UTILIZANDO AS TÉCNICAS DE “NUVEM DE PALAVRAS” E CLUSTERIZAÇÃO APLICADAS AS ENTREVISTAS DOS ATLETAS OLÍMPICOS DA CIDADE DE SÃO CARLOS

Resumo - Esse artigo pretende demonstrar formas de se descobrir, utilizando técnicas da mineração de textos, assuntos relevantes dentro de conjuntos de textos não estruturados. Foram consideradas as entrevistas de quatro atletas olímpicos nascidos na cidade de São Carlos. Essas entrevistas foram realizadas pelo Grupo de Estudos Olímpicos, da Universidade de São Paulo (USP). Como aplicação, foi utilizada a técnica da nuvem de palavras (*WordCloud*) e da clusterização (agrupamento).

Palavras-chave: Nuvem de palavras, atletas Olímpicos, entrevistas.

USING THE "WORDCLOUD" AND CLUSTERIZATION TECHNIQUES APPLIED TO THE INTERVIEWS OF THE OLYMPIC ATHLETES OF THE CITY OF SÃO CARLOS

Abstract – This paper aims to demonstrate ways of discovering, using text mining techniques, relevant subjects within sets of unstructured texts. The interviews of four Olympic athletes born in the city of São Carlos were considered. These interviews were conducted by the Group of Olympic Studies, University of São Paulo (USP). As a tool, it was used “WordCloud” and clusterization (grouping) techniques.

Keywords: Wordcloud, Olympic athlets, interviews.

UTILIZANDO LAS TÉCNICAS DE "NUBE DE PALABRAS" Y CLUSTERIZACIÓN APLICADAS EN LAS ENTREVISTAS DE LOS ATLETAS OLÍMPICOS DE LA CIUDAD DE SÃO CARLOS

Resumen - Este artículo pretende demostrar formas de descubrir, utilizando técnicas de la minería de textos, asuntos relevantes dentro de conjuntos de textos no estructurados. Se consideraron las entrevistas de cuatro atletas olímpicos nacidos en la ciudad de São Carlos. Estas entrevistas fueron realizadas por el Grupo de Estudios Olímpicos, de la Universidad de São Paulo (USP). Como aplicación, se utilizó la técnica de la nube de palabras (*WordCloud*) y de la clusterización (agrupación).

Palabras-clave: Nube de palabras, atletas Olímpicos, entrevistas.

Rovilson de Freitas

Instituto de Ciências Matemáticas e de Computação

Universidade de São Paulo

rovilson.freitas@usp.br

Ruan Felipe de Oliveira Neves

Instituto de Ciências Matemáticas e de Computação

Universidade de São Paulo

rfo.neves@gmail.com

Victor Henrique Gonçalves

Instituto de Ciências Matemáticas e de Computação

Universidade de São Paulo

victorhenriquenator@gmail.com

http://dx.doi.org/10.30937/2526-6314.v2n2.id41

Introdução

Ao longo dos últimos vinte anos, o Grupo de Estudos Olímpicos da Escola de Educação Física da USP, liderado pela Professora Doutora Katia Rubio, entrevistou grande parte dos atletas que estiveram em pelo menos uma edição olímpica. Foram mais de 1200 entrevistas, dos mais de 2000 olímpicos brasileiros.

Essas entrevistas, gravadas em vídeo e transcritas em formato texto, representam um rico acervo da história do esporte brasileiro. Sob vários aspectos, esse material inexplorado pode trazer diversas reflexões. E não apenas no campo esportivo ou competitivo, mas trazer a luz outras questões importantes sobre o país. Rubio[1] afirma que busca em suas histórias de vida suas singularidades, suas origens sociais e culturais. Também podemos considerar que as entrevistas podem demonstrar diversos aspectos, como políticas públicas e até mesmo, características regionais e temporais do esporte no país.

Para tanto, é impensável não utilizar alguma ferramenta computacional para auxiliar o processo. O volume de textos é muito grande, e por sua característica não estruturada, as dificuldades são ainda maiores. Pensando num trabalho exclusivamente manual, seria necessário a utilização de uma grande quantidade de recursos (humanos, financeiros e de tempo) para que algum resultado objetivo fosse atingido. Além disso, um trabalho manual implica, necessariamente, de uma maior dependência do fator humano. Isso traz, como consequência, uma menor precisão nos resultados.

Algumas técnicas da mineração de textos podem contribuir para que esse trabalho seja feito de forma mais inteligente, trazendo resultados de maneira automatizada, de forma mais rápida, e com maior precisão.

