Universidade Federal de São Carlos
Centro de Educação e Ciências Humanas
Programa de Pós-Graduação em Ciência da Informação

GUILHERME FRANCO SILVA PINTO

COMPORTAMENTO INFORMACIONAL E MINERAÇÃO TEXTUAL NO TWITTER

São Carlos
2018

GUILHERME FRANCO SILVA PINTO

COMPORTAMENTO INFORMACIONAL E MINERAÇÃO TEXTUAL NO TWITTER

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de São Carlos como requisito parcial para a obtenção do título de Mestre em Ciência da Informação, na Área de Concentração Conhecimento, Tecnologia e Inovação.

Orientadora: Profa. Dra. Ariadne Chloe Mary Furnival

São Carlos

2018

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Educação e Ciências Humanas
Programa de Pós-Graduação em Ciência da Informação

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Dissertação de Mestrado do candidato Guilherme Franco Silva Pinto, realizada em 24/08/2018:

Profa. Dra. Ariadne Chloe Mary Furnival
UFSCar

Prof. Dr. Ronaldo Ferreira de Araújo
UFAL

Profa. Dra. Ana Carolina Simionato
UFSCar

Certifico que a defesa realizou-se com a participação à distância do(s) membro(s) Ronaldo Ferreira de Araújo e, depois das arguições e deliberações realizadas, o(s) participante(s) à distância está(ao) de acordo com o conteúdo do parecer da banca examinadora redigido neste relatório de defesa.

Profa. Dra. Ariadne Chloe Mary Furnival

RESUMO

Este trabalho busca explorar o comportamento informacional dos usuários de mídias sociais, especificamente o Twitter, quando buscam e compartilham informações sobre Direito Autoral e suas vertentes. Tendo em vista as mudanças sociais, culturais e econômicas que acarretaram do uso das mídias sociais, exploramos como as adaptações da legislação de Direito Autoral às novas plataformas digitais afetam a utilização destas plataformas por seus usuários e como os usuários buscam informações relacionadas. Utilizamos o API do Twitter para coletar as postagens (*tweets*) feitas na plataforma que falem sobre Direito Autoral e suas áreas afins, como licenças livres, Creative Commons, domínio público etc. A seguir, utilizando análises de Mineração Textual dentro do ambiente para computação estatística R, foi possível examinar e relacionar os termos e palavras das postagens com as dúvidas e perguntas mais comuns sobre Direito Autoral. Finalmente, foi possível identificar exemplos de comportamento informacional dos usuários do Twitter durante suas interações com outros usuários e as informações disponíveis no Twitter.

Palavras-Chave: Comportamento Informacional. Twitter. Mineração Textual. Direito Autoral

ABSTRACT

This work aims to explore the information behavior of social media users, on Twitter specifically, at the moment they are searching and sharing information on Copyright Laws and its related subjects. Taking in consideration the social, cultural and economic changes that social media has made possible, this project will explore how the adapted Copyright Laws to digital platforms affect the usability of these platforms, and how its users seek information related to those adaptations. The Twitter API was used to collect the posts (tweets) made on the platform that are about Copyright and its related subjects, such as open licenses, Creative Commons, public domain, etc. Then, using Text Mining analysis in the statistical computing environment R, it was possible to assess and relate the terms and words used on these posts with the most common doubts and questions regarding Copyright. Finally, in this data it was possible to identify instances of information behavior of Twitter users during their interactions with each other and the information available on Twitter.

Keywords: Information behavior. Twitter. Text Mining. Copyright.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

CI – Ciência da Informação

TICs – Tecnologias de Informação e Comunicações

IDF – *Inverse Document Frequency* (Frequência Inversa do Documento)

SUMÁRIO

1 INTRODUÇÃO

Neste estudo colocamos em foco a intersecção entre dois campos temáticos da Ciência da Informação (CI), a saber: o Comportamento Informacional e o uso de Redes/Mídias Sociais como meios para obtenção de informações relevantes ao usuário inserido neste contexto digital. Especificamente, estudamos como usuários da rede social Twitter buscam e compartilham informações sobre direitos autorais e para tal, empregamos as técnicas de mineração textual.

O Comportamento Informacional, já amplamente discutido na CI, foi definido por Wilson como a "totalidade do comportamento humano em relação às fontes de informação, incluindo buscas passivas e ativas por informação, bem como seu uso" (2000, p. 49). Abrange situações como a necessidade de informação e sua identificação, diversos métodos de busca, como: conversas pessoais; capitação passiva da informação através de mídias como rádio ou televisão; buscas em bibliotecas ou plataformas especializadas. Passa também pela geração, compartilhamento, acesso e uso da informação, com interesse nas dimensões afetivas e emocionais das experiências informacionais e no processo decisório no momento da busca. São diversos e variados os modelos e abordagens epistemológicas elaborados e adotados na área de Comportamento Informacional. Assim como as bases de dados e bibliotecas digitais proporcionaram as primeiras pesquisas sobre informação online, agora, como observado por Khoo (2014, p. 90), "o crescimento das mídias sociais parece anunciar uma nova era na pesquisa em comportamento informacional"; contudo, ainda existe uma escassez de pesquisas que discutam esta perspectiva na Ciência da Informação.

A fim de preencher parte desta carência, buscamos apoio na literatura que aborda os usuários de mídias sociais em um sentido duplo, como criadores/autores e como leitores e disseminadores de informação, podendo, assim, buscar as informações que necessitam e compartilhar as informações que já possuem. Neste trabalho, tratamos mídias sociais como um conceito que cobre uma ampla variedade de aplicações na Internet que suportem comunicação entre indivíduos, sendo de forma direta ou indireta, síncrona ou assíncrona. Em especial quando valorizam a interação entre usuários, a geração de conteúdo por usuários e a criação de comunidades online (KHOO, 2014).

Deste modo, serão abordadas as razões para a utilização de mídias sociais e suas vantagens sobre outros métodos de busca (CONNAWAY; DICKEY; RADFORD, 2011; JANSEN et al., 2011; JAVA et al., 2007; JOHNSTON et al., 2013; SAVOLAINEN, 2001),

bem como, o comportamento informacional durante a busca e o compartilhamento de informações entre os usuários (ADAMIC et al., 2008; FOSTER, 2006).

A utilização do termo mídia social também é justificada pela classificação de alguns autores, como Khoo (2014) e Kaplan e Haenlein (2010), que identificaram treze serviços disponíveis na Internet como mídias sociais, o Twitter é alocado dentro da categoria blog/microblog. Os treze tipos são:

- Sites de redes sociais (e.g., Facebook);
- Projetos colaborativos (e.g., Wikipedia);
- Blogs e microblogs (e.g., blogspot.com e Twitter), que podem suportar jornalismo online;
- Comunidade de criação de conteúdo (e.g., YouTube and Flickr), que suportam compartilhamento de arquivos e outros conteúdos;
- Mundos virtuais de jogos (e.g., World of Warcraft);
- Mundos virtuais sociais (e.g., Second Life);
- Fóruns de discussão online;
- Sites de crítica e avaliação por consumidores (e.g., TripAdvisor e RateAdrug);
- Sites de perguntas e respostas (e.g., Yahoo! Respostas);
- Sites de gerenciamento de favoritos e tags (e.g., CiteULike e Delicious), podendo ser colaborativo;
- Sites de vendas e/ou leilões online (e.g. eBay e Mercado Livre);
- Serviços de comunicação por texto (e.g. e-mail e mensagens instantâneas);
- Serviços de comunicação por voz e vídeo (e.g. Skype).

Buscamos compreender, então, como os usuários da mídia social Twitter utilizam a plataforma para obter ou compartilhar informações relacionadas ao tema Direito Autoral e temas relacionados, bem como, quais são as falhas de entendimento dentre as diversas vertentes e especificidades do direito autoral brasileiro, especialmente em meios digitais.

A população, de modo geral, apoderou-se do potencial de criar, consumir e compartilhar conhecimento a partir da interação com as tecnologias de informação e comunicação (TICs). Condição que acarretou numa revolução cultural possibilitando que qualquer pessoa na Internet possa distribuir seus trabalhos, formais ou criativos, sem a necessidade de adaptação para as grandes mídias ou o intermédio das indústrias culturais. E sobretudo, passou a ser possível imaginar arquivos que concentrariam toda a cultura

produzida e distribuída publicamente (BRANCO, 2011; LESSIG, 2005). Contudo, ao lado das novas possibilidades de criação e divulgação em ambientes digitais, estão responsabilidades sobre o uso e disseminação consciente e responsável destas obras, tendo em vista que a legislação de Direito Autoral permanece sem uma adaptação conveniente para tais ambientes.

Como elucidado por Sérgio Branco, o direito autoral e o domínio público estão relacionados com

> todos os ramos clássicos do direito civil. Os direitos de personalidade (por conta do direito moral do autor), o direito de propriedade (na discussão acerca da natureza do direito autoral), os negócios jurídicos (em razão dos direitos decorrentes da exploração econômica das obras), os laços familiares e os direitos sucessórios (ao tratarmos dos direitos transmissíveis aos herdeiros) devem ser todos considerados para a perfeita compreensão do tema (BRANCO, 2011).

Tais relações indicam a complexidade e importância do assunto. Contudo, contar com o entendimento jurídico tão detalhado não é uma exigência para a utilização plena e correta de obras autorais. A lei brasileira de direitos autorais (Lei 9.610/98) determina quem é o portador dos direitos de autor (pessoa ou entidade criadora da obra), por quanto tempo este portador terá o direito de usufruir exclusivamente da obra e em que casos e condições a população pode usar desta, sendo obrigada a pedir autorização para tanto em certos caso, mas não em todos.

A preocupação aqui não é, entretanto, apenas sobre o entendimento da lei e suas nuances, mas sobre os efeitos sociais dela, esclarecendo como a proteção do criador e a divulgação da cultura e do conhecimento podem entrar em conflito e como o texto da lei busca o equilíbrio nessa disputa (ULLMANN, 2015).

O conceito de Web 2.0 encabeça as alterações nas noções de autor, autoria e propriedade intelectual que acontecem no início do século XXI, pois é neste momento que surgem e consolidam-se os recursos que viabilizam e promovem o compartilhamento de conteúdos nos ambientes virtuais entre criadores e consumidores. Como exemplos de plataformas colaborativas temos: o Flickr, o Youtube, a Wikipedia, o Google Earth e o Soundcloud. Foi nelas, como Lemos (2005) observa, que a prática de remixagem foi concebida e difundida entre os usuários, criando uma transformação cultural, de modo que a audiência deixa de existir no sentido literal da palavra, mas passa a incluir sua participação na

própria ação de ouvir ou apreciar outras obras. Deste modo, é possível perceber que a noção de copyright, ou direito autoral, que existia apenas de modo abstrato e indireto na vida cotidiana do cidadão comum, passou a ser algo recorrente e relevante quando, assim como Bailey descreve, "o acesso e o uso de informações e ideias expressadas por outros funcionam como blocos para a construção de criações e expressões futuras, convertendo os espectadores de hoje nos criadores de amanhã" (BAILEY, 2005).

Tendo em vista a população que, seja por puro interesse ou necessidade, procura informar-se sobre direito autoral, será necessário abranger e especificar termos que podem causar dúvidas durante uma pesquisa, como *copyright*, licenças de divulgação, acesso aberto e Creative Commons. Todos representam conceitos distintos, mas estão relacionados com o assunto geral dos direitos autorais e, portanto, também são de interesse de alguns destes usuários.

A importância do acesso aberto e das licenças Creative Commons já foi discutida em outros trabalhos, mas normalmente com o foco nos usuários que são pesquisadores acadêmicos e poderiam utilizar este conhecimento no momento de publicar suas pesquisas a fim de alcançar maior abrangência (BAPTISTA et al., 2007; BJÖRK; SOLOMON, 2012; FROSIO, 2014), ou criando comparações entre os métodos de divulgação clássicos e as novas revistas ou repositórios de acesso aberto (BJÖRK, 2004; BOMFÁ et al., 2008; HARNAD et al., 2008; NASSI-CALÒ, 2016), ou também em análises específicas das publicações brasileiras (JACOB, 2015; MACÊDO et al., 2014).

Nas últimas décadas, foram desenvolvidos diversos estudos que tratam de variados aspectos do comportamento informacional envolvendo fatores socioeconômicos, culturais, tecnológicos e cognitivos que podem favorecer o acesso à informação ou criar barreiras e incertezas (ADAMS, 2009; AGOSTO; HUGHES-HASSELL, 2005; DAVENPORT, 2010; DAVIES; BATH, 2002). Muitos destes estudos geraram modelos e teorias sobre o comportamento informacional e suas nuances, e um destes modelos foi apresentado por Wilson em 1981 no artigo "*On User Studies and Information Needs*" que, mesmo quase quatro décadas depois, ainda possui uma abordagem relevante e que detêm uma influência grande nos pesquisadores do campo, até hoje. De acordo com tal modelo, pessoas com necessidades informacionais buscam saná-las utilizando sistemas de informação convencionais, como bibliotecas e bases de dados, ou outras fontes humanas ou institucionais, e este processo continua até que achem o que estão procurando ou decidam parar a busca

(WILSON, 1981). Uma consideração relevante a se fazer é que a sociedade atual, envolvida por ambientes digitais, acessa as duas possibilidades de Wilson através do mesmo meio: a Web. É uma situação distinta da descrita por Wilson, pois através da Internet, é possível acessar as fontes de informação em páginas de bibliotecas tradicionais, bases de dados, bibliotecas digitais, repositórios, páginas de instituições e mídias sociais. Gerando, afinal, uma adaptação deste modelo por Chowdhury e Chowdhury (2011), na qual atestam que as opções de "outras fontes" descritas por Wilson tornaram-se infinitas nos ambientes digitais.

Fica clara, portanto, a abrangência do campo do Comportamento Informacional na CI e é possível comprová-la através de uma busca na plataforma *Web of Science* por termos relacionados ao Comportamento Informacional. Na primeira busca, utilizando os termos: "information behavio*" OR "information seek*" OR "information search*" OR "information use", recuperamos 11.036 documentos, sendo 3.416 específicos da Ciência da Informação e Biblioteconomia, que datam desde a década de 50 e sempre com um crescente número de publicações até o ano de 2015. Contudo, se refinarmos a pesquisa para documentos que envolvam mídias sociais em sua discussão, com os termos: "social media" OR "social net*" OR "facebook" OR "twitter". Então, o número de resultados diminui para 620 documentos, sendo apenas 170 da CI, mas que também apresentam crescente número de publicações a partir do ano 2000. Pode-se observar que a popularização da Internet nos anos 2000 é, obviamente, fator importante nestes números, acompanhado da criação das mídias sociais: o Myspace (2003), o Orkut (2004), o Facebook (2004), o Twitter (2006), entre outras.

Focaremos no Twitter, uma plataforma de interação social em ambiente virtual, que é a mais popular plataforma para o que ficou conhecido como *microblogging* contando com 383 milhões de usuários em 2012, sendo que 33 milhões são contas do Brasil, o segundo país com mais usuários cadastrados (SEMIOCAST, 2012), e um fluxo contínuo e imenso de mensagens. Há pesquisas que visam compreender quais os objetivos dos usuários do Twitter quando usam a plataforma, que passam por educação, pesquisas científicas, discussões políticas, de saúde, culturais, sociais, entre outras (EBNER; SCHIEFNER, 2008; ELSWEILER; HARVEY, 2015; SAVOLAINEN, 2011).

O Twitter, como abordado por Ferreira e Araújo (2015) e por Recuero e Zago (2010), é uma plataforma de publicações dinâmicas e interativas, um local de discussões num ambiente colaborativo, com a presença constante de perguntas e respostas, fornecendo também um impacto social significativo em relação à qualidade e ao conteúdo veiculado.

Sendo assim, perguntamo-nos Quais os meios utilizados pelos usuários do Twitter para obter ou compartilhar informações relacionadas ao Direito Autoral e suas vertentes dentro da plataforma?

Para averiguar se tais aspectos de utilização da plataforma realmente se concretizam, este estudo buscou explorar os dados textuais extraídos do Twitter a fim de identificar quais os meios utilizados pelos usuários para obter ou compartilhar informações, especificamente as relacionadas à temática do direito autoral. Buscamos investigar o comportamento informacional dos usuários do Twitter mediante a captação e análise de *tweets* sobre o direito autoral. Esta análise foi feita no software estatístico R com a utilização de métodos de Mineração Textual adaptados ao conjunto de dados obtidos. Para tanto, os seguintes objetivos específicos foram traçados:

- Consultar a literatura científica atual sobre o comportamento informacional, em especial para necessidades informacionais em redes ou mídias sociais e análise de textos com Mineração textual;
- Identificar conceitos teóricos sobre pesquisas em mídias sociais, suas ferramentas e metodologias empregadas. Bem como, os métodos de captação de dados e análise utilizando técnicas de Mineração Textual;
- Desenvolver um procedimento experimental para captar *tweets*, com o auxílio do API do Twitter, relacionados com direito autoral e suas vertentes;
- Contribuir para a compreensão acerca do Comportamento Informacional de usuários do Twitter, em especial, quando relacionados ao Direito Autoral;
- Contribuir para a disseminação do uso de ferramentas computacionais em pesquisas da Ciência da Informação, sobretudo em Mídias Sociais;
- Analisar as potencialidades e limitações do procedimento proposto e sua aplicabilidade em outras áreas da Ciência da Informação ou áreas correlatas.

