. *Literary and Linguistic Computing*, 7(1),1-16.

Batarseh, F. A., & Yang, R. (2020). *Data Democracy. At the nexus of artificial intelligence, software development, and knowledge engineering*. Elsevier Academic Press.

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. O'Relly.

Cavique, L. (2014). Big data e data science. *Boletim da APDIO, 51*, 11-14. http://hdl.handle.net/10400.2/3918

Cheng, J., & Lapata, M. (2016). Neural summarization by extracting sentences and words. *Proceedings of the Association for Computational Linguistics (ACL)*, 484-494.

Davenport, T. H., & Patil, D. J. (2012). *Data Scientist: The Sexiest Job of the 21st Century*. Harvard Business Review. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

European Comission (2016). *Open Innovation. Open Science. Open to the World: A Vision for Europe.* Publications Office of the European Union. https://data.europa.eu/doi/10.2777/552370

Hartley, J. (2016). Is time up for the Flesch measure of reading ease? *Scientometrics, 107(3)*, 1523-1526.

Layton, R. (2015). *Learning Data Mining with Python*. Packt Publishing.

Lloret, E. (2021). Enfoques y retos para la generación automática de resúmenes. *Archiletras Científica, 6*, 87-103.

Masuzzo, P. (2017). *Do You Speak Open Science? Resources and Tips to Learn the Language*. PeerJ Prints. https://doi.org/10.7287/peerj.preprints.2689v1

Mitchell, D. (2015). Type-token models: a comparative study. J*ournal of Quantitative Linguistics, 22(1)*, 1-21. https://doi.org/10.1080/09296174.2014.974456

Mirowski, P. (2018). The future(s) of open science. *Social Studies of Science, 48(2)*, 171-203. https://doi.org/10.1177/0306312718772086

Newman, R., Chang, V., Walters, R. J., & Wills, G. B. (2016). Web 2.0. The past and the future. *International Journal of Information Management, 36*, 591-598. http://dx.doi.org/10.1016/j.ijinfomgt.2016.03.010

Provost, F., & Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *BigData, 1(1)*, 51-59. https://doi.org/10.1089/big.2013.1508

Ribeiro, C., Rodrigues, E., Matos, M. E., & Saraiva, R. (2010). Os Repositórios de Dados Científicos: Estado da Arte. *Projeto RCAAP. D24 – Relatório*. https://hdl.handle.net/10216/2