Para esse trabalho, foram selecionadas as entrevistas de quatro atletas olímpicos nascidos na cidade de São Carlos. Eles representam a totalidade de são-carlenses que disputaram, pelo menos, uma edição olímpica.

Como hipótese o trabalho visa, de maneira inicial, buscar assuntos que poderiam ser estudados dentro desse universo. Parte-se do princípio que, para esse contexto, os temas que podem ser relevantes para um estudo mais aprofundado são desconhecidos. Características desse grupo em especial podem sugerir assuntos que poderiam ser estudados tanto dentro do próprio contexto ou em comparação a outros grupos diversos (como de outras cidades da região, ou da mesma modalidade em outros locais, etc.).

Portanto, de que maneira o uso de algumas técnicas da mineração de textos poderia contribuir para sugerir alguns temas relevantes para pesquisas posteriores?

Foram considerados para a análise as seguintes técnicas: primeiramente foi gerado a partir dos textos processados, uma nuvem de palavras. Em seguida, nos mesmos textos foi aplicado o recurso de clusterização (agrupamento).

A priori, tanto a nuvem de palavras quanto a clusterização foram realizadas de forma individual. Ao final, os textos dos quatro atletas foram agrupados, para que uma nova análise fosse realizada, agora considerando a totalidade, e não apenas as partes.

Grupo escolhido e seleção dos textos

Os quatro atletas selecionados para esse estudo, foram entrevistados entre os anos de 2012 e 2015. São dois atletas do futebol: Fábio Aurélio e Mônica. Do triatlo, Carla Moreno e do atletismo, Maurren Maggi.

Considerando o fator gênero, 75% dos atletas são mulheres (3 no total). Se o critério for a modalidade, metade (50%) são atletas do futebol. Em comum eles têm outros fatores. Os quatro atletas nasceram num intervalo de 3 anos (1976-1979). Por terem nascido num mesmo período histórico, a chance de terem competido em edições olímpicas próximas aumenta. Os quatro estiveram na mesma edição (Sydney-2000). Todos fizeram sua estreia olímpica nessa edição.

Duas atletas ainda competiram em 2004. Outra atleta disputou os Jogos de 2008 e 2012. Duas atletas foram medalhistas. O jogador de futebol Fábio Aurélio, disputou uma única edição. No caso do futebol masculino, é comum que o atleta dispute uma única edição, visto que essa prova olímpica tem uma característica específica: apenas atletas com menos de 23 anos podem disputar a competição. Existe uma exceção para atletas com mais de 23 anos: três podem ser inscritos. Soma-se também ao fato de que um atleta do futebol se profissionaliza muito mais cedo que outras modalidades. Com isso, existe a dificuldade de liberação de seus clubes, já que os Jogos Olímpicos não são considerados campeonatos oficiais entre os homens (como são os torneios continentais ou Copa do Mundo). Portanto, é raro que atletas olímpicos do futebol masculino disputem mais de uma competição olímpica. Não há restrições no futebol feminino.

As entrevistas foram gravadas em vídeo, e posteriormente transcritas em formato docx (Microsoft Word). Dessas transcrições, surgiram os textos que serão utilizados no trabalho. Juntas, totalizaram 35715 palavras (nessa contagem, ainda estão incluídas as

stopwords). As entrevistas não tinham roteiro pré-definido: os atletas eram convidados e contar sua história de vida, de maneira livre. Não havia nenhum questionário ou direcionamento, o que caracteriza os textos analisados como altamente orais. Eles decidiam por onde começar, como começar e a sequência a partir disso.

Ferramenta Utilizada

Para efetuar as análises das entrevistas selecionadas, foi utilizado como ferramenta a linguagem de programação Python. O ambiente de desenvolvimento escolhido foi o Jupyter Notebook. Esse ambiente, permite ao analista de dados modularizar o processo de análise, usando o recurso de células. Além disso, também permite a exibição dos resultados em diferentes formatos, podendo ser texto, tabelas ou gráficos.

Em todos, o primeiro passo foi o pré-processamento dos textos. Todos eles tiveram retiradas termos que são considerados desnecessários: as chamadas *stopwords*. Em praticamente todo processo de mineração, são removidas as preposições, artigos, adjetivos, advérbios, alguns verbos e substantivos. Cada processo vai direcionar a remoção, visto que dependendo da necessidade, algum grupo específico de palavras é necessário.