A estrutura deste trabalho está organizada da seguinte forma: a primeira (presente) seção é composta pela Introdução, apresentando a justificativa do tema, definição do problema de pesquisa, os objetivos e a estrutura do trabalho. Na segunda seção será apresentada a Revisão de Literatura, tendo como temas o Comportamento Informacional, o Twitter e a Mineração de Textos, método escolhido para esta análise. Na terceira seção deste estudo, a Metodologia detalhará a caracterização e o universo da pesquisa, o instrumento de coleta de dados e as delimitações da pesquisa. Os Resultados e Discussão serão descritos na

quarta seção. Em seguida, serão tecidas as Conclusões, seguidas das Referências que serviram de base para esta pesquisa.

2 REVISÃO DE LITERATURA

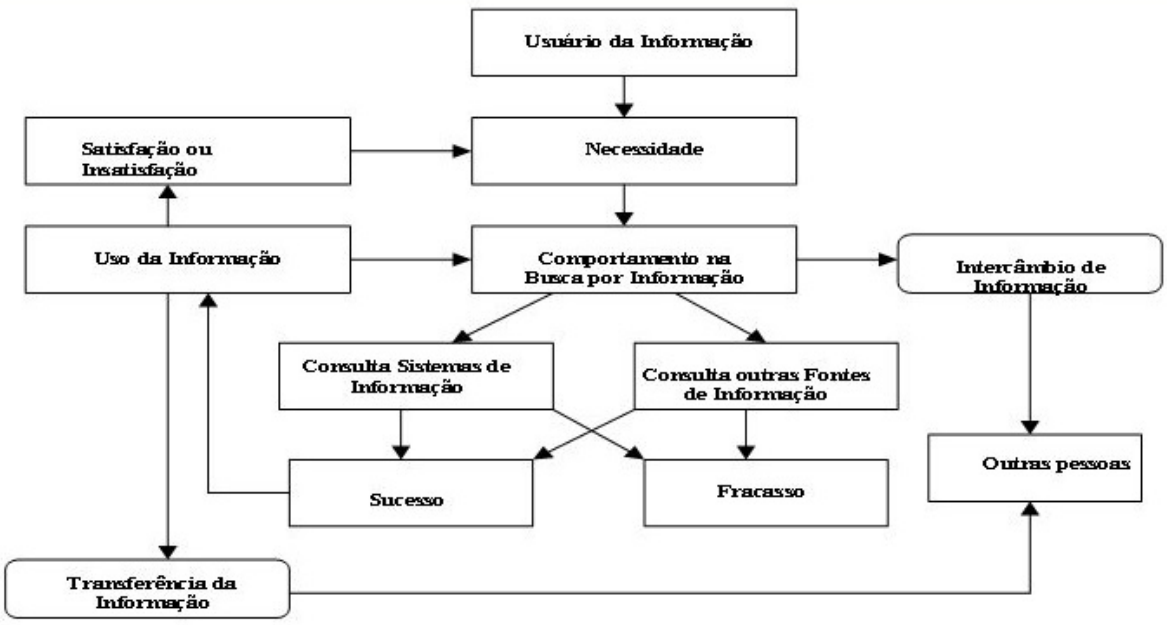
Neste capítulo serão abordados os conceitos e temas que norteiam esta pesquisa, a fim de que se possa ter um entendimento mais amplo sobre quais são tais conceitos, como e porquê foram escolhidos e de que modo estão relacionados.

Em primeiro plano tem-se o Comportamento Informacional, tema recorrente da Ciência da Informação, e que engloba os principais conceitos que serão utilizados nesta pesquisa. A seguir, o Twitter, mídia social escolhida como fonte de dados. A Mineração Textual, processo escolhido como método de análise dos dados, pois trata especificamente de informações textuais e em grandes escalas, será apresentada em seguida. E, finalmente, será abordado o software R, ferramenta estatística utilizada para uma vasta gama de análises estatísticas, dentre elas, as de mineração textual.

2.1 COMPORTAMENTO INFORMACIONAL

Chowdhury e Chowdhury (2011) afirmam que um usuário, ao iniciar um processo de busca por informações, realiza a maior parte de suas interações, seja com sistemas informacionais ou com pessoas, por meio da Web e das mídias sociais. Isso diverge consideravelmente da abordagem original de Wilson (1981), disponível na Figura 1.

Figura 1 – Modelo de Wilson (1981)



Fonte: adaptado de Wilson (1981)

Desde então, a internet tornou-se o principal canal para acessar diversos recursos informacionais, como bibliotecas clássicas ou digitais, bases de dados, conteúdos pessoais e mídias sociais. Desse modo, "as opções de 'outras fontes de informação' mencionadas no modelo de Wilson são, agora, incontáveis" (CHOWDHURY; CHOWDHURY, 2011). Uma das grandes diferenças na necessidade informacional que surgiu com o advento da internet foi a possibilidade de instituições e empresas utilizarem o mesmo recurso que seus clientes para conseguir informações sobre eles. No momento em que passou a ser natural que cada empresa possuísse um site próprio, as informações sobre os clientes que acessam este site ganham importância para a empresa, caracterizando uma necessidade informacional da empresa como um todo.

Vale observar que o próprio Wilson já adaptou seu modelo de Comportamento Informacional em 1999 para adicionar "variáveis intermediárias" ao processo de busca (WILSON, 1999). Tais variáveis seriam derivadas da necessidade primária do usuário, que vem de um problema ou tarefa e originou o processo de busca, podendo ser relacionada com a pessoa que realiza a busca, seu trabalho ou ambiente (político, social, econômico etc.) em que vive. E em 2008, novamente, quando foca na Teoria da Atividade (*Activity Theory*), original da Psicologia, mas que pode ser aplicada à Ciência da Informação, especificamente em desenvolvimento de sistemas informacionais, interação entre humanos e computadores, letramento informacional, dentre outras possíveis áreas de aplicação (WILSON, 2008).

O modelo original de Wilson (1981) e adaptado por Chowdhury e Chowdhury (2011) também toca, mesmo que indiretamente, num outro fator importante: a conveniência durante a busca por informações. As pessoas têm preferências pelas fontes de informação que conhecem e utilizam com frequência, e desde Wilson (1981), foi evidenciada a preferência por fontes humanas em conversas diretas, sobre fontes "clássicas". Na sociedade atual, passou a ser possível ter acesso à opinião e discutir com diversas pessoas, com as quais exista um relacionamento pessoal direto ou não, por meio das mídias sociais. O fator da conveniência é amplamente discutido no contexto do comportamento informacional por Connaway, Dickey e Radford (2011) e nas mídias sociais especificamente por Lee e Oh (2013), que descrevem que diferentemente das mídias de notícias, por exemplo, disponibilizam o mesmo conteúdo em massa para sua audiência, contudo, as mensagens que indivíduos veem no Twitter são determinadas por suas próprias conexões dentro da plataforma, que provavelmente refletem

seus interesses pessoais e motivações. Conveniência é um critério nas escolhas e ações das pessoas durante o processo de busca por informações, e um dos quatro principais fatores no momento de escolha de fontes de informação, ao lado de qualidade do recurso, custo e confiabilidade. Pode ser aferida pela facilidade de uso de uma ferramenta, satisfação dos usuários e o tempo gasto para acessá-la (CONNAWAY; DICKEY; RADFORD, 2011).

Outro aspecto considerado como aspecto do comportamento informacional envolve as atividades passivas e comportamentos não intencionais que podem trazer informações relevantes, como encontrar informações inesperadas ou de relance (*glimpsing* ou *encountering information*) (ERDELEZ, 1999), assim como, comportamentos opostos à busca, quando o usuário evita ativamente encontrar a informação (*information avoidance*) (CASE, 2007).

2.2 COMPORTAMENTO INFORMACIONAL NO TWITTER

Apesar da escassez de pesquisas que versem sobre comportamento informacional em mídias sociais, sua relação é inquestionável. As mídias sociais estão removendo as fronteiras entre buscar informações e socializar-se, e fazendo com que seus usuários consultem também as mídias sociais, além dos meios de busca tradicionais (MATTHEWS, 2008; MUTULA, 2010; SAVOLAINEN, 2008). Porém, encontrar informações relevantes nas mídias sociais pode ser uma tarefa complexa, pois a informação fica difusa, pode aparecer em muitas formas e estar disponível por diferentes meios. Deste modo, procurar e avaliar informações apropriadamente tornou-se uma importante competência informacional (CHOWDHURY; CHOWDHURY, 2011).

Pesquisas anteriores mostram que os usuários do Twitter falam sobre suas leituras, sobre o que pensam e compartilham links relacionados a assuntos de seu interesse (GOECKS; MYNATT, 2004), mas além de ser uma plataforma para divulgar ideias e opiniões, o Twitter também representa uma fonte de informação, criada pelos próprios usuários e suas experiências. Mostram, ainda, que muitos usuários fazem perguntas diretas a seus seguidores (outros usuários que escolhem receber as postagens deste usuário) na esperança de que possam respondê-la, ou compartilhar informações que colaborem com a descoberta da resposta (MORRIS; TEEVAN; PANOVICH, 2010).

A Twitter Inc. fornece acesso às suas informações, além de pela interface dos usuários, também por meio de um API (*application programing interface*) no qual é possível buscar e filtrar *tweets* utilizando palavras-chave, operadores de busca e outros filtros. Um *tweet* é o

nome dado à mensagem postada pelos usuários do Twitter. Até novembro de 2017, os *tweets* possuíam um limite de até 140 caracteres; a partir de dezembro de 2017, este limite foi dobrado. Algo interessante a se considerar é que o aumento não modificou muito os modos de uso da ferramenta: apenas 9% dos *tweets* chegaram até os 140 caracteres e, agora, apenas 1% dos *tweets* chega a 280 caracteres (NEWTON, 2017). Por outro lado, a empresa tomou a decisão com o objetivo de aumentar o fluxo de postagens na plataforma, pois a hipótese era de que alguns usuários talvez deixassem de postar no Twitter por receio de não conseguir expressar o que desejam em apenas 140 caracteres. Logo, com o aumento do limite, espera-se que mais usuários sejam atraídos a utilizar a plataforma.

2.3 MINERAÇÃO DE DADOS

um conjunto de métodos e ferramentas, conhecido como *Knowledge Discovery in Databases* (KDD) ou "descoberta de conhecimento em bases de dados", foi desenvolvido com o objetivo de identificar padrões e relações que seriam complexos de perceber em leituras comuns e que podem fornecer conhecimento e informações úteis, de acordo com sua aplicação (FELDMAN; DAGAN, 1995). O processo de descoberta de conhecimento investiga e cria métodos, algoritmos e mecânicas para recuperar conhecimento potencialmente relevante de conjuntos de dados ou textos (NORTON, 1999).

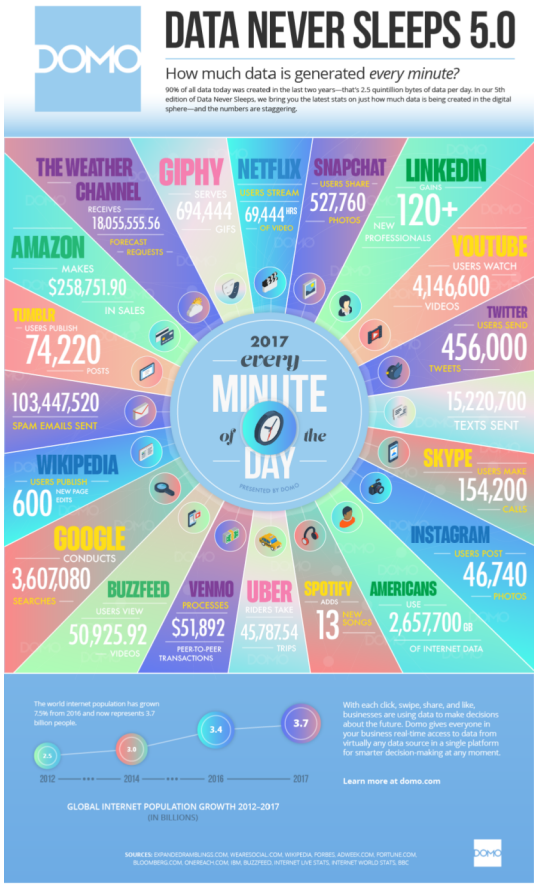
Um dos principais objetivos é descrever e diferenciar as possibilidades de uso destas técnicas de descoberta de conhecimento, tendo em foco as bases de dados textuais e as perspectivas de uso na Ciência da Informação. Como descrito por Fayyad et al. (1996), o processo KDD já recebeu uma variedade de nomes, como extração de conhecimento, descoberta de informações, arqueologia de dados, processamento de padrões de dados e mineração de dados (*data mining*). Contudo, a mineração de dados também pode ser vista como uma etapa particular do processo geral de descoberta de conhecimento e, quando aplicada sem os devidos cuidados, como a preparação, seleção e limpeza dos dados ou sem a interpretação apropriada dos resultados, pode acarretar em descobertas falsas ou irrelevantes.

Uma quantidade massiva de dados é criada e distribuída em diferentes mídias sociais populares entre públicos distintos, no Twitter, por exemplo, são 456 mil *tweets* enviados por minuto, como podemos ver na Figura 2. Tal fenômeno tem sido chamada de *Big Data,* sendo visto responsável por uma revolução em como empresas, governos e organizações coletam e analisam dados no momento de tomada de decisão, bem como mudanças no campo científico (CONEGLIAN; SANTAREM SEGUNDO; SANT'ANA, 2017), que acarretaram no

desenvolvimento de diversas novas técnicas e tecnologias para análise destes dados.

A fonte de objetos de estudo que utilizamos aqui são as mídias sociais, uma categoria de serviços da Web que atraiu milhões de usuários nos últimos anos, tornaram-se fontes de recursos interessantes para pesquisas sociocomportamentais (REIPS; GARAIZAR, 2011). São nelas que os usuários compartilham experiências, sentimentos, informações, imagens e vídeos e onde criam a maior parte desses novos documentos.

Figura 2 – Dados criados em serviços disponíveis na Internet a cada minuto



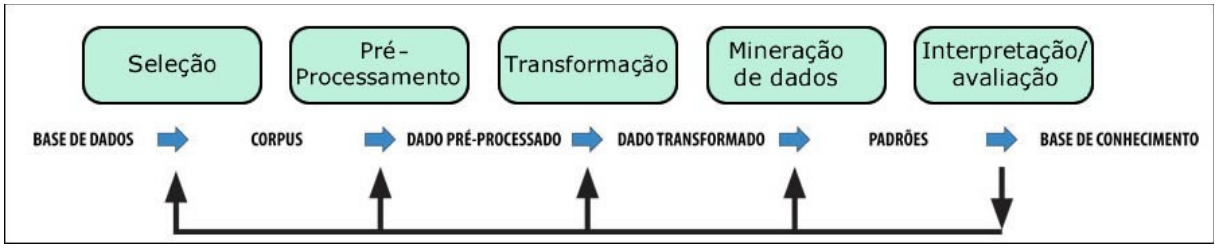
Fonte: James (2017)

Considerando, então, as bases de dados e repositórios científicos como os objetos clássicos de estudo da CI, pode-se dizer que a descoberta de conhecimento já é parte integrante das pesquisas em CI, mas os novos métodos, como mineração de dados, de textos e automatização, contudo, ainda estão em seus primeiros passos.

É importante ter em mente que a mineração de dados é apenas uma etapa do processo de descoberta de conhecimento em base de dados e que, apesar de plenamente baseada em tecnologia, depende em última instância da avaliação e direção de agentes humanos para que

os padrões e relações descobertos estejam além da simples contagem e discriminação dos dados (NORTON, 1999). O processo de descoberta de conhecimento é dividido em cinco etapas por Fayyad et al. (1996), são elas: seleção, pré-processamento, transformação, mineração de dados e interpretação/avaliação do conhecimento extraído. Estas fases e suas interações podem ser vistas de maneira geral na Figura 3.