No caso das entrevistas, por se tratarem das falas individuais dos atletas sobre sua prática pessoal, é comum encontrar uma série de termos repetidos, comuns na oralidade. Foi necessário, portanto, um cuidado maior na remoção das *stopwords*, para garantir que os resultados não mostrassem termos desnecessários e que poderiam atrapalhar o processo de análise desses resultados.

Nuvem de Palavras

A primeira estratégia utilizada foi a de Nuvem de palavras (*WordCloud*). Em uma nuvem de palavras, cada palavra tem seu tamanho e intensidade de cor regidos pela relevância em determinado corpus[2]. A nuvem de palavras não deve ser a única ferramenta utilizada, mas pode apontar para que sentido o analista pode caminhar. Um ponto de partida, para direcionar as análises. Cruz e Bonfante[3] apontam como exemplo de utilização da *WordCloud* a análise de diferentes textos, na busca de similaridade para identificação do autor de um determinado grupo de livros.

Na nuvem de palavras, quanto maior a frequência de um determinado termo, maior será sua apresentação na imagem. Sua cor também é destacada. Um termo citado 100 vezes, terá seu tamanho maior que um outro termo citado 50 vezes.

Para elaborar nuvens de palavras de arquivos de textos no *Jupyter Notebook*, foram utilizadas as bibliotecas *WordCloud* e *Matplotlib.pyplot*.

Resultados das Nuvens de Palavras

Considerando os quatro atletas olímpicos de São Carlos, temos as seguintes nuvens.

Figura 1 - Nuvem de palavras da atleta Carla Moreno

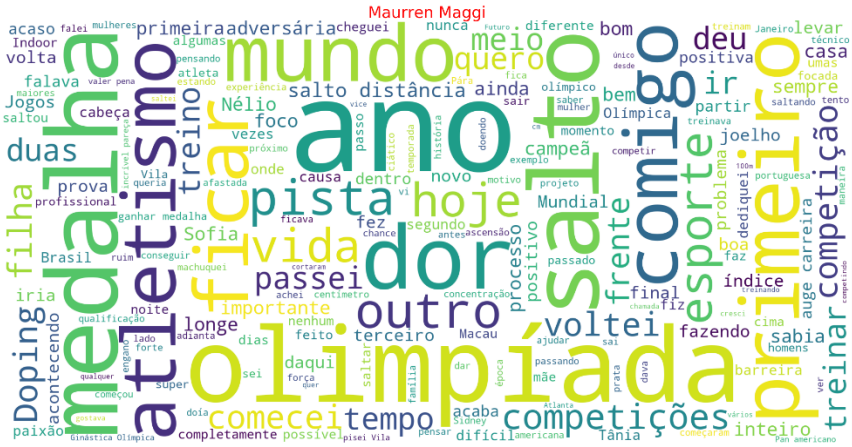
Na nuvem de palavras da triatleta Carla Moreno (Figura 1), notamos que um grande número de palavras com mais ou menos o mesmo número de termos. Prova, vai/vou, sei, fazer, bom, sabe, quero/queria, esporte, pode, nada (nesse contexto, relacionado ao verbo nadar), hoje, são termos que aparecem com frequência mais estável. São raras as menções a locais (cidade, estado, país), e mesmo a questões esportivas (a palavra triátlon, sua modalidade, aparece em quantidade menor que as citadas anteriormente). os termos “Jogos Olímpicos”, ou “Olímpiadas”, aparecem de forma discreta. Percebe-se também, muitos termos relacionadas com a palavra “tempo” (presente, passado, tempo, momento, sempre, amanhã, semana). Outro aspecto importante é a presença de termos relacionados a dor e/ou lesões. Dor, sangue, cirurgia, lesão, corpo, são termos que aparecem, embora sua frequência não seja alta. Uma

hipótese é de que a entrevista dessa atleta tenha focado em questões mais genéricas, não se aprofundando em nenhum aspecto mais específico.

Figura 2 - Nuvem de palavras do atleta Fábio Aurélio

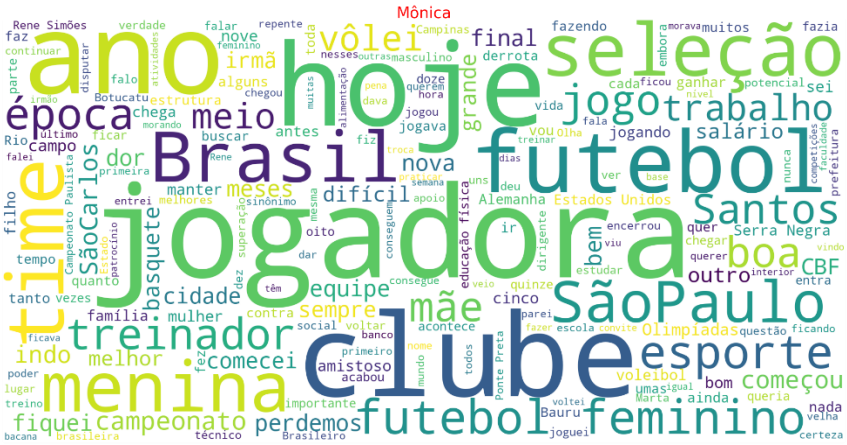
No caso do segundo atleta, Fábio Aurélio (Figura 2), fica claro que os termos específicos de sua modalidade e prática esportiva, foram dominantes na entrevista. Termos comuns a modalidade, como jogo/jogar/jogando/joguei, temporada, gol, treino, titular, clube, final, time, seleção, treinador, campeão, copa, convocação, profissional, futebol, bola, contrato tem frequência destacada. Ainda no campo esportivo, mas considerando a experiência do atleta, surgem os termos mais objetivos: São Paulo, Grêmio, Liverpool, Valência, clubes por onde o jogador passou, Corinthians como provável adversário. São Paulo (agora enquanto cidade), Inglaterra, São Carlos, Brasil, Americana, também aparecem com destaque, indicando uma provável indicação de deslocamento, comum a jogadores de futebol. Assim como na análise anterior, termos relacionados com lesões também estão presentes: dor, lesão, joelho, fisicamente, esquerdo. O tempo também surge como assunto relevante: tempo, meses/mês, semana, ano, momento. O termo olímpiada, também teve destaque ao longo do texto. Pré olímpico (campeonato que serve como classificatório para a participação nos Jogos), também aparece com alguma frequência. Como hipótese nessa avaliação, podemos considerar que o atleta focou sua entrevista em sua prática e trajetória esportiva, não abordando de maneira mais incisiva questões mais genéricas, ou pessoais.

Figura 3 - Nuvem de palavras da atleta Maurren Maggi



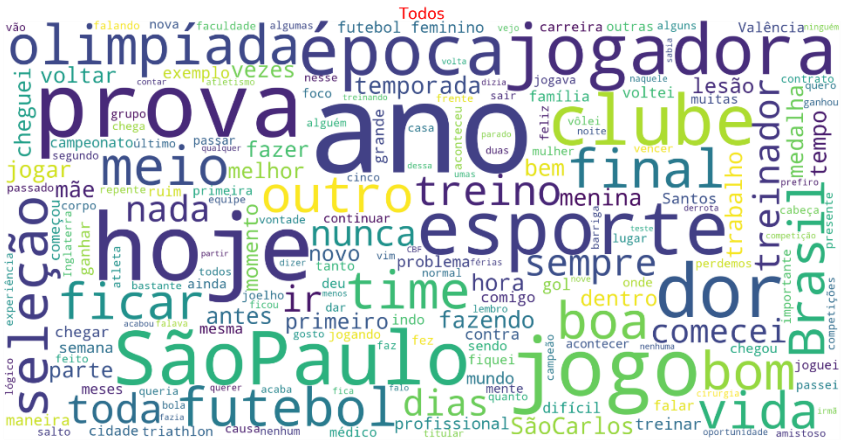
Na nuvem de palavras da atleta Maurren Maggi (Figura 3), notamos a presença importante de termos relacionados a sua modalidade, de maneira objetiva: atletismo, olímpiada, salto, medalha, esporte, primeiro, competição/competições, treinar, pista, campeã, doping. Assim como nas nuvens dos atletas anteriores, é marcante a presença de marcas de tempo: noite, ano, dias, janeiro, hoje, futuro. Também percebemos a presença de palavras que podem remeter a lesões, como dor, machuquei, joelho. Não percebemos a presença de lugares (cidade, estado ou país). Para essa atleta, fica claro em seu discurso a importância de sua prática esportiva, as modalidades e tudo que a cerca. A nuvem deduz que foi uma entrevista muito focada em sua carreira atlética, com poucas marcas subjetivas .