Figura 3 – Etapas do processo de descoberta de conhecimento em bases de dados

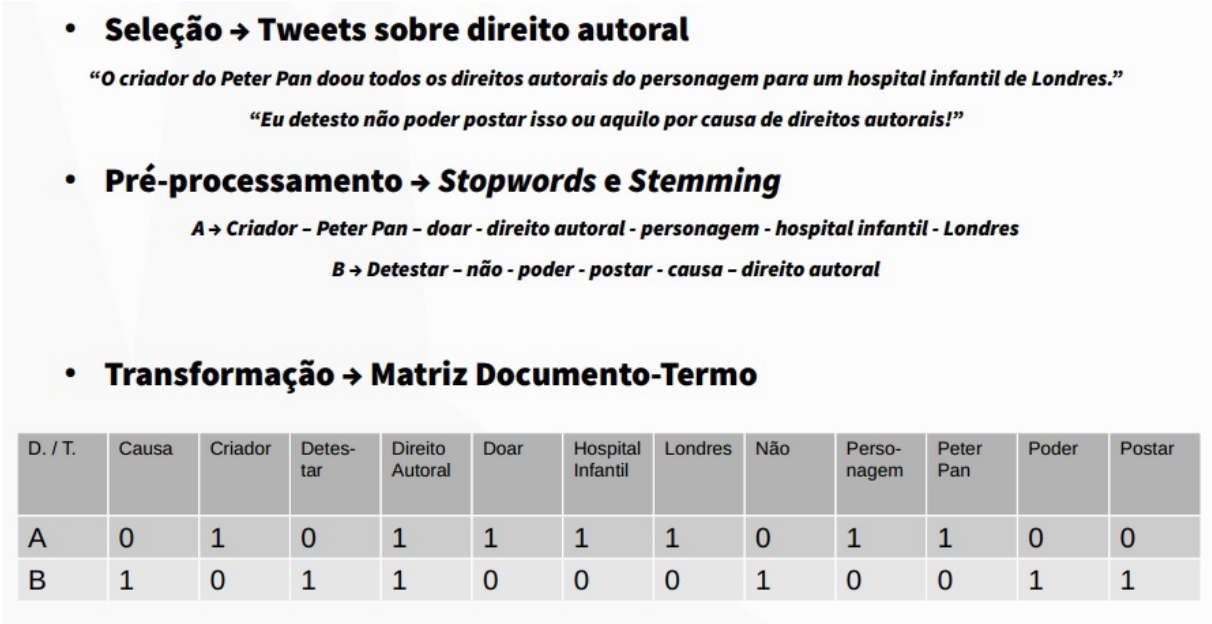


Fonte: Fayyad et al. (1996, p. 41)

Para melhor compreensão, vamos descrever as etapas ilustradas na Figura 2. Primeiro vemos a Seleção, etapa onde o analista decide que base ou conjunto de dados utilizará, para a qual é preciso ter em mente qual será o objetivo final da análise e se algum conhecimento é necessário para lidar com os dados brutos. A seguir, o Pré-processamento, também conhecido como limpeza, etapa em que é necessário observar se os dados estão completos, como será possível lidar com os campos que faltam, ou com o ruído presente, e removê-los, se for apropriado. A Transformação, em seguida, envolve achar aspectos que possam representar os dados, para que seja possível diminuir o número de variáveis sob consideração na análise, a fim de realizá-la efetivamente. A Mineração de Dados acontece a seguir, em que são escolhidos os métodos e algoritmos que serão utilizados na análise, bem como o modo ideal de aplicá-los no conjunto de dados em questão, para que os padrões de interesse sejam extraídos. E, finalmente, a interpretação dos padrões adquiridos na etapa anterior, é o passo que pode envolver um retorno a qualquer das etapas anteriores num processo iterativo, bem como desenvolver visualizações que facilitem o entendimento dos padrões ou modelos extraídos (FAYYAD et al., 1996)

Adicionamos mais uma descrição das etapas de Seleção, Pré-processamento e Transformação com exemplos de textos que foram capturados nesta pesquisa, na Figura 4.

Figura 4 – Etapas com exemplos



D. / T.	Causa	Criador	Detes-tar	Direito Autoral	Doar	Hospital Infantil	Londres	Não	Perso-nagem	Peter Pan	Poder	Postar
A	0	1	0	1	1	1	1	0	1	1	0	0
B	1	0	1	1	0	0	0	1	0	0	1	1

Fonte: elaborado pelo autor

Assim, ao considerarmos novamente a enorme quantidade de documentos que são criados e armazenados em ambientes Web, devemos atentar que tais documentos muitas vezes não possuem classificações mínimas ou sequer metadados relacionados. Para estas situações, foi desenvolvido o processo de mineração de textos (*text mining*). Neste processo, que envolve diversas técnicas desenvolvidas e refinadas nas últimas décadas, a análise automatizada possibilita a descoberta e criação de padrões que só seriam identificáveis por meio da leitura exaustiva de uma enorme quantidade de documentos. Cabe ressaltar que o funcionamento da mineração de textos é equivalente à mineração de dados. As principais diferenças ficam nas etapas de seleção e pré-processamento, pois são nestas fases que os dados originais (em linguagem natural no caso da mineração de textos) são tratados a fim de que as ferramentas computacionais possam atuar (UPSHALL, 2014).

2.4 MINERAÇÃO DE TEXTOS

Para compreender melhor o funcionamento do processo de mineração de textos, esta seção descreve alguns dos componentes mais comuns utilizados para extração de conhecimento de bases de dados de documentos textuais. Esta não é uma listagem exaustiva e não pretende dar uma descrição completa sobre o funcionamento de cada componente, mas apenas exemplificar o funcionamento do processo e a complexidade envolvida, especialmente levando em consideração que a combinação destes componentes pode trazer resultados distintos e mesmo inesperados.

- **Identificar Estrutura** – é importante identificar a estrutura textual dos documentos a serem analisados e reconhecer qualquer metadado que possa estar associado ao documento. Segundo Zhang e Gu (2011), 90% dos dados disponíveis na Web estão em formatos não estruturados Desse modo, e motivados por perspectivas de problemas específicos do mundo, diversos algorítimos e sistemas de gerenciamento de conhecimento foram desenvolvidos. A data de publicação, ou o nome do autor são exemplos de metadados que, mesmo se não descritos especificamente, podem ser descobertos a fim de melhorar a análise.
- **Seleção do texto** – nem todas as palavras do texto são de suma importância para a análise. Como grande parte dos documentos disponíveis na Web estão escritos em linguagem natural, muitas vezes é vantajoso para a mineração de textos levar em consideração apenas as palavras mais relevantes que compõe o texto, removendo as palavras irrelevantes (*stopwords*), como artigos e verbos, de acordo com o tipo de documento e com o objetivo final da análise.
- **Combinação de elementos** – momento em que as palavras de mesmo radical são agrupadas para que contem como uma única unidade semântica durante a análise. É importante perceber que palavras diferentes podem ter a mesma significação no corpo do texto. Para melhorar o resultado da análise é necessário definir esses sinônimos, seja antes da análise, dependendo do nível de conhecimento prévio dos assuntos tratados nos documentos, ou iterativamente durante a análise, conforme as relações sejam reconhecidas pelo analista.
- **Coocorrência** – a ocorrência de duas ou mais palavras específicas juntas pode ter sentido diferente do que se ambas fossem consideradas como não relacionadas e modificar o entendimento do texto. É possível comparar a ocorrência e a proximidade de palavras durante a análise, de modo que, normalmente, palavras próximas têm mais probabilidade de estarem relacionadas do que palavras distantes.
- **Análise de sentimento** – uma análise desenvolvida especialmente para documentos publicados em massa nas mídias sociais, normalmente depoimentos ou descrições de situações vividas ou produtos e serviços que foram utilizados pelo autor. Pela contagem de palavras “positivas”, “negativas” ou “neutras” é possível determinar a reação do autor em relação ao produto/serviço. Este tipo de informação é de suma importância para empresas, no momento de tomada de decisões e na melhoria geral da

relação com seus clientes. Como discutido por Yu et al. (2012), a publicação de críticas amadoras é um modo popular de divulgar opiniões e sentimentos sobre produtos ou serviços que o usuário tenha utilizado. A relevância destes comentários vai depender da quantidade de seguidores da pessoa que fez a crítica, bem como da quantidade de críticas sobre tal produto ou serviço e, sendo assim, é de suma importância para as empresas terem conhecimento sobre tais postagens, além de uma comunicação direta com os clientes.

- **Indexação automática** – em documentos científicos, é possível utilizar métodos de mineração textual para determinar os assuntos do texto. Há controvérsias sobre a qualidade da indexação automática em comparação à feita por classificadores humanos mas, segundo Ribeiro (2009), é evidente que se utilizada em conjunto com os vocabulários controlados e com a mediação de um agente, a indexação automática pode render bons resultados.

2.5 MINERAÇÃO DE TEXTOS E SUAS APLICAÇÕES

Para a *Royal Society of Science* em seu relatório "*Science as an open enterprise*", a mera divulgação de dados científicos não garante que sejam, de fato, dados ou documentos abertos. Para tanto, é necessário que os dados sejam acessíveis, ou seja, disponíveis de maneira que possam ser acessados facilmente; inteligíveis, isto é, divulgados numa linguagem clara; avaliáveis, todos os interessados poderiam julgar e avaliar estes dados e; usáveis, ou que possam ser reutilizados, frequentemente para outros propósitos (BOULTON et al., 2012).

Desse modo, consideramos nesta discussão que a mineração textual é uma ferramenta apta a auxiliar na avaliação e reutilização de dados ou documentos científicos, facilitando o reconhecimento de padrões e a criação de hipóteses para justificá-los e compreendê-los.

Witten e Frank (2005) caracterizam a mineração de dados como a extração de informações potencialmente úteis de um conjunto de dados, onde estavam implícitas e permaneceriam desconhecidas. No *Text Mining*, contudo, a informação que será extraída está explícita no texto. Ainda assim, ela pode ser considerada desconhecida devido ao tempo que seria necessário para ler todos os textos e, deste modo, o *Text Mining* facilitaria este processo de descoberta. Trazemos, a seguir, as etapas do processo de *Text Mining*, segundo a divisão de Dixon (1997):

- Recuperação da Informação – isto é, a localização dos documentos relevantes e filtragem, para garantir que contenham as informações de interesse;
- Extração da Informação – momento no qual certifica-se de que os dados textuais tenham sido corretamente estruturados e que palavras irrelevantes (*stopwords*) tenham sido retiradas;
- Mineração da Informação – uma vez que os textos selecionados já estejam transformados em dados estruturados, inicia-se a etapa de mineração a fim de descobrir ou encontrar padrões nos textos;
- Interpretação da Informação – a última etapa é a interpretação dos padrões descobertos na etapa de mineração.

Apresentaremos alguns exemplos de aplicação das ferramentas de mineração de textos, de modo a demonstrar quão útil pode ser a utilização de textos em análises estatísticas e identificar a diversidade de seu uso, seja em Ciência da Informação, outros campos científicos ou em empresas.

2.5.1 Estudo de caso: mineração textual em bibliotecas acadêmicas

Para as organizações de pesquisa, o conhecimento tornou-se um importante recurso. Zhang e Gu (2011) consideram que a geração, codificação, gerenciamento e compartilhamento de conhecimento são essenciais para o processo de inovação que acontece nestas organizações. Sendo assim, a mineração textual tem muito potencial para auxiliar as bibliotecas e profissionais da informação que almejam inovar seus processos e funções.

A automatização de processos como classificação e indexação pode ser feita com o reconhecimento de termos (*term recognition*), agrupamento de documentos (*document clustering*) e sumarização automática (*automatic summarization*), todas técnicas utilizadas em mineração textual. Também podem ser usados no gerenciamento de pessoas, a fim de identificar os usuários da biblioteca e suas expectativas e reações sobre os serviços oferecidos e padrões de uso do acervo, melhorando a relação entre o usuário e a biblioteca.

As bibliotecas acadêmicas muitas vezes são parte de organizações de pesquisa, como universidades e laboratórios, mas podem também agir com a autonomia de uma organização, e as ferramentas de mineração textual podem ser utilizadas para alcançar os patamares de inovação e desenvolvimento que são esperados de organizações de pesquisa atuais.

2.5.2 Estudo de caso: problemas de saúde em site de perguntas e respostas

O site Yahoo Respostas é um dos mais frequentados sites de perguntas e respostas, onde os usuários podem interagir uns com os outros a fim de perguntar ou responder perguntas sobre diversos assuntos, umas das áreas com mais postagens é a de saúde. Oh, Zhang e Park (2016) coletaram mais de 80.000 perguntas feitas entre 2009 e 2014 na plataforma, todas da categoria de "câncer."

A análise resultou em seis aspectos mais comuns no assunto das perguntas feitas, são eles: aspectos demográficos dos autores, como idade e sexo; aspectos cognitivos, como tratamentos ou sintomas específicos; aspectos afetivos, como sentimentos positivos e negativos; aspectos situacionais, como fatores de exposição e dietas; aspectos sociais, como relacionamentos com amigos e familiares e; aspectos técnicos, como fontes de informação.

As frequências em que foram reconhecidos cada um destes aspectos nas perguntas podem determinar os principais assuntos e dúvidas deste público-alvo. Tais resultados poderiam ser utilizados no desenvolvimento de cartilhas e outros meios de divulgação especializados em portadores de câncer e, assim, garantir que as dúvidas e aspectos mais comuns dessa doença sejam abordados.

2.5.3 Estudo de caso: análise competitiva em mídias sociais da indústria de pizzas

Este estudo coletou dados de mídias sociais, como Facebook e Twitter, sobre as três maiores franquias de pizzas nos EUA: *Pizza Hut*, *Domino's Pizza* e *Papa John's Pizza*. O objetivo era comparar a utilização das mídias sociais pelos clientes de cada franquia e pelas próprias franquias durante a prestação do serviço ou o contato com os clientes. He, Zha e Li (2013) demostram que a análise competitiva pode ser feita em mídias sociais e que a mineração de textos, especialmente pelo uso da análise de sentimentos, é uma ferramenta útil para este tipo de estudo.

Para cada franquia de pizza foram encontrados cinco temas principais no Twitter: pedido e entrega, qualidade da pizza, retorno dos clientes para decisão de compra, *tweets* sociais e *marketing tweets*. Ficou claro que as três companhias empenham-se em engajar os clientes em suas mídias sociais e tiram vantagens sobre os conteúdos criados ou compartilhados por eles, além de poderem utilizar estes dados no momento de tomada de decisão sabendo de antemão os desejos e anseios de seus clientes. Ao mesmo tempo em que

há também vantagem para os clientes que passam a ter um contato mais aberto com as empresas que lhes prestam serviços.

2.5.4 Estudo de caso: determinação de assuntos e indexação automática para documentos jurisprudenciais

Samuel Ribeiro (2009) propõe neste trabalho um método que utiliza em conjunto uma análise de mineração de textos e a indexação clássica de documentos. Em sua comparação, fica claro que a mineração de textos, num primeiro momento, coleta simplesmente as palavras com mais frequência no texto, ao passo que num processo de indexação clássico existem diversos outros fatores que influenciam as decisões e o resultado final da determinação do assunto do documento.

Deste modo, a pesquisa sugere a utilização de um tesauro jurídico da área para identificar os termos importantes esperados dentro dos documentos e atribui pesos distintos aos termos preferidos, relacionados ou ausentes do tesauro. Quando a análise é refeita com a utilização destes pesos, o resultado é mais satisfatório, em comparação à indexação feita por agentes humanos.

3 PROCEDIMENTOS METODOLÓGICOS

Nesta seção, descreveremos os procedimentos metodológicos realizados no decorrer do estudo para obtenção dos dados e durante sua análise. Esta pesquisa é de natureza qualitativa e quantitativa concomitantemente, de modo que os dados serão analisados e descritos estatisticamente a fim de agrupá-los para facilitar a identificação padrões ou e tendências (WILDEMUTH, 2009). Primeiramente, nos textos coletados e, possivelmente, também nos próprios usuários que os publicaram ou compartilharam.

3.1 R – SOFTWARE ESTATÍSTICO

Segundo Blake (2011, p. 126), o processo de *Text Mining* como a identificação de padrões originais, interessantes e compreensíveis dentro de uma coleção de textos. Para descobrir tais padrões utilizamos o *software* R, escolhido por ser um *software* que contém recursos para análise de *Text Mining* e por ser gratuito. O software R é uma ferramenta estatística poderosa que possui um amplo conjunto de funções e pode ser aperfeiçoado para análises específicas com o uso de pacotes com novas funções.

O pacote “tm” é o que possui as principais funções utilizadas em *Text Mining*, e é o mais utilizado neste tipo de análise. Alguns pacotes que foram utilizados neste trabalho são: “tidytext”, “fpc”, “wordcloud”, “topicmodels” e “cluster”, que auxiliaram no pré-processamento do texto, na análise de *cluster*, modelagem por tópicos e construção de gráficos. Estes pacotes, bem como outros que podem ser relevantes, são descritos nos livros *Text Mining with R* (SILGE; ROBINSON, 2017) e *Text Mining in Practice with R* (KWARTLER, 2017) que utilizamos como base em nossas análises. Também foram realizadas consultas constantes em fóruns e comunidades online a fim de buscar respostas a dúvidas que surgiam no decorrer do processo e que não eram tratadas especificamente nos livros.

3.2 PROCESSO DE COLETA

Utilizamos o API do Twitter para obter os dados em formato de texto. Buscamos por *tweets* que falem sobre copyright e direitos autorais a fim de identificar quais vertentes deste assunto são mais discutidas e geram mais dúvidas entre os usuários da plataforma. Deste modo, compreenderíamos as lacunas de conhecimento sobre o assunto e seria possível analisar o comportamento dos usuários em face a estas questões.

Nele é possível inserir *queries* a fim de recuperar as postagens feitas de acordo com os parâmetros estabelecidos. Todos os parâmetros da busca avançada da plataforma estão disponíveis, alguns deles são: palavras-chave, *hashtags*, idioma, localização e data. Cabe destacar que utilizamos termos de busca que fazem parte do vocabulário comum e que representem elementos de possível interesse dos usuários do Twitter e que sejam abrangentes o bastante para garantir a recuperação de um número relevante de *tweets*. Os termos foram: “acesso aberto”, “copyleft”, “copyright”, “creative commons”, “direito autoral”, “direitos autorais” e “domínio público”. Também inserimos filtros para limitar os resultados apenas aos escritos em língua portuguesa e que fossem apresentados os resultados do mais recente ao mais antigo.

Foram capturados 8.367 *tweets* utilizando os termos de busca definidos anteriormente. Utilizamos filtros para captar apenas resultados em língua portuguesa e que apresentassem uma das palavras-chave escolhidas, como vemos na Tabela 1.

Tabela 1 – Quantidade de documentos recuperados por palavra-chave

Palavras-Chave	Nº de Resultados
Acesso Aberto	113
Copyleft	12
Copyright	1533
Creative Commons	78
Direito Autoral	498
Direitos Autorais	5700
Domínio Público	433

Fonte: Elaborado pelo autor.

Podemos dividir os dados coletados em grupos de acordo com o termo de busca detectado em cada um, desse modo, vemos que os termos que trouxeram maiores resultados são “domínio público” - 433, “copyright” - 1.533, “direito autoral” - 498 e “direitos autorais” - 5.700. Vistos como os termos mais comuns para tratar de assuntos relacionados ao direito autoral no vocabulário destes usuários. Os outros termos: “acesso aberto” - 133, “creative commons” - 78 e “copyleft” - 12, caracterizam conceitos mais complexos e modernos, o que pode justificar a menor quantidade de discussões sobre eles na plataforma.

3.2 PRÉ-PROCESSAMENTO

O API do Twitter recupera somente os *tweets* dos últimos sete dias, função que foi incluída devido à enorme quantidade de textos postados na plataforma. Sendo assim, caso a coleta fosse realizada em menos de sete dias, possivelmente seriam capturados *tweets* duplicados. Para removê-los basta conhecer a identificação do *tweet*, dado que o API também fornece e que é um valor único para cada *tweet*, uma sequência de números de 18 dígitos atualmente, e remover os *tweets* com o mesmo número de identificação.

Partindo dos 8.367 documentos iniciais, foi possível reconhecer que 222 possuíam o mesmo número identificador. Para isso, foi utilizado o pacote "plyr" do software R, que conta a frequência de valores repetidos num conjunto de dados. Em seguida, para removê-los, bastou adicionar um filtro ao conjunto de dados inicial que removesse as linhas identificadas como repetidas. Restando, então, 8.145 documentos diferentes. Vale ressaltar que ainda dentro desses 8.145 restantes existem *retweets*, mensagem que um usuário simplesmente copia o conteúdo de outro usuário e publica novamente. Nesse caso, o documento é relevante para a análise pois pode representar a relevância de uma mensagem que o usuário decidiu repetir, ou uma dúvida comum entre os usuários que republicaram a mensagem.

A seguir, para facilitar a análise de Mineração Textual, apenas o conteúdo dentro do documento, isto é, o texto escrito pelo usuário, será considerado, e as colunas com nomes de usuários e outros dados relacionados ao documento foram removidas, permanecendo apenas as colunas com o número identificador, a data de publicação e o texto a ser analisado. Finalmente, é necessário remover as pontuações e as *stopwords*. *Stopwords* são palavras que não contém significado relevante para a análise, como artigos, preposições, conjunções e outras palavras específicas.

Para remover a lista de *stopwords,* foi necessário modificar o conjunto de dados para uma matriz documento-termo. Isto é uma matriz onde as linhas representam todos os documentos do conjunto, isto é, os 8.145 *tweets*; as colunas são uma lista de todas as palavras presentes em todos os *tweets*; e o conteúdo da matriz são apenas valores "1" ou "-1" que simbolizam "positivo" ou "negativo", caso a palavra esteja presente naquele documento, ou não, respectivamente.

Nesse ponto foi necessário contar a quantidade de palavras em todos os documentos. O valor encontrado foi de 129.589 palavras no total, sendo 16.467 únicas. Também foi possível calcular que 28.644 palavras do total possuem apenas 2 caracteres, fato que as

caracteriza como não relevantes e removidas também.

A lista de *stopwords* que utilizamos, criada por membros do Núcleo Interinstitucional de Linguística Computacional (HARTMANN et al., 2017), possui 443 palavras em língua Portuguesa que recomenda-se remover antes de realizar análises textuais. Para removê-las é necessário filtrar a lista com todas as palavras pela lista de *stopwords*. Terminamos o processo com 77.138 palavras com 3 ou mais caracteres e que não são *stopwords*, sendo 16.111 únicas.

Finalmente, já durante a etapa de análises foram reconhecidas algumas dificuldades que poderiam ser remediadas com algumas modificações na lista de *stopwords*, bem como no texto original das mensagens coletadas. Estas modificações foram feitas iterativamente e de acordo com a necessidade e o momento em que foram reconhecidas as dificuldades. São elas:

- Adicionar os termos “tá”, “já”, “só” e “até” à lista de *stopwords;*
- Localizar e substituir os termos:
  - “yt” e “ytb” por “youtube”;
  - “tt” por “twitter”;
  - “rt” por “retweet”;
  - “msc” por “música”;
  - “pq” por “porque”;
  - “vc” por “você”;
  - “to” por “estou”.

Estes foram termos que afetavam os resultados de maneira considerável e que a substituição pôde ser feita de maneira a garantir a integridade dos dados. Para tanto, foi necessário localizar cada mensagem que possuía algum dos termos, lê-la e realizar a mudança apenas nos casos em que não havia dúvidas de que o termo presente era de fato a abreviação do novo termo.

Com os documentos preparados, é possível criar matrizes de termos, calcular a frequência das palavras, a associação entre elas, agrupá-las e determinar assuntos iterativos e modelar tópicos.

4 RESULTADOS E DISCUSSÃO

Este capítulo apresenta os experimentos realizados durante a análise dos textos. O primeiro deles é a criação de uma nuvem de palavras. A partir da lista de palavras preprocessadas, iniciou-se a etapa de mineração textual e análise dos resultados.

Existem inúmeras análises possíveis de serem feitas com dados textuais. Talvez a mais conhecida seja a Nuvem de Palavras, em que as palavras são apresentadas de acordo com a frequência que aparecerem nos documentos. As palavras que formam a nuvem são, normalmente, distribuídas aleatoriamente na imagem, mas o tamanho da fonte e a cor podem ser relacionadas à frequência ou outras características dos dados, como documento de origem, data ou sentimento.

A Figura 5 apresenta uma Nuvem de Palavras com as palavras mais frequentes dentre os documentos que recuperamos. As palavras foram distribuídas na imagem aleatoriamente, mas o tamanho da fonte utilizada representa a frequência que a palavra aparece nos textos, quanto maior a fonte, maior sua frequência de uso. Destacamos que as palavras-chave utilizadas nas buscas, e que consequentemente são as mais comuns, foram removidas a fim de facilitar a visualização dos outros termos mais frequêntes.

Figura 5 – Nuvem de Palavras

Fonte: elaborado pelo autor

De modo mais simples, também é possível identificar as palavras mais comuns a partir de uma simples contagem. Apresentamos na Tabela 2 o número de aparições de cada palavra em *tweets* distintos. É interessante reparar que apenas 27 palavras são repetidas mais do que 150 vezes, o que demostra uma grande variedade de abordagens sobre o assunto, e também o uso de um vocabulário amplo pelos usuários do Twitter.

Tabela 2 – Frequência das palavras

Rank	Palavra	Nº de Resultados
1	direitos	5525
2	autorais	5478
3	https	2844
4	t.co	2756
5	video	1866
6	retweet	1631
7	copyright	1478
8	youtube	1443
9	musica	890
10	gostei	770
11	direito	539
12	vou	518
13	autoral	499
14	dominio	440
15	publico	439
16	cobrar	301
17	conta	263
18	causa	246
19	tauz	232
20	twitter	205
21	raptributo	203
22	homem	199
23	adicionei	183
24	cade	171
25	palavra	165
26	foto	160
27	playlist	151

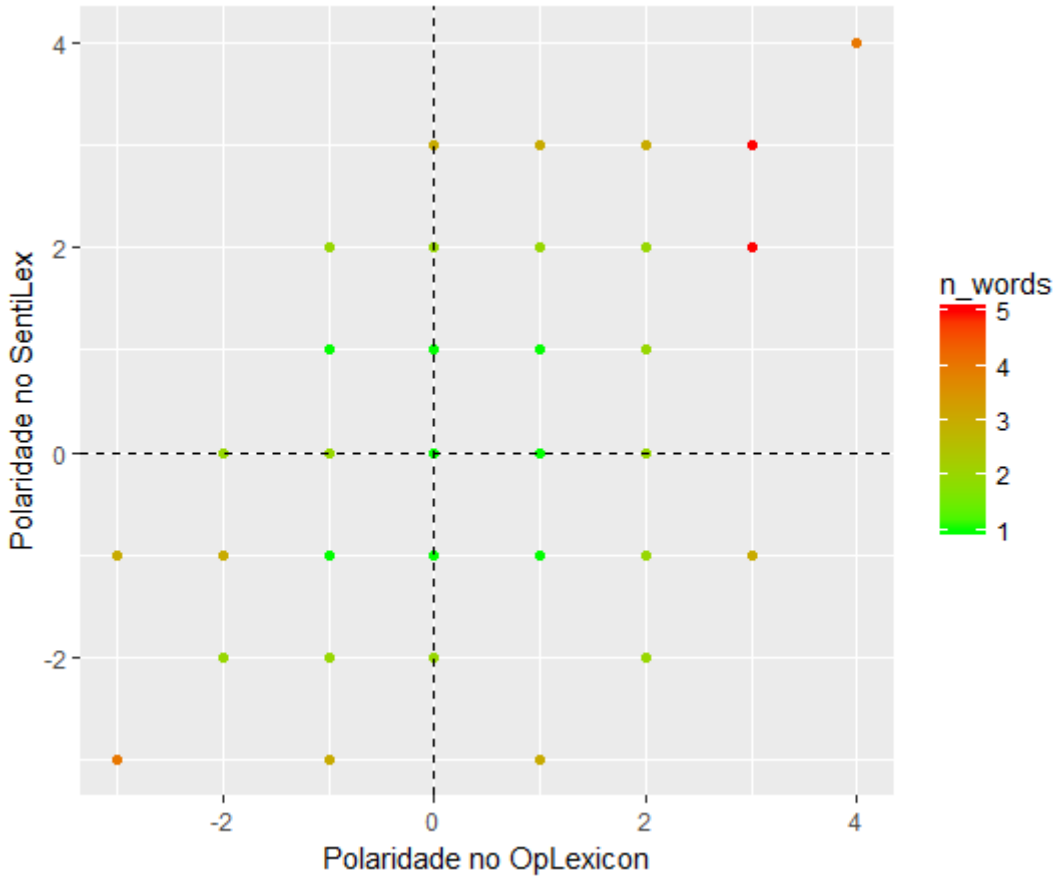
Fonte: elaborado pelo autor.

Na Análise de Sentimentos, um índice criado anteriormente é usado para classificar as palavras de acordo com o sentimento que geralmente representam quando são utilizadas em textos. Desse modo, é possível relacionar um sentimento aos documentos analisados e também aos usuários, no momento que escreveram e publicaram a mensagem. Os sentimentos são divididos de modo geral apenas em "positivos" e "negativos", em diferentes graus. Assim como em qualquer análise textual, a quantidade de documentos e a quantidade de palavras por documento são fatores importantes para a obtenção de resultados representativos.

Para tal análise, utilizamos dois léxicos de palavras, são eles: OpLexicon v3.0 (SOUZA; VIEIRA, 2012) e SentiLex-PT02 (SILVA et al., 2010). Ambos são léxicos específicos para a Língua Portuguesa que classificam as palavras em valores numéricos de acordo com o sentimento que remetem, isto é, palavras positivas têm o valor "1", neutras o valor "0" e negativas o valor "- 1". Uma maneira de determinar o sentimento de cada *tweet* seria, então, somar o valor correspondente a cada palavra utilizada e o total seria o valor do sentimento presente no *tweet* como um todo. Para essa análise e para comparação dos dois léxicos, contudo, utilizamos apenas *tweets* que possuem palavras presentes simultaneamente em ambos os léxicos; no total, então, foram considerados 2.169 *tweets*.

A fim de exemplificar as diferenças entre os léxicos, criamos um gráfico de polaridade (Figura 6). Nele estão representados todos os *tweets* analisados por ambos os léxicos numa comparação que mede a quantidade de resultados comuns entre os dois léxicos, ou seja, quantas vezes ambos os léxicos classificam o sentimento de um determinado *tweet* com o mesmo valor. O valor "n_words" representado equivale a quantidade de *tweets* classificados igualmente pelos dois léxicos.

Figura 6 – Polaridade entre os léxicos



Fonte: elaborado pelo autor.

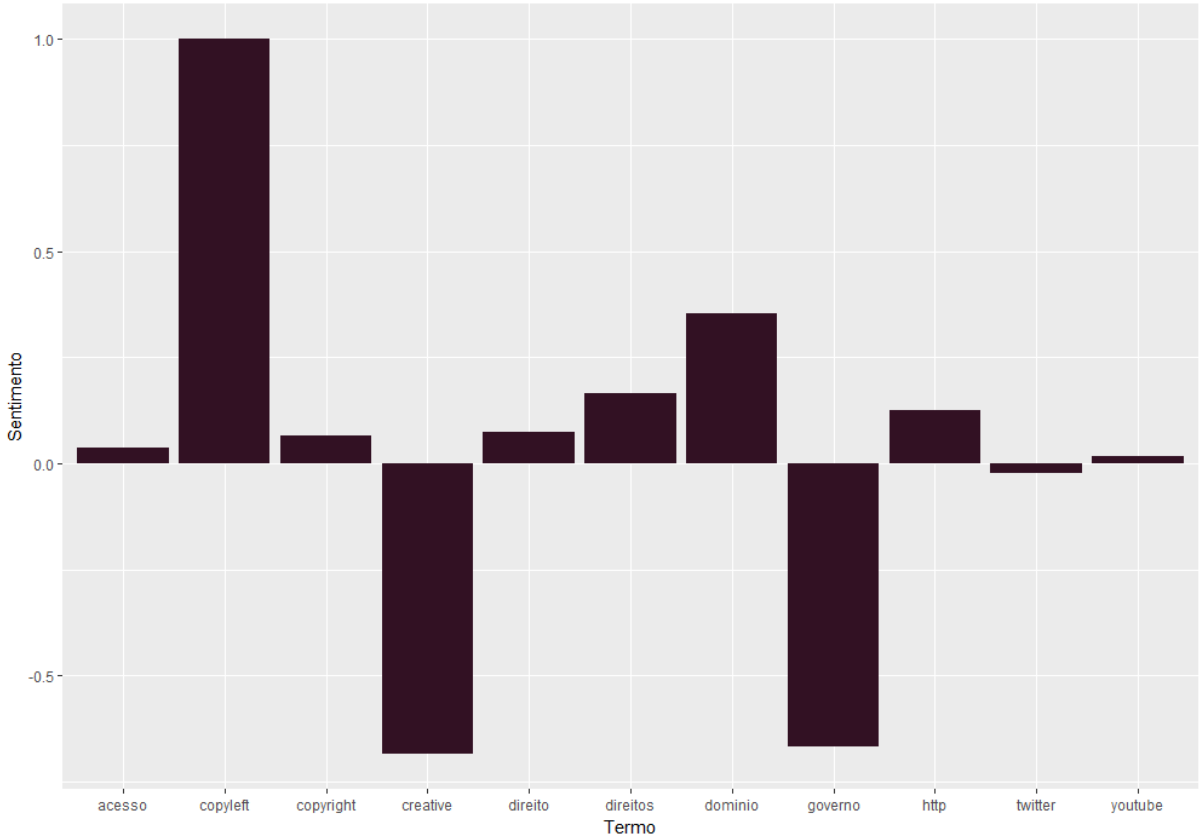
É possível perceber que os pontos com mais valores coincidentes (vermelhos e laranjas) estão nas bordas do gráfico, ao passo que, no centro ocorreram menos coincidências. Apesar das diferenças de classificação, marcantes especialmente em *tweets* classificados com valores próximos de 0 (neutros), isso demonstra que os léxicos são diferentes e podem classificar palavras de pouca carga emocional diferentemente, mas que podem trazer resultados concordantes para exemplos mais carregados emocionalmente.

Mostramos a seguir dois gráficos que apresentam a média da somatória do "valor do sentimento" das palavras de cada *tweet* utilizando um dos léxicos e, para facilitar a compreensão, dividimos as colunas de acordo com as palavras-chave presentes. Ainda, como comparação, adicionamos colunas com palavras comuns e que poderiam ter carga emocional. São elas "governo", "twitter", "youtube" e "http", sendo que esta última foi adicionada para representar os *tweets* nos quais os usuários compartilham links.

Esta é uma atitude comum entre os usuários do Twitter e é possível identificar tais links pois todos possuem o prefixo "http". Twitter e Youtube são duas das palavras mais

comuns dentre os *tweets* coletados, e plataformas muito utilizadas, assim decidimos adicioná-las na análise.

Figura 7 – Sentimento por termo (OpLexicon)



Fonte: elaborado pelo autor.

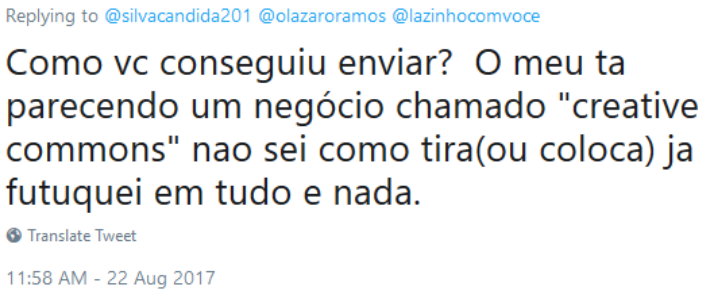
A Figura 7 foi elaborada com o léxico OpLexicon v3.0, que possui mais de 30 mil palavras e símbolos classificados. Os valores variam entre 1 e – 0,75, tais valores são a média dos valores de sentimento de cada *tweet* que contem a palavra-chave da coluna em questão. Podemos ver que para a maior parte das palavras-chave os comentários são considerados positivos, sendo que os únicos com médias negativas são “creative commons”, “governo” e “twitter”. No termo “twitter”, o valor ainda é muito próximo de 0, o que representa comentários neutros. No caso do termo “governo”, a negatividade já era esperada e, de fato, foi escolhido pela hipótese de que se esperava encontrar em comentários negativos, o que foi confirmado.

Já para o termo “creative commons”, a expectativa era justamente contrária ao resultado de fato obtido. Por ser um conjunto de normas que tenta facilitar o compartilhamento e a distribuição ética de obras, especialmente em formato digital, esperava-

se que fizesse parte de comentários positivos. É possível que por serem recentes, e de importância para o meio digital, os usuários encontrem mais dificuldade.

Abaixo (na Figura 8), vemos um exemplo de *tweet* que ilustra o comportamento informacional de um usuário que não sabe o que são as licenças Creative Commons descreve certa dificuldade em ter que utilizá-las. Acompanhando as respostas a este *tweet,* vemos que o usuário tentava publicar um vídeo para participar de um concurso, contudo, tinha dificuldade de lidar com a licença Creative Commons exigida pelos organizadores.

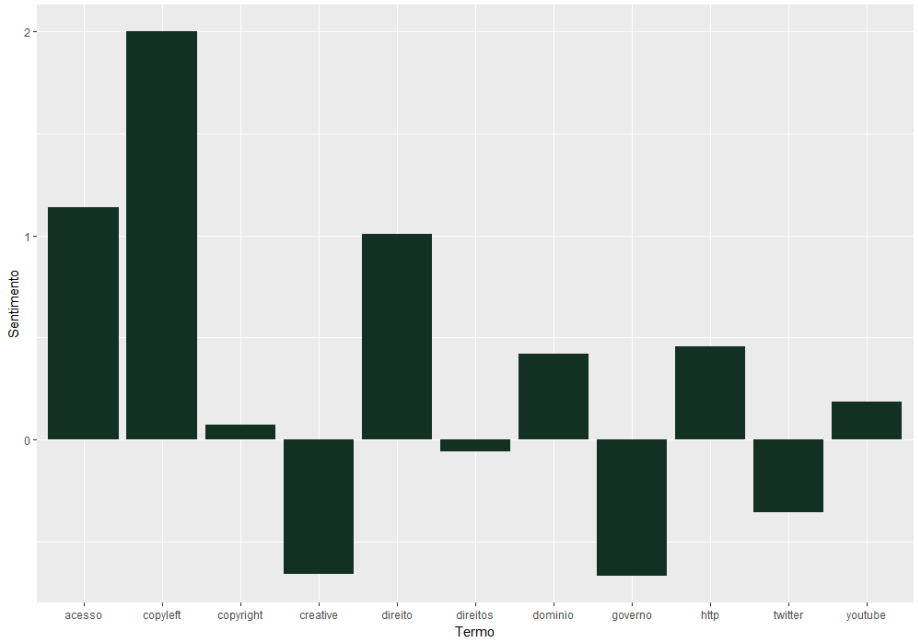
Figura 8 – Exemplo de tweet



Fonte: Twitter. Disponível em: <https://twitter.com/i/web/status/900009189304487940>.

A Figura 9 foi elaborada a partir do léxico SentiLex-PT02. Nele estão classificadas mais de 7 mil palavras em língua portuguesa. Utilizando este léxico, vemos que os valores variam entre 2 e – 0,6, mas ainda com médias de resultados positivos para a maioria das palavras-chave. Os termos com comentários mais negativos são “creative commons”, “governo” e “twitter”, novamente, além de “direitos autorais” que também é negativo neste gráfico, mas ainda muito próximo de 0.

Figura 9 – Sentimento por termo (SentiLex)



Fonte: elaborado pelo autor.

Numa comparação pode-se dizer que os dois léxicos trouxeram resultados coerentes entre si. De modo geral, o léxico OpLexicon traz resultados mais próximos da neutralidade, com apenas 3 termos com valores além de 0,5, para mais ou para menos. Já o SentiLex traz resultados com valores mais distintos, especialmente para os termos “acesso aberto” e “direito autoral”, que estão próximos da neutralidade no primeiro gráfico, mas são consideravelmente positivos no segundo.

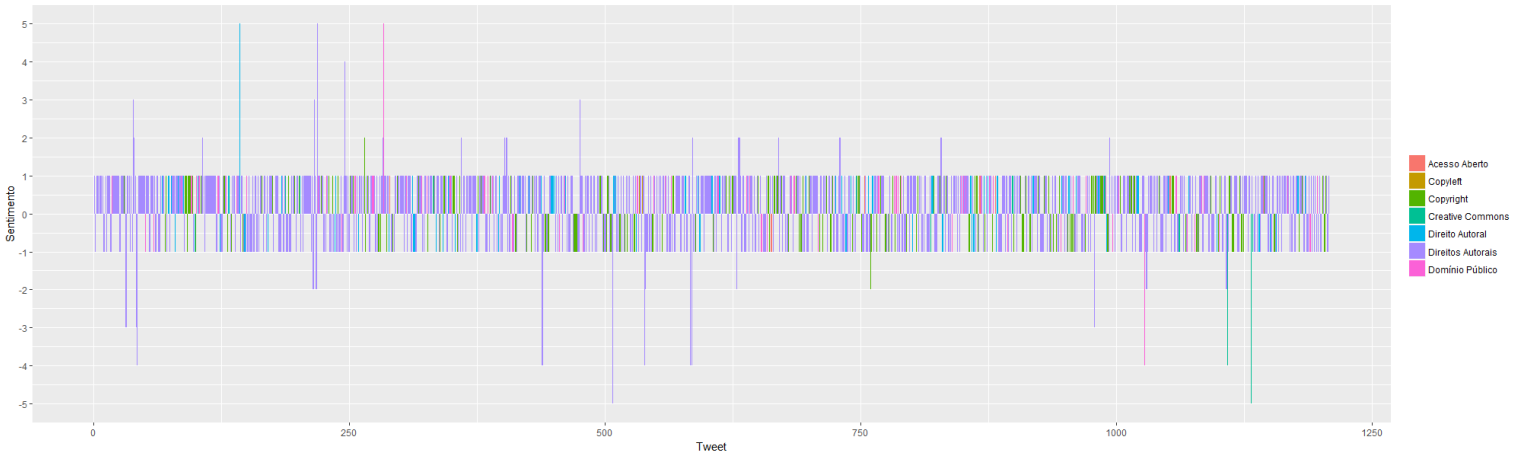
Os termos “copyleft” e “domínio público” são os únicos com altos valores positivos em ambos os léxicos. É de se esperar que sejam assuntos que agradem os usuários e que seja possível encontrar informações de maneira relativamente fácil. Ambos são conceitos que visam a divulgação de conteúdos sem barreiras, o que imagina-se ser de interesse para a maior parte dos usuários comuns. Contudo, o termo “creative commons”, que também relaciona-se à divulgação aberta e compartilhamento de informações, está conectado a comentários negativos em ambos os gráficos. Uma das possíveis explicações para isso seria justamente a crescente obrigatoriedade de uso das licenças Creative Commons por algumas instituições, websites, revistas científicas etc, o que pode causar dúvidas entre os usuários e insatisfação no momento de sua utilização.

Numa tentativa de ilustrar a dimensão temporal dos *tweets* coletados entre os dias 1º e 31 de Agosto de 2017, foram desenvolvidos os gráficos a seguir, para o léxico OpLexion

(Figura 10 – disponível em tamanho original no Apêndice A) e para o SentiLex (Figura 11 – disponível em tamanho original no Apêndice B), que apresentam os *tweets* em ordem cronológica em relação aos sentimentos expressos neles. Estão presentes apenas os *tweets* cujos valores de sentimento são diferentes de 0, pois não seria possível visualizar os elementos de valor 0 no gráfico.

Com o OpLexicon são apresentados 1.208 *tweets* com valores que variam entre "- 13" e "11". Para facilitar a visualização, os limites do gráfico foram definidos como "- 5" e "5", o que acabou por esconder 7 *tweets* de valores que vão além destes limites: são 4 positivos de valores "6", "7", "8" e "11"; e 3 negativos de valores "- 9", "- 10" e "-13".

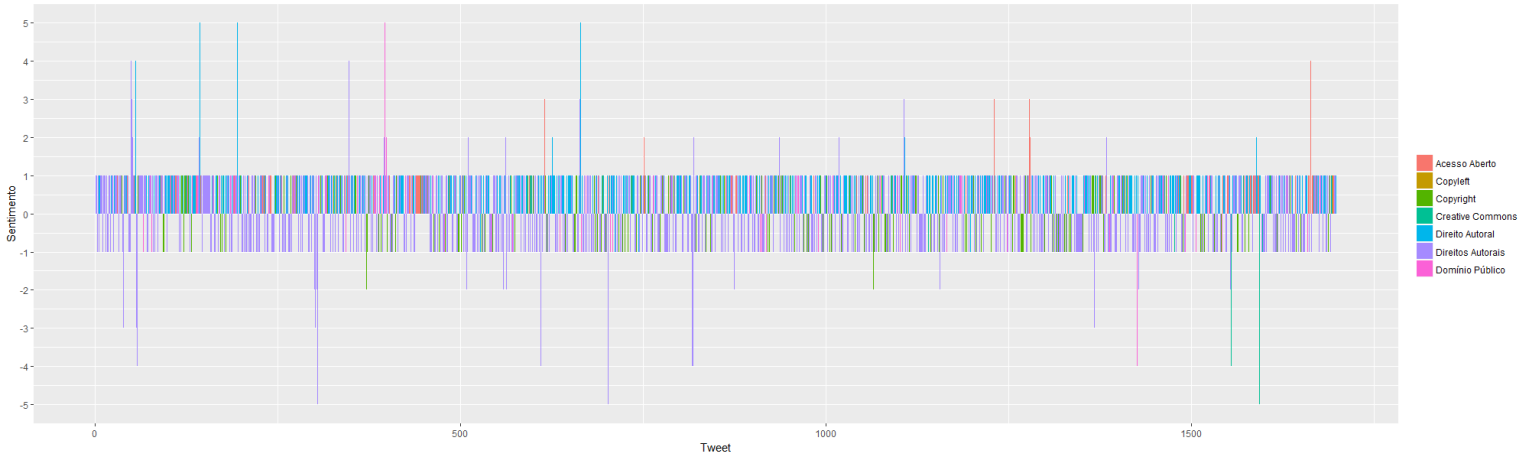
Figura 10 – Evolução do sentimento (OpLexicon)



Fonte: elaborado pelo autor.

Com o SentiLex são apresentados 1.698 *tweets* com valores variando entre "- 19" e "16". Como no gráfico anterior, mantivemos os limites de "- 5" e "5", o que escondeu 13 *tweets* de valores que ultrapassam os limites.

Figura 11 – Evolução do sentimento (SentiLex)



Fonte: elaborado pelo autor.

A partir deste gráfico, seria possível identificar algum tema ou evento que poderiam influenciar o comportamento informacional dos usuários. Como, por exemplo, a divulgação de uma lei controversa que afetaria os usuários negativamente, proporcionando uma grande quantidade de *tweets* negativos sobre o assunto durante determinado espaço de tempo. Contudo, não é possível identificar um padrão neste sentido. Por outro lado, pode-se ver que a maioria dos *tweets* tem sentimentos de valores entre "- 1" e "1", o que indica que, de fato, pouca carga emocional é expressada nas publicações analisadas.

Trazemos a seguir alguns exemplos transcritos de *tweets* que ilustrem as variadas cargas emocionais e que demonstrem as abordagens distintas que os usuários tomam quando falam sobre direito autoral e suas vertentes. Em verdade, o processo de escolha destes *tweets* não seguiu nenhum método específico, devido à grande quantidade coletada, os *tweets* não foram lidos um a um, mas, durante o processo de análise, alguns eram lidos aleatoriamente ou de relance. Dentre os lidos desta maneira, os mais interessantes foram salvos para que pudessem servir de exemplo, caso necessário. Estes são tweets que foram classificados como positivos:

- @irispalo disse que não tinha entendido direito. quem assina um contrato de direitos autorais no meio de uma festa??? depois diz que não tinha entendido
- video tava pronto, editado, lindo, maravilhoso, dai chega papum pirulitão, tá com direito autoral pqp grrrrrrrrrr
- retweet @democraciascfc: foi um papo bastante interessante sobre cultura livre, software livre, copyleft e contracultura. https://t.co/6nfebgfrin

- @Ucasberto Mano, tenho um canal, mas sempre que eu ponho algum episódio de alguma série ou algo assim na edição eu tomo direitos autorais, como fazer?

A seguir alguns tweets de valor negativo:

- @portalselenabr: Por motivos de direitos autorais e por muitas contas estarem sendo suspensas, achamos melhor apagar os vídeos da LIVE q…
- A música "Parabéns Pra Você" é protegida por direitos autorais e qualquer uso comercial sem autorização é ilegal (até 2017) aqui no Brasil.
- A música "Parabéns pra você" não é de domínio público. A Warner, que comprou os direitos autorais, recebe milhões p/ ano pelo uso da canção. https://t.co/GWytF6KskH
- Polícia Civil prende dois homens por tráfico e associação ao tráfico de drogas e uma mulher por violação de direito autoral.
- Posso usar o mesmo nome de uma música que já existe? Vem conferir 3 dicas para proteger seus direitos autorais https://t.co/HBKmgzibmP
- como eu queria que não tivesse isso de direitos autorais, ia ser mto daora usar as músicas nos vídeos, vários conceitos

Alguns exemplos de tweets neutros:

- @arthurrocha @felipeneto @umusicbrasil Isso de direitos autorais em youtube é ultrapassado, nenhuma outra gravadora faz isso, e ele tava ajudando na divulgação do clipe, isso é bom
- @felipeneto @YouTube Meus últimos 7 reaction foram bloqueados, @UMG, já passou da hora de evoluir, Copyright a troco de que? pq? #umgempresaultrapassada #UMG
- juiz usa princípio da adequação social e absolve acusado de violação de direito autoral https://t.co/hncmjvf4el

Até agora, apresentamos a contagem simples de palavras como parte de todas as análises descritas, o que exige o uso das *stopwords* para remover as palavras mais comuns e que são consideradas sem relevância. Contudo, outra abordagem é a frequência inversa nos documentos (*Inverse Document Frequency* – IDF), que diminui o peso das palavras que são comuns em todos os documentos e aumenta o peso de palavras que não são tão repetidas na coleção. Desse modo, não é necessário passar os filtros de *stopwords* e todas as palavras presentes passam a ser consideradas na análise. O IDF é, portanto, uma medida estatística que

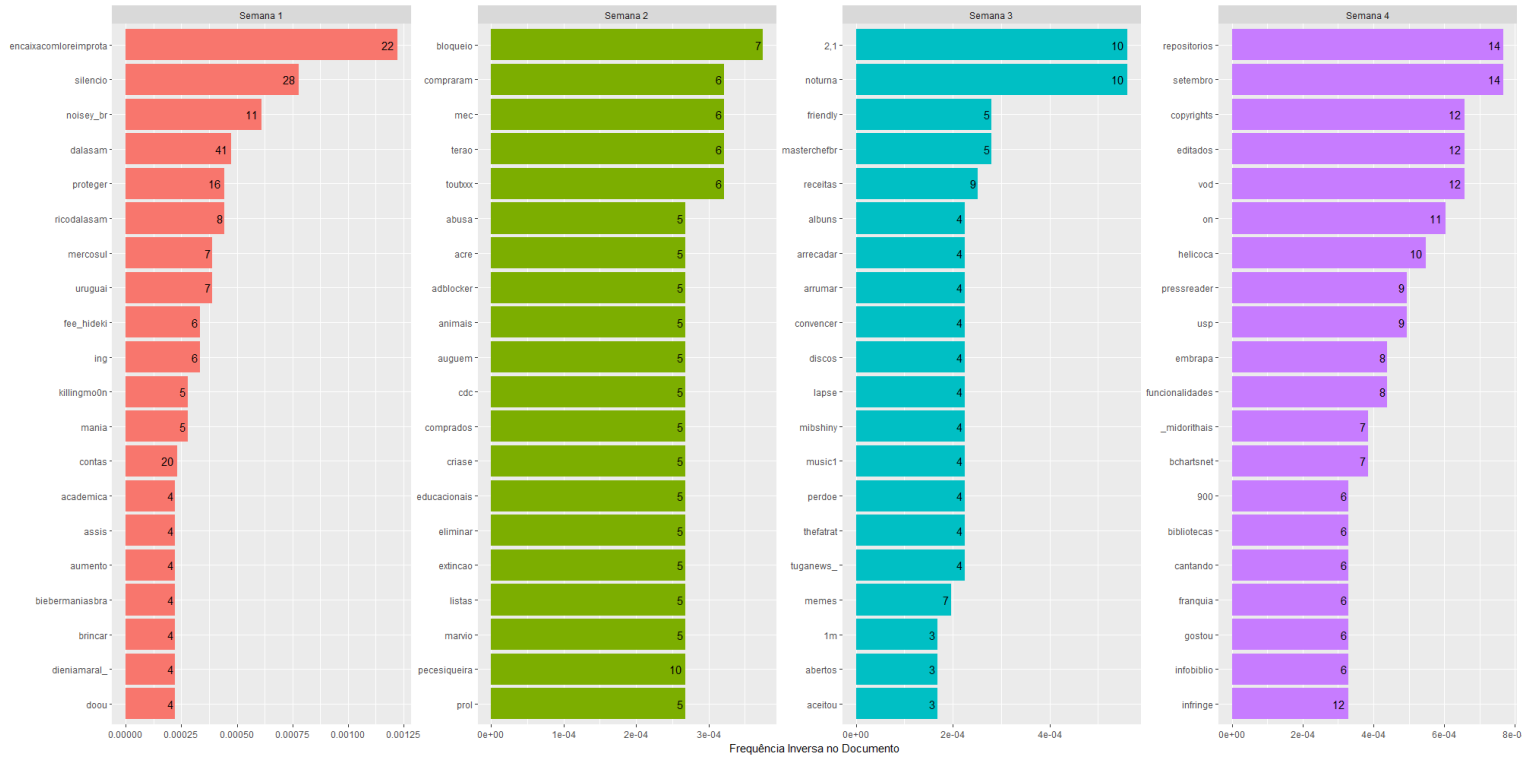
visa indicar a importância de uma palavra de um documento em relação a uma coleção de documentos, ou corpus (ALDAHAWI, 2015; GODFREY et al., 2014).

A técnica IDF pode ser aplicada em diversos tipos de corpus, mas uma coleção com diversos documentos, que sejam aproximadamente da mesma extensão, pode trazer resultados mais relevantes, como é o caso de uma coleção de *tweets*. A ideia principal da IDF é que se uma palavra aparece com frequência em um *tweet*, mas raramente em outros, ela tem uma grande habilidade para distinguir os assuntos dos *tweets*. A técnica utiliza a frequência de uma palavra num determinado *tweet* juntamente à frequência da palavra na coleção inteira para calcular seu peso.

A seguir apresentamos gráficos que elencam as palavras de acordo com suas frequências inversas nos documentos. Em cada linha é apresentado um número, que é o número de vezes que a palavra foi encontrada no total; contudo, o comprimento da linha é relacionado a outros fatores, são eles: o número de vezes que a palavra aparece num documento (*tweet*), em relação ao número de vezes que a palavra aparece em todo o corpus.

Na Figura 10 (Disponível em tamanho original no Apêndice C), as colunas foram divididas cronologicamente. A primeira relaciona apenas os *tweets* feitos na primeira semana, de 1º a 8 de Agosto de 2017; a segunda coluna para a segunda semana e assim sucessivamente. Nesta divisão é possível identificar as palavras mais importantes a cada semana, e como elas mudaram no decorrer da coleta, assim representando como os assuntos e discussões dos usuários também mudou no decorrer da coleta dos dados.

Figura 12 – IDF por semana



Fonte: elaborado pelo autor.

Dentre as palavras indicadas na Figura 10 é possível reconhecer algumas que ilustram tendências no Twitter. Por tendências entende-se assuntos de grande repercussão e que são comumente discutidos durante certo espaço de tempo, podendo ser identificadas a partir do uso de hashtags e/ou *retweets*. Alguns deles são: “silêncio” na primeira semana, “pecesiqueira” na segunda, “masterchefbr” e “receita” na terceira e “repositórios” e “embrapa” na quarta. Apresentamos a seguir alguns dos *tweets* que colaboraram com estas tendências:

- O vídeo do YouTube que recebeu uma notificação de copyright pelo uso de… silêncio https://t.co/EhrC1eFzSo https://t.co/D36nCCsFnk – Publicado em 04/08/2017.

Nesse caso a tendência aconteceu pela frequência de *retweets* compartilhando a mesma informação que soa absurda.

- @sergiuolivera @pecesiqueira Ein. Mas isso não pode dar problema de direito autoral?

- @pecesiqueira a unica coisa é que tem que ver os direitos autorais, mas de resto é tranquilo, venderia sim

Neste caso a tendência foi criada por usuários que responderam a dúvida do usuário @pecesiqueira, que foi: “Amigos ilustradores: voces vendem commissions de personagens de *desenho*/anime/filmes/séries, certo? *Venderiam camisetas c esses* desenhos?”

- Diante da polêmica da sobremesa do #MasterchefBR: Receita tem dono e direito autoral? - Via Estadão: https://t.co/DvacYsRYWJ

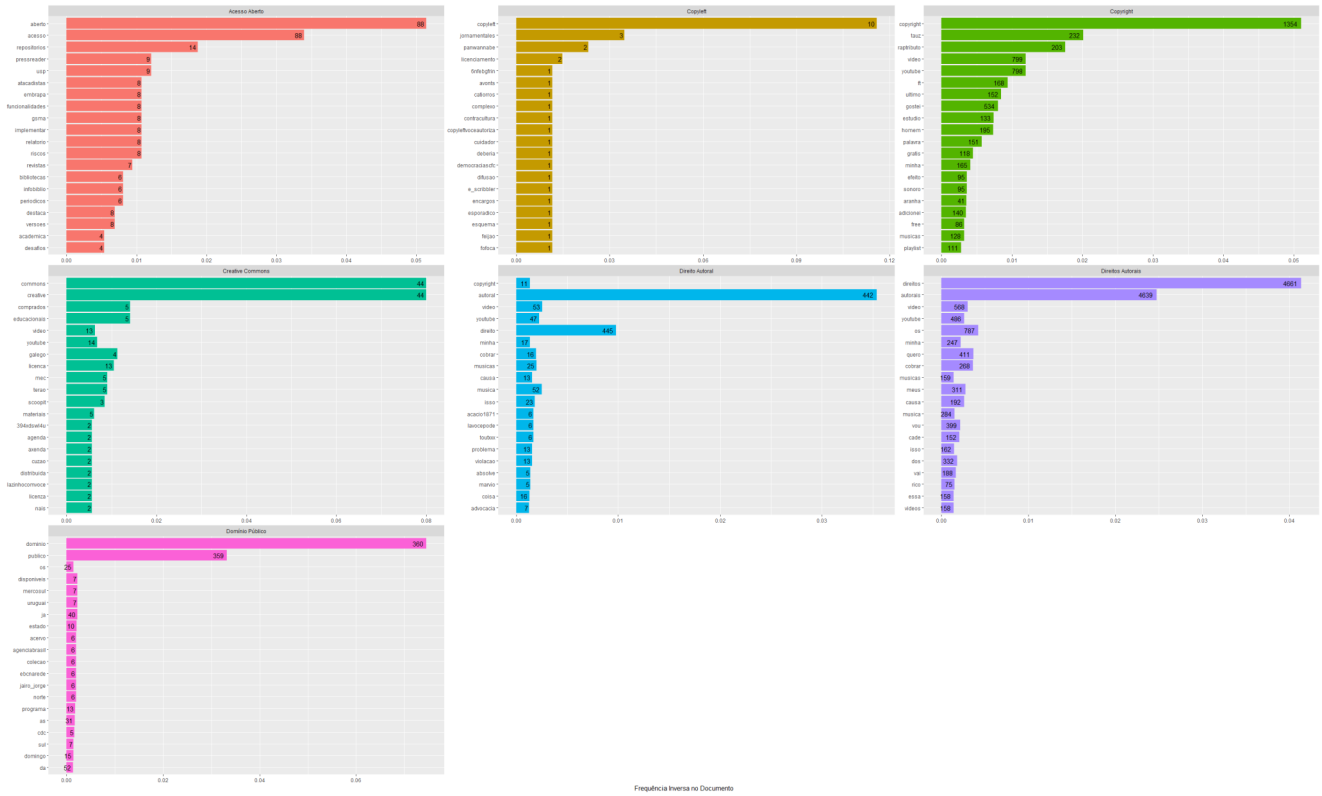
Este foi um caso que criou repercussão pois discutia-se a possibilidade de se tomar medidas legais contra a utilização de uma receita culinária criada por outro.

- Notícia @embrapa: Repositórios de acesso aberto da Embrapa ganham novas versões e funcionalidades https://t.co/5u6Bi7PpSj

Finalmente, aqui a tendência foi criada pelo compartilhamento da notícia de atualização dos repositórios da Embrapa.

A seguir apresentamos, na Figura 11 (Disponível em tamanho original no Apêndice D), um gráfico que também utiliza a IDF, mas onde as palavras foram agrupadas de acordo com a palavra-chave presente em seu texto. Assim, espera-se reconhecer assuntos e tendências que sejam relacionados às vertentes específicas de Direito Autoral escolhidas como filtros no momento da coleta.

Figura 13 – IDF por palavra-chave



Fonte: elaborado pelo autor.

Neste gráfico também é possível reconhecer algumas tendências para cada palavra-chave, sejam elas por assuntos de grande repercussão, uso de hashtags etc. Apresentaremos a seguir alguns exemplos destas tendências:

- PressReader na USP – acesso aberto até 30 de setembro https://t.co/xOkOVSPIIW #usp #bibliotecas
- Nem todo "conteúdo aberto" ou um Recurso Educacional Aberto é gratuito! "Acesso gratuito" é diferente de "acesso aberto". #edc287

Este *tweet*, compartilhado diversas vezes, anuncia um novo recurso disponível por tempo limitado aos usuários das bibliotecas do sistema SIBi-USP.

- @jornamentales @panwannabe Copyleft tem nada a ver com orientação política, é tipo de licenciamento de conteúdo.

Este é um trecho de uma conversa no Twitter onde 3 usuários discutem a relação entre "copyleft" e os movimentos políticos de esquerda.

- Gostei de um vídeo @YouTube https://t.co/jVwuVvreB3 de Terror #4 / Efeito Sonoro Grátis e Sem Copyright
- @yurizandotv NOSSA DA MT RAIVA E QUANDO C COLOCA SÓ UMA PEQUENA PARTE D UMA MUSICA E MESMO ASSIM PEGA DIREITO AUTORAL

Na coluna de "Copyright" vemos algumas tendências de compartilhamento, especialmente de vídeos que não possuam copyright, ou que tragam músicas, efeitos sonoros e outros elementos que poderiam ser utilizados livremente para criação de novos conteúdos. As tendências sobre youtube, vídeos e músicas são reconhecíveis também para as palavras-chave "Direito Autoral" e "Direitos Autorais".

- Materiais educacionais comprados pelo MEC terão licença Creative Commons #YearofOpen https://t.co/n5n3NvSfYh
- O que é Creative Commons? https://t.co/l1aGkeFel4
- @seekpedro Quando apresentei, tive o feedback que poderia pegar a "propriedade intelectual" como o Creative Commons e produzir artigos sobre o tema.

Para os *tweets* sobre Creative Commons é possível identificar tendências sobre divulgações de materiais disponíveis em Creative Commons, assim como dúvidas ou divulgações sobre o uso das licenças abertas.

- China quer negociar tratado de livre-comércio com Mercosul e Uruguai. https://t.co/BmepDkUvcX Domínio Público https://t.co/sThhXfpiT4
- Disponíveis no portal Domínio Público a Coleção Educadores e a Coleção História Geral da África para download… https://t.co/ReOmcHqg1m
- O filme *Um Barco e Nove Destinos* é de domínio público e está disponível online no acervo de filmes antigos que... https://t.co/RsWW1kLpbG

Já as tendências para os *tweets* sobre domínio público podemos identificar casos de divulgação de materiais e também casos onde o termo "Domínio Público" é adicionado como fonte da imagem ou informação compartilhada na mensagem.

5 CONSIDERAÇÕES FINAIS

Nesta pesquisa, pretendeu-se investigar o comportamento informacional dos usuários do Twitter no que diz respeito a buscar e compartilhar informações relacionadas ao Direito Autoral, mediante a captação e análise de *tweets* sobre o assunto.

A plataforma Twitter e seu conceito de *microblogging* fornecem uma oportunidade interessante para estudar tal comportamento e as tendências dos usuários dentro deste ambiente específico. A enorme quantidade de textos produzidos e disseminados em plataformas como o Twitter impossibilitam uma leitura simples como método para determinar seus conteúdos; desse modo, as ferramentas de *Text Mining* oferecem um método automatizado capaz de extrair estes conteúdos e facilitar sua análise.

Nas discussões e exemplos apresentados, foi possível constatar a versatilidade dos processos de descoberta de conhecimento em bases de dados e textos. A aplicação desses processos é limitada apenas pela quantidade de dados e metadados disponível sobre o assunto de interesse. Ademais, o desenvolvimento das tecnologias de informação e comunicação influencia a produção e disseminação de conteúdos em ambientes virtuais e mídias sociais, caracterizando um cenário ideal para ferramentas como a mineração de textos.

Mais especificamente, pudemos reconhecer que as interações entre os usuários do Twitter neste período foi alinhada às possibilidades de uso da plataforma e às expectativas de compartilhamento de informação na Web de modo geral. A possibilidade para adicionar links, assim como a prática dos *retweets*, são os principais exemplos de compartilhamento presente nas mensagens coletadas. Também reconhecemos alguns casos de perguntas diretas feitas com a perspectiva de que seriam respondidas pelos seguidores do usuário em questão, porém, para contabilizar e reconhecer de fato esta tendência de uso seria necessário ler cada *tweet* especificamente. Por outro lado, foi possível estabelecer que os usuários podem sim utilizar a plataforma como meio de obtenção de informações.

Espera-se, também, ter abordado os métodos e técnicas de mineração textual de forma a auxiliar futuros trabalhos e pesquisas que visem atingir objetivos semelhantes. Para tanto, apresentamos as etapas do desenvolvimento de modo que servissem de exemplo para a utilização do software R e suas características. Dentre as dificuldades encontradas, podemos listar diversos pontos durante a etapa de pré-processamento, etapa que é, de fato, mais propícia a erros e de difícil aplicação. Seria interessante aplicar outras listas de stopwords, ou mesmo criar uma específica para o uso nesta coleção de tweets, pois a Língua Portuguesa

permite diversos regionalismos e adaptações da linguagem formal, especialmente quando utilizada em ambientes virtuais. Além das conjugações verbais, plurais e sufixos de gênero que afetam os resultados durante análises estatísticas.

Desse modo, apesar de ainda não fazerem parte do arsenal de ferramentas da Ciência da Informação, tais processos podem certamente ser aplicados nos diversos campos de pesquisa da CI. De modo geral, é possível identificar a aplicação de técnicas de mineração textual usualmente em trabalhos das áreas de computação e linguística, então, faz-se interessante que a comunidade da CI esteja aberta e disposta a dialogar com estes pesquisadores a fim de trazer contribuições à evolução da área.

Aplicações clássicas da CI, como a indexação de assuntos, podem utilizar da mineração de textos para auxiliar na determinação de assuntos, em especial para documentos de alta complexidade ou especificidade e documentos de produção em massa. Passando também por aplicações mais modernas como o estudo de usuários e necessidades informacionais, a fim de identificar os métodos de busca destes usuários e suas dúvidas mais comuns, com foco em mídias sociais, ou ainda, com foco em empresas privadas e estratégias de marketing, para potencialmente auxiliar nas tomadas de decisão e na resposta às necessidades dos clientes.

Assim, fica evidente a necessidade de mais estudos que explorem as diversas aplicações da mineração de dados e de textos nos contextos digitais de comportamento informacional. Entretanto, também há a necessidade de adaptação dos profissionais até que estas ferramentas sejam utilizadas em larga escala, além de treinamentos e especializações aos que estão interessados mas não possuem o conhecimento e prática necessários para utilização das ferramentas.

Nesta perspectiva, mostra-se oportuno prosseguir com investigações que reconheçam o usuário das mídias sociais como criador, leitor e disseminador de informações, a fim de alcançar entendimento sobre seu comportamento, de modo a possibilitar mudanças em seu favor no que diz respeito ao uso da informação em ambientes digitais.

REFERÊNCIAS

ADAMIC, L. A. et al. Knowledge sharing and yahoo answers: everyone knows something. Proceeding of the 17th international conference on World Wide Web - WWW '08. **Anais**...Beijing, China: 2008. Disponível em: <http://portal.acm.org/citation.cfm?doid=1367497.1367587>.

ADAMS, S. S. What Games Have to Offer: Information Behavior and Meaning-Making in Virtual Play Spaces. **Library Trends**, v. 57, n. 4, p. 676–693, 2009.

AGOSTO, D. E.; HUGHES-HASSELL, S. People, places, and questions: An investigation of the everyday life information-seeking behaviors of urban young adults. **Library & Information Science Research**, v. 27, p. 141–163, 2005.

ALDAHAWI, H. A. **Mining and Analysing Social Network in the Oil Business : Twitter Sentiment Analysis and Prediction Approaches**. [s.l.] Cardiff University, 2015.

BAILEY, J. Deflating the Michelin Man : Protecting Users ' Rights in the Canadian Copyright Reform Process. In: **Public Interest: The Future of Canadian Copyright Law**. Toronto: Irwin Law Book, 2005. p. 125–166.

BLAKE, C. Text Mining. **Annual Review of Information Science and Technology**, v. 45, n. 1, p. 123–155, 2011.

BOULTON, G. et al. **Science as an open enterprise**. London: The Royal Society of Science, 2012.

BRANCO, S. **O domínio público no direito autoral brasileiro**. Rio de Janeiro: Lumen, 2011.

CASE, D. O. **Looking for Information: a Survey of Research on Information Seeking, Needs, and Behavior**. 2nd. ed. Amsterdam: Elsevier Inc., 2007.

CHOWDHURY, G. G.; CHOWDHURY, S. **Information Users and Usability**. London: Facet Publishimg, 2011.

CONEGLIAN, C. S.; SANTAREM SEGUNDO, J. E.; SANT'ANA, R. C. G. Big Data: fatores potencialmente discriminatórios em análise de dados. **Em Questão**, v. 23, n. 1, p. 62–86, 2017.

CONNAWAY, L. S.; DICKEY, T. J.; RADFORD, M. L. "If it is too inconvenient I'm not going after it:" Convenience as a critical factor in information-seeking behaviors. **Library and Information Science Research**, v. 33, n. 3, p. 179–190, 2011.

DAVENPORT, E. Confessional Methods and Everyday Life Information Seeking. **Annual Review of Information Science and Technology**, v. 44, n. 1, p. 533–562, 2010.

DAVIES, M. M.; BATH, P. A. Interpersonal sources of health and maternity information for Somali women living in the UK. **Journal of Documentation**, v. 58, n. 3, p. 302–318, 2002.

DIXON, M. An overview of document mining technology. **Computer Based Learning Unit, University of Leeds**, 1997.

EBNER, M.; SCHIEFNER, M. Microblogging - more than fun ? Procceding of IADIS Mobile Learning Conference 2008. **Anais**...Algarve, Portugal: 2008.

ELSWEILER, D.; HARVEY, M. Engaging and Maintaining a Sense of Being Informed: Understanding the Tasks Motivating Twitter Search. **Journal of the Association for Information Science and Technology**, v. 66, n. 2, p. 264–281, 2015.

ERDELEZ, S. Information encountering: It's more than just bumping into information. **Bulletin of the American Society for Information Science**, v. 25, n. 3, p. 25, 1999.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI magazine**, p. 37–54, 1996.

FELDMAN, R.; DAGAN, I. Knowledge Discovery in Textual Databases (KDT). KDD'95 Proceedings of the First International Conference on Knowledge Discovery and Data Mining. **Anais**...Canada: 1995

FERREIRA, J. T. L.; ARAÚJO, R. F. DE. Compartilhamento de Informação Ambiental e a Repercussão do Código Florestal no Twitter. **Ciência da Informação em Revista**, v. 2, n. 1, p. 44–54, 2015.

FOSTER, J. Collaborative lnformation Seeking and Retrieval. **Annual Review of Information Science and Technology**, v. 40, n. 1, p. 329–356, 2006.

GODFREY, D. et al. A Case Study in Text Mining: Interpreting Twitter Data From World Cup Tweets. **arXiv**, v. 08, n. 5427, p. 1–11, 2014.

GOECKS, J.; MYNATT, E. Leveraging social networks for information sharing. Proceedings of the 2004 ACM conference on Computer supported cooperative work (CSCW '04). **Anais**...Chicago, USA: 2004.

HARTMANN, N. et al. Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks. **arXiv**, v. 1708.06025, n. Section 3, 2017.

HE, W.; ZHA, S.; LI, L. Social media competitive analysis and text mining: A case study in the pizza industry. **International Journal of Information Management**, v. 33, p. 464–472, 2013.

JAMES, J. **Data Never Sleeps 5.0**. Disponível em: <https://www.domo.com/blog/data-never-sleeps-5/>. Acesso em: 28 out. 2017.

JANSEN, B. J. et al. Real time search on the web: Queries, topics, and economic value. **Information Processing and Management**, v. 47, n. 4, p. 491–506, 2011.

JAVA, A. et al. Why We Twitter: Understanding Microblogging Usage and Communities. Procedings of the Joint 9th WEBKDD and 1st SNA-KDD Workshop 2007. **Anais**...San Jose, USA: Springer, 2007.

JOHNSTON, K. et al. Social capital: the benefit of Facebook “friends”. **Behaviour and Information Technology**, v. 32, n. 1, p. 24–36, 2013.

KAPLAN, A. M.; HAENLEIN, M. Users of the world, unite! The challenges and opportunities of Social Media. **Business Horizons**, v. 53, n. 1, p. 59–68, jan. 2010.

KHOO, C. S. G. Issues in Information Behaviour on Social Media. **Library and Information Science Research**, v. 24, n. 2, p. 75–96, 2014.

KWARTLER, T. **Text Mining in Practice with R**. Hoboken, NJ: John Wiley & Sons, 2017.

LEE, E. J.; OH, S. Y. Seek and You Shall Find? How Need for Orientation Moderates Knowledge Gain from Twitter Use. **Journal of Communication**, v. 63, n. 4, p. 745–765, 2013.

LEMOS, A. Ciber-Cultura-Remix. **Sentidos & processos**, v. 21, n. 7, p. 1–9, 2005.

LESSIG, L. **Cultura Livre:** Como a Grande Mídia Usa a Tecnologia e a Lei Para Bloquear a Cultura e Controlar a Criatividade. São Paulo: Trama Universitário, 2005.

MATTHEWS, P. Search Delegation, Synthesists and Expertise on Social Media. **Library and Information Science Research**, v. 24, n. 2, p. 97–107, 2008.

MORRIS, M. R.; TEEVAN, J.; PANOVICH, K. What Do People Ask Their Social Networks, and Why? Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI 2010. **Anais**...Atlanta, USA: 2010. Disponível em: <http://dl.acm.org/citation.cfm?id=1753587>.

MUTULA, S. M. Policy gaps and technological deficiencies in social networking environments : Implications for information sharing. **Journal of Information Management**, v. 15, n. 1, p. 1–9, 2010.

NEWTON, C. Twitter is rolling out 280-character tweets around the world. **The Verge**, p. 1–13, nov. 2017.

NORTON, M. J. Knowledge Discovery in Databases. **Library Trends**, v. 48, n. 1, p. 9–21, 1999.

OH, S.; ZHANG, Y.; PARK, M. S. Cancer information seeking in social question and answer services: identifying healthrelated topics in cancer questions on Yahoo! Answers. **Information Research**, v. 21, n. 3, p. 1–26, 2016.

RECUERO, R.; ZAGO, G. “RT, por favor”: considerações sobre a difusão de informações no Twitter. **Fronteiras – estudos midiáticos**, v. 12, n. 2, p. 69–81, 2010.

REIPS, U.-D.; GARAIZAR, P. Mining twitter: A source for psychological wisdom of the crowds. **Behavior Research Methods**, v. 43, n. 3, p. 635–642, 2011.

RIBEIRO, S. F. Análise De Documentos Jurisprudenciais : Uma Abordagem Utilizando Text Mining Para Determinação De Assuntos e Seleção de Termos Para Indexação. XXIX Encontro Nacional de Engenharia de Produção. **Anais**...Salvador: 2009.

SAVOLAINEN, R. "Living encyclopedia" or idle talk? Seeking and providing consumer information in an Internet newsgroup. **Library and Information Science Research**, v. 23, n. 1, p. 67–90, 2001.

SAVOLAINEN, R. Source preferences in the context of seeking problem-specific information. **Information Processing and Management**, v. 44, n. 1, p. 274–293, 2008.

SAVOLAINEN, R. Requesting and providing information in blogs and internet discussion forums. **Journal of Documentation**, v. 67, n. 5, p. 863–886, 2011.

SEMIOCAST. Brazil becomes 2nd country on Twitter, Japan 3rd, Netherlands most active country. **Semiocast**, 2012.

SILGE, J.; ROBINSON, D. **Text Mining with R: A Tidy Approach**. [s.l.] O'Reilly Media, 2017.

SILVA, M. J. et al. Automatic Expansion of a Social Judgment Lexicon for Sentiment Analysis. Technical Report - LASIGE. **Anais**...Lisboa, Portugal: FC-DI - University of Lisbon, 2010. Disponível em: <http://hdl.handle.net/10451/14068>.

SOUZA, M.; VIEIRA, R. Sentiment Analysis on Twitter Data for Portuguese Language. (P. F. Caseli H., Villavicencio A., Teixeira A., Ed.)Computational Processing of the Portuguese Language. PROPOR 2012. Lecture Notes in Computer Science. **Anais**...Berlin: Springer, 2012. Disponível em: <http://link.springer.com/10.1007/978-3-642-28885-2>.

UPSHALL, M. Text mining: Using search to provide solutions. **Business Information Review**, v. 31, n. 2, p. 91–99, 2014.

WILDEMUTH, B. M.. Descriptive Statistics. In: Applications of Social Research Methods to Questions in Information and Library Science. Westport: Libraries Unlimited, 2009. p. 338-347.

WILSON, T. D. On User Studies and Information Needs. **Journal of Documentation**, v. 37, n. 1, p. 3–15, 1981.

WILSON, T. D. Models in Information Behaviour Research. **Journal of Documentation**, v. 55, n. 3, p. 249–283, 1999.

WILSON, T. D. Human information behavior. **Informing Science**, v. 3, n. 2, p. 49–55, 2000.

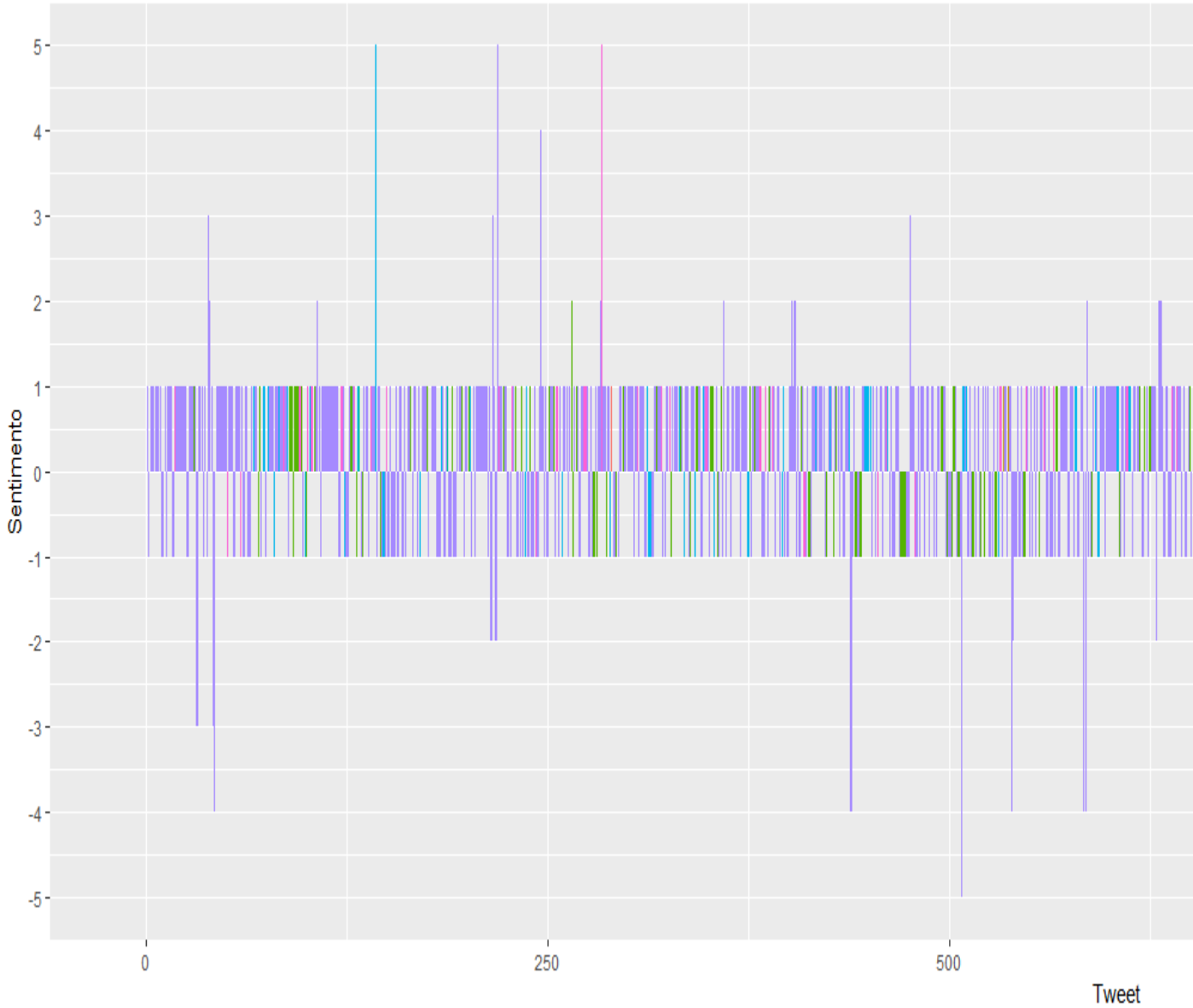
WILSON, T. D. Activity Theory and Information Seeking. **Annual Review of Information Science and Technology**, p. 119–161, 2008.

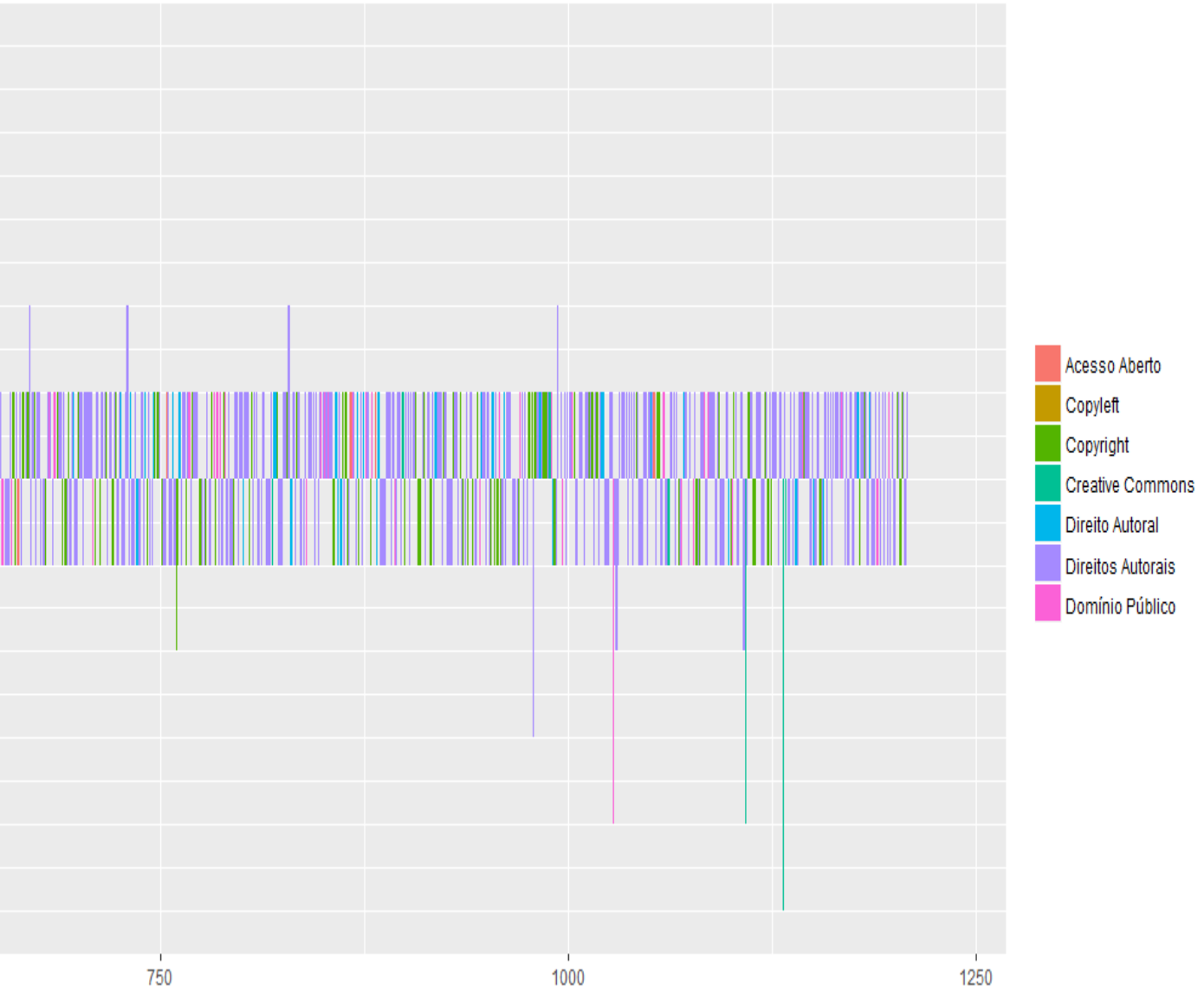
WITTEN, I. H.; FRANK, E. **Data Mining:** Practical Machine Learning Tools and Techniques. 2nd. ed. San Francisco, USA: Elsevier, 2005.

YU, X. et al. Mining online reviews for predicting sales performance: A case study in the movie domain. **IEEE Transactions on Knowledge and Data Engineering**, v. 24, n. 4, p. 720–734, 2012.

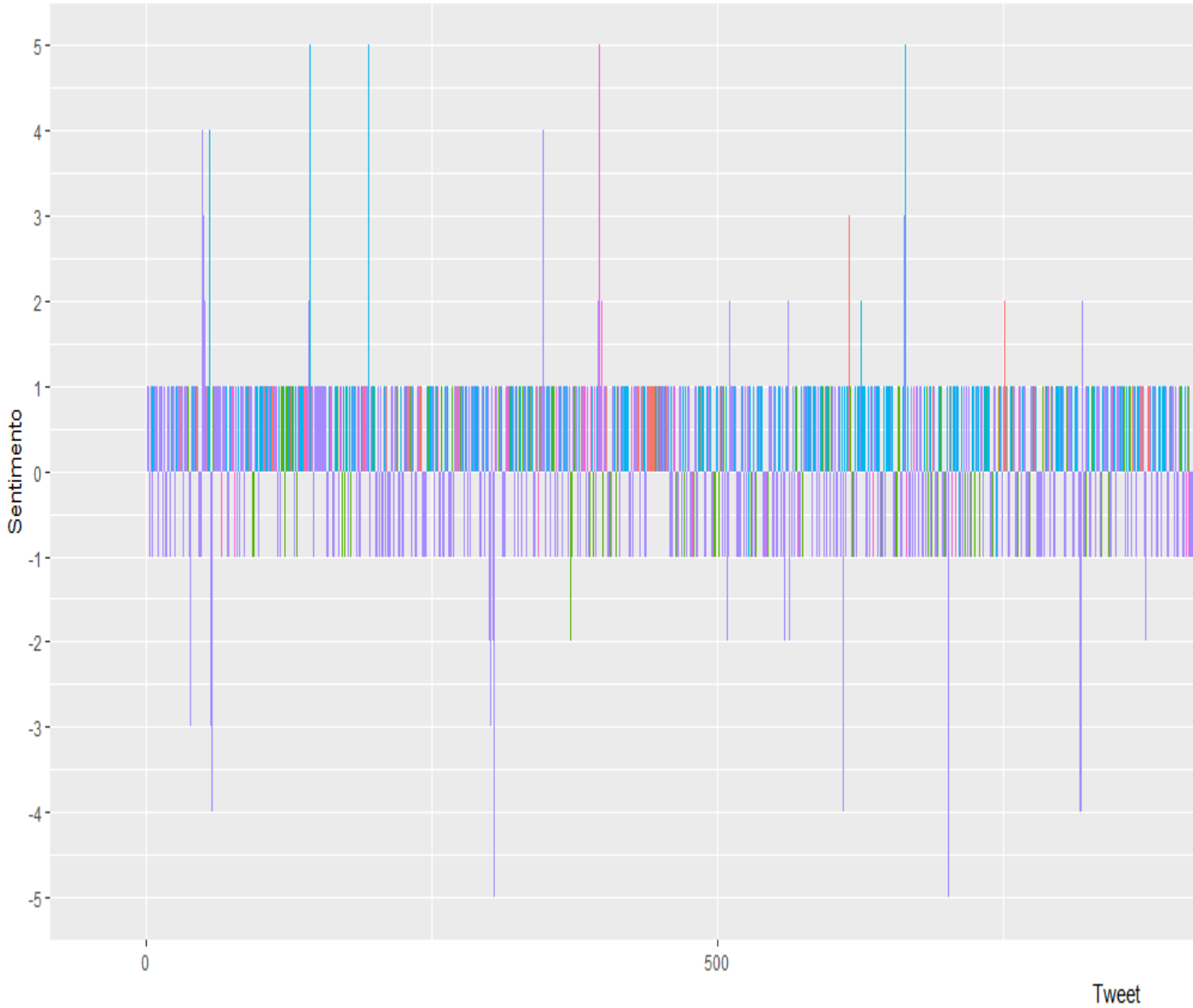
ZHANG, Y.; GU, H. Text Mining with Application to Academic Libraries. In: **Computer Science for Environmental Engineering and EcoInformatics**. Kunming, China: Springer, 2011. p. 200–205.

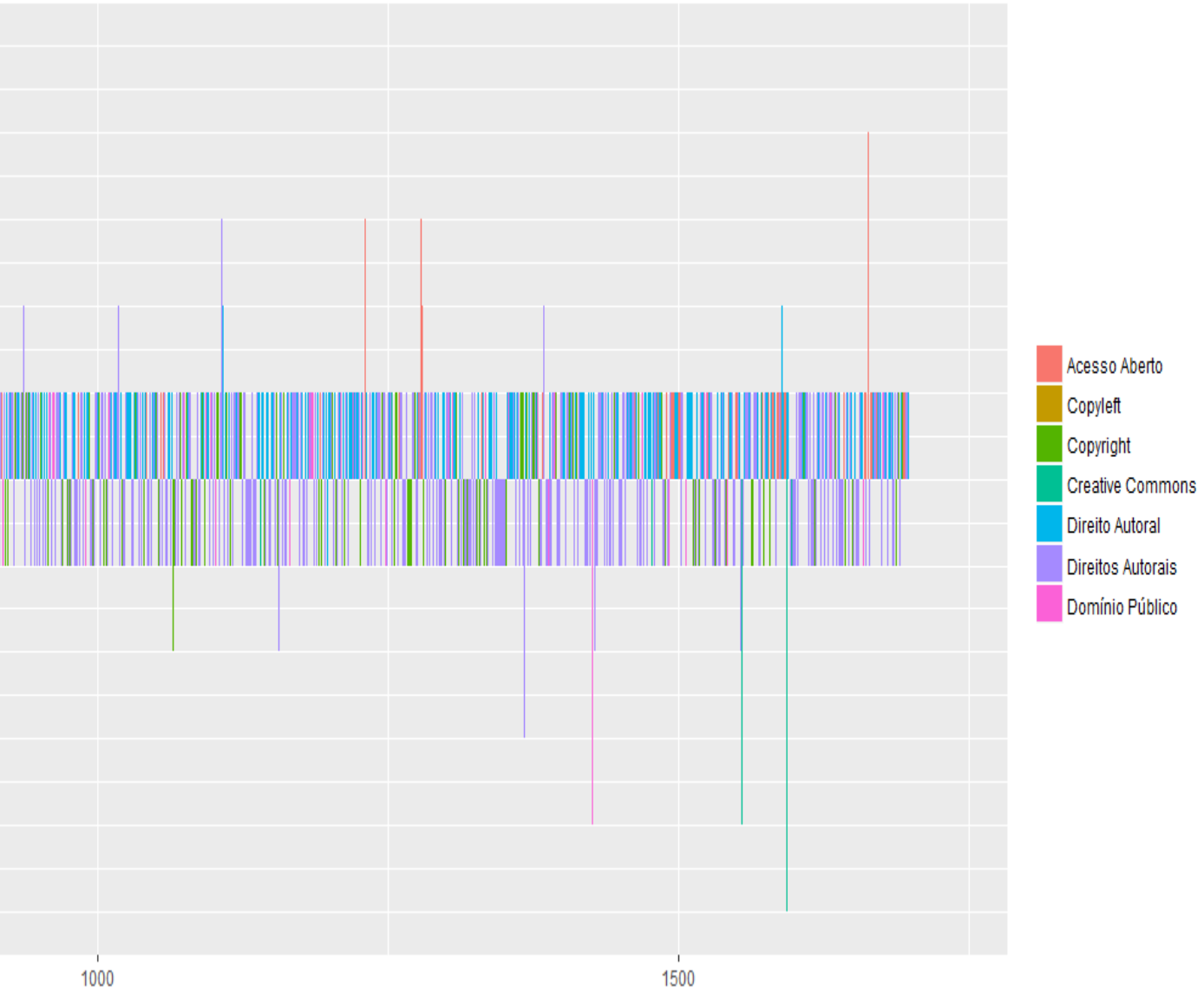
APÊNDICE A – FIGURA 8: EVOLUÇÃO DO SENTIMENTO (OPLEXICON)



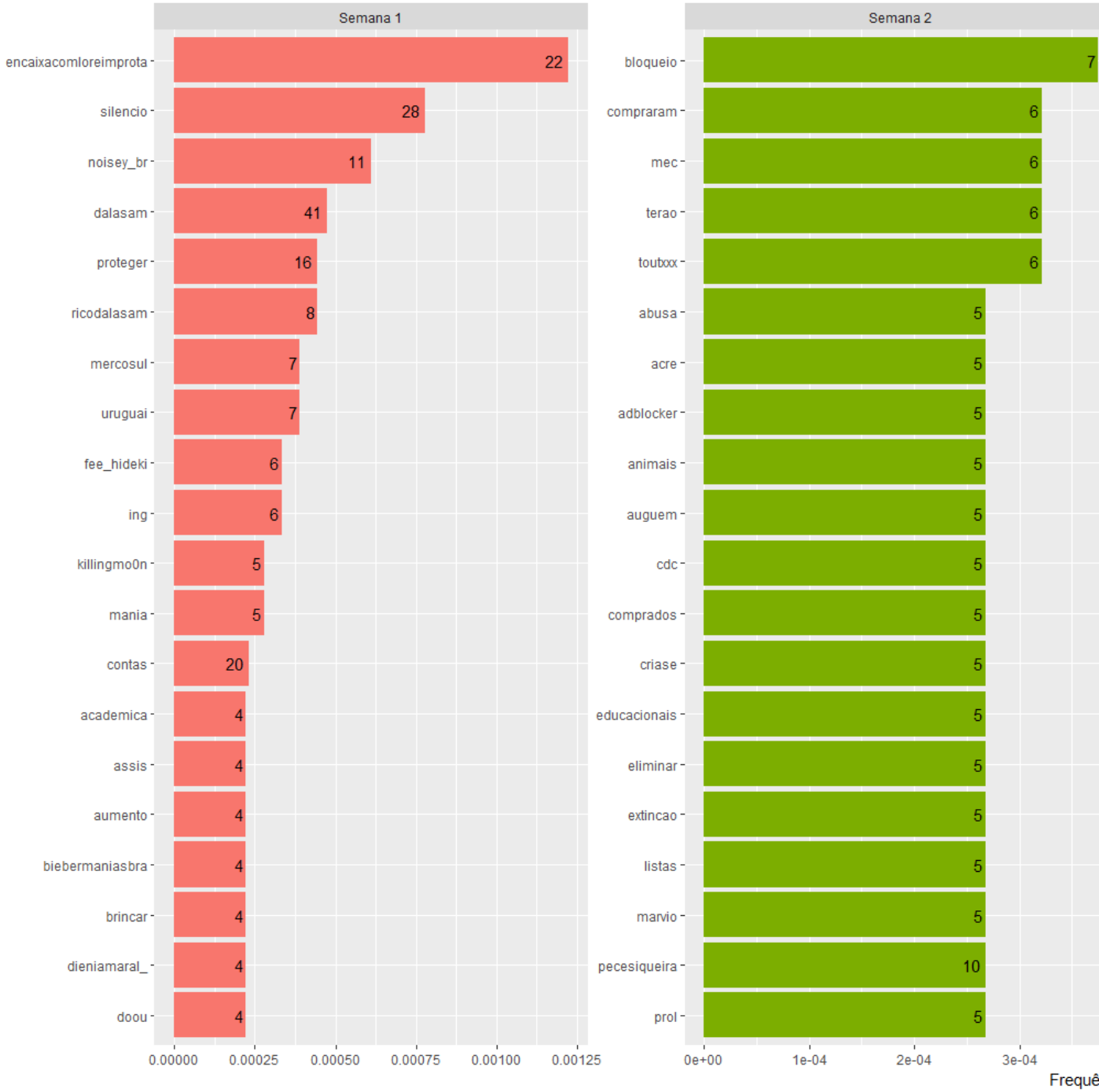

750
1000
1250
Acesso Aberto
Copyleft
Copyright
Creative Commons
Direito Autoral
Direitos Autorais
Domínio Público

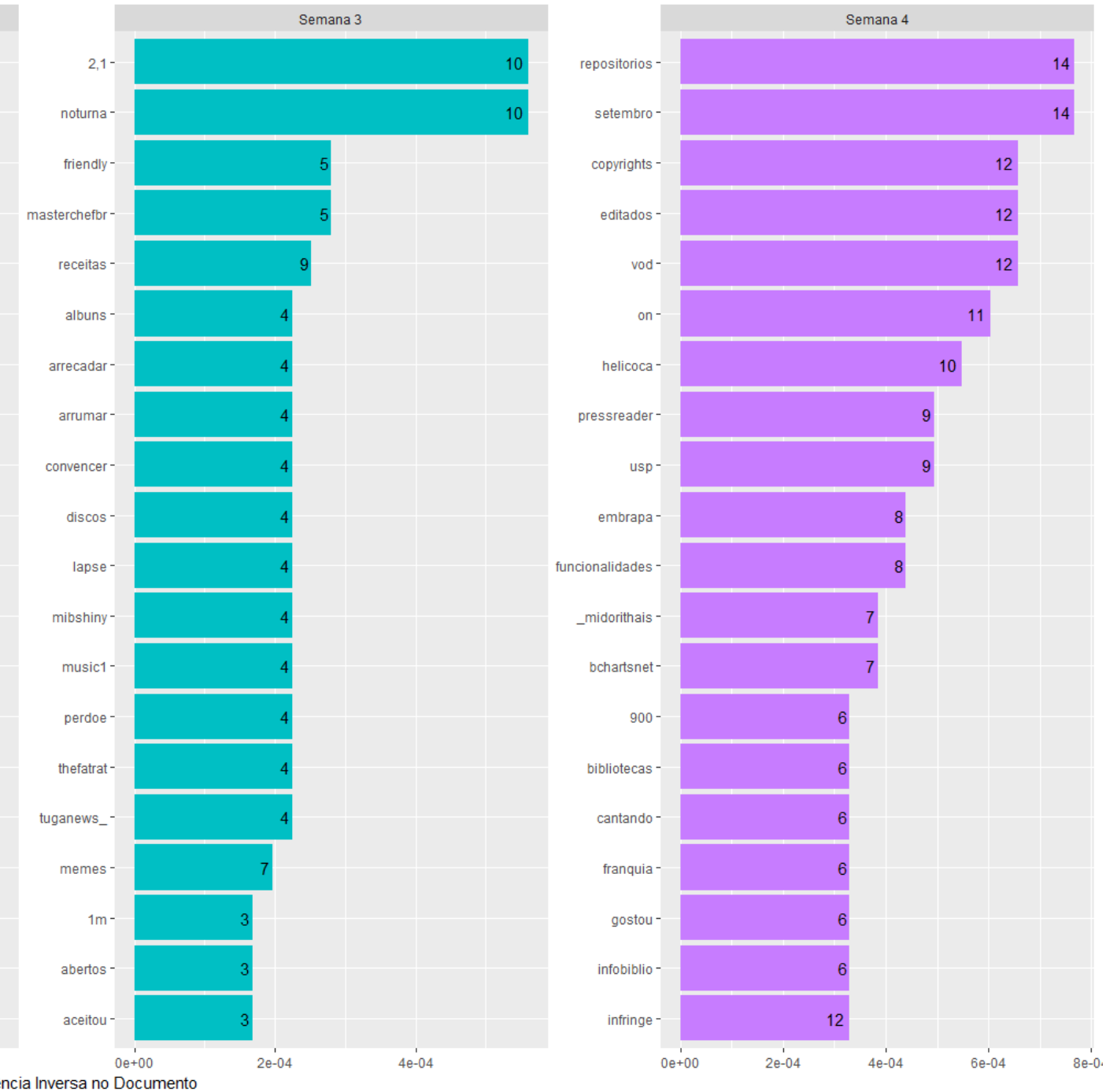
APÊNDICE B – FIGURA 9: EVOLUÇÃO DO SENTIMENTO (SENTILEX)




1000
1500
Acesso Aberto
Copyleft
Copyright
Creative Commons
Direito Autoral
Direitos Autorais
Domínio Público