Figura 4 - Nuvem de palavras da atleta Mônica



Já na nuvem gerada pela entrevista dada pela jogadora de futebol Mônica (figura 4), nota-se a presença de termos que remetem ao gênero feminino: jogadora, mulher, futebol feminino, menina, mãe, irmã. Uma hipótese é de que exista ao longo da entrevista uma oposição ao futebol masculino, muito provavelmente comparativo. Encontramos também termos relacionados a sua modalidade, o futebol. Porém, ao contrário da atleta Maurren Maggi, do atletismo, os termos estão ligados a um aspecto mais gerencial: futebol seleção, esporte, treinador, campeonato, derrota, jogo, trabalho, equipe, amistoso, time. Obviamente, deve-se levar em conta, também, as diferenças entre as modalidades. Nota-se também a presença do termo “treinador “em oposição a “técnico”, numa frequência muito maior. Assim como aconteceu com o atleta Fábio Aurélio, percebe-se várias menções a lugares. Essa pode ser uma característica especial dessa modalidade, visto que é comum o deslocamento dos atletas em busca de oportunidades em novas equipes: Brasil, Alemanha, Estados Unidos, São Paulo, Santos, São Carlos, Botucatu, Bauru. Mais uma vez, assim como em todos os anteriores, o tempo é tema recorrente. Termos como ano, hoje, sempre, meses, reforçam que a questão temporal, pelo menos para esse grupo, é muito frequente.

Figura 5 - Nuvem de palavras dos atletas Olímpicos da cidade de São Carlos



Considerando a nuvem de palavras contendo as entrevistas de todos os atletas (figura 5), reforçamos o que já havíamos visto nas nuvens individuais. A questão temporal, os termos relacionados a lesões, menções a alguns locais (destacando-se São Paulo em oposição a própria cidade de São Carlos), além dos termos específicos do futebol (visto que metade dos atletas é dessa modalidade) dominam a representação. A

palavra olímpiada, um dos termos que poderia ser comum a todos eles, aparece numa frequência menor que os temas anteriores.

Clusterização

Podemos definir clusterização como o agrupamento de termos que apresentam alguma similaridade. Segundo Beltrame e Fonseca[4], os clusters descobertos podem ser usados para explicar as características da distribuição dos dados subjacentes e assim servir como base para várias técnicas de análise e mineração de dados.

Uma característica importante do processo de clusterização é que possível ter um entendimento do conjunto de dados inicial. Pode ser usada, inclusive, como parte do pré-processamento dos dados. O agrupamento permite identificar correlações que não seriam vistas de maneira elementar, direcionando a mineração e a análise para caminhos mais objetivos. Também fundamental reforçar que essa técnica tem característica não supervisionada, isso significa que os grupos são gerados sem nenhuma suposição anterior, não existe predefinição das classes, gerando um aprendizado sem supervisão prévia.

Nesse trabalho, foram utilizadas as seguintes bibliotecas do *Jupyter Notebook*: *TfidfVectorizer* e *K-means*. A biblioteca, permite transformer os textos brutos de maneira que possam ser utilizados pelo *K-means*, que, de fato, realiza o agrupamento. Normalmente, o agrupamento é realizado com os dados organizados num formato de tabela. Nesse caso, arquivos com extensão .CSV são os mais utilizados. Para esse caso, como os dados não eram estruturados, foi necessário esse tratamento prévio.

A título de exemplo, ficou determinado que em cada aplicação, seriam gerados dois clusters. E que cada cluster teria um total de 10 palavras.

Resultados

Na tabela abaixo (tabela 1), podemos ver quais foram os termos encontrados para cada um dos clusters.

Tabela 1: Resultados da clusterização das entrevistas

Atleta	Cluster 0	Cluster 1
Carla Moreno	prova, dia, nunca, triathlon, vida, cachorro, dias, bom, anos, maneira	não, nada, hoje, prova, bom, isso, dor, maneira, dia, mente
Fábio Aurélio	ano, Valência, Seleção, Inglaterra, Joelho, psicológico, agradeço, esquerdo, época, anos	temporada, São Paulo, pré, final, ano, jogo, olimpíadas, Seleção, olímpico, dor
Maurren Maggi	dia, dar, cima, volta, exemplo, saber, doía, passado, final, dor	ano, medalha, salto, mundo, fora, dor, olimpíada, atletismo, primeiro, olimpíadas
Mônica	futebol, São Paulo, Seleção, feminino, esporte, dor, treinador, sete, difícil, anos	não, hoje, dia, ano, futebol, anos, prova, dor, bom, esporte
Todos	jogadoras, clube, meninas, Santos, umas, novas, boas,bem, seleção, época	não, hoje, dia, ano, futebol, anos, prova, dor, bom, esporte

Os resultados são coerentes com aqueles previamente encontrados nas nuvens de palavras, com menções a lesões, lugares / deslocamentos, termos específicos das modalidades esportivas, e algumas questões subjetivas, provavelmente relacionadas ao indivíduo.

Por exemplos, no caso do cluster 0 do atleta Fábio Aurélio, notamos que existe similaridade entre os termos Valência/Inglaterra (locais) com joelho/esquerdo (parte do corpo) e ano/época. Pode significar uma lesão, em um desses locais de sua prática atlética, em determinado período de tempo.

Chama atenção a presença do termo dor. Ele é presente em todas as entrevistas. Gonçalves[5] discutiu o tema em sua dissertação de mestrado, considerando atletas da

ginástica rítmica. Um assunto que poderia ser mais explorado, considerando outras modalidades, como mostrado no agrupamento das entrevistas dos atletas de São Carlos.

Conclusões

Ao selecionar esse grupo restrito de atletas, sabíamos que poderia existir alguma coisa em comum entre eles. Nas discussões preliminares, cogitamos a possibilidade de, por terem nascido num mesmo período histórico, existido alguma ação externa (como alguma política pública para o esporte, ou até mesmo a presença de um projeto social ou organização não-governamental – ONG na cidade no processo de formação desses atletas) que justificasse o fato de que seus primeiros atletas olímpicos estivem juntos na mesma edição olímpica. Pelas entrevistas, não foi possível comprovar essa hipótese inicial. Por seu caráter específico, trabalhar com textos nos traz essa incerteza em uma primeira análise. Portanto, fazer uso de ferramentas que possam mostrar tendências dentro de um conjunto de textos, pode economizar o dispêndio de tempo e dinheiro em algo que, provavelmente, não teria trazido resultados objetivos para uma provável pesquisa. Em compensação, outras informações pertinentes surgiram, que podem resultar em análises que podem contribuir de maneira mais importante para outras pesquisas.

Considerações finais

A mineração de textos pode ser uma ferramenta importante para auxiliar na tomada de decisões, sob vários aspectos. Ao trabalhar com textos, em muitas situações não temos uma hipótese inicial sobre quais informações relevantes podem ser consideradas dos mesmos.

Muitas técnicas e algoritmos podem ser aplicados para direcionar e facilitar essa análise. A nuvem de palavras (*WordCloud*) e a clusterização (agrupamento), podem ajudar o analista no sentido de permitir que, dentro da realidade daqueles textos, possa perceber o que é relevante. Por consequência, direcionar seus esforços na análise, com maiores ganhos de tempo e performance.

No exemplo dos atletas olímpicos de São Carlos, ficou claro quais foram os termos mais frequentes das entrevistas dadas ao Grupo de Estudos Olímpicos. Isso permitirá que, em outras situações e contextos, o Grupo possa explorar em suas pesquisas aquilo que é mais relevante em sua fonte principal de estudos (os atletas

olímpicos), saindo de um campo apenas intuitivo e passando para uma realidade mais embasada em dados concretos.

Referências

1 Rubio K. Atletas Olímpicos Brasileiros. São Paulo. Editora Sesi; 2015.
2 Schwartz FP. Análise do discurso parlamentar por meio da técnica do processamento de linguagem natural: abordagem estatística e aprendizagem de máquina [relatório de pesquisa de estágio pós-doutoral]. Brasília: Universidade de Brasília, Departamento de Engenharia Elétrica; 2018.
3 Cruz FM, Bonfante AG. TideneTextVis: um módulo Python de visualização de textos baseado nas técnicas de tag clouds. Anais Escola Regional de Informática da Sociedade Brasileira de Computação (SBC) – Regional de Mato Grosso; 2016 [citado 03 dez. 2018]. Disponível em: http://anaiserimt.ic.ufmt.br/index.php/erimt/article/view/72/94.
4 Beltrame WAR, Fonseca FCS. Aplicações práticas dos algoritmos de clusterização K-means e Bisecting K-means. 2010 [citado 03 dez. 2018]. Disponível em: https://www.researchgate.net/profile/Walber_Beltrame/publication/327121358_Aplicacoes_Praticas_dos_Algoritmos_de_Clusterizacao_K-means_e_Bisecting_K-means/links/5b7b53a6299bf1d5a718d785/Aplicacoes-Praticas-dos-Algoritmos-de-Clusterizacao-K-means-e-Bisecting-K-means.pdf.
5 Gonçalves GCM. O significado da dor em atletas da ginástica rítmica [dissertação]. São Paulo: Universidade de São Paulo, Escola de Educação Física e Esporte; 2017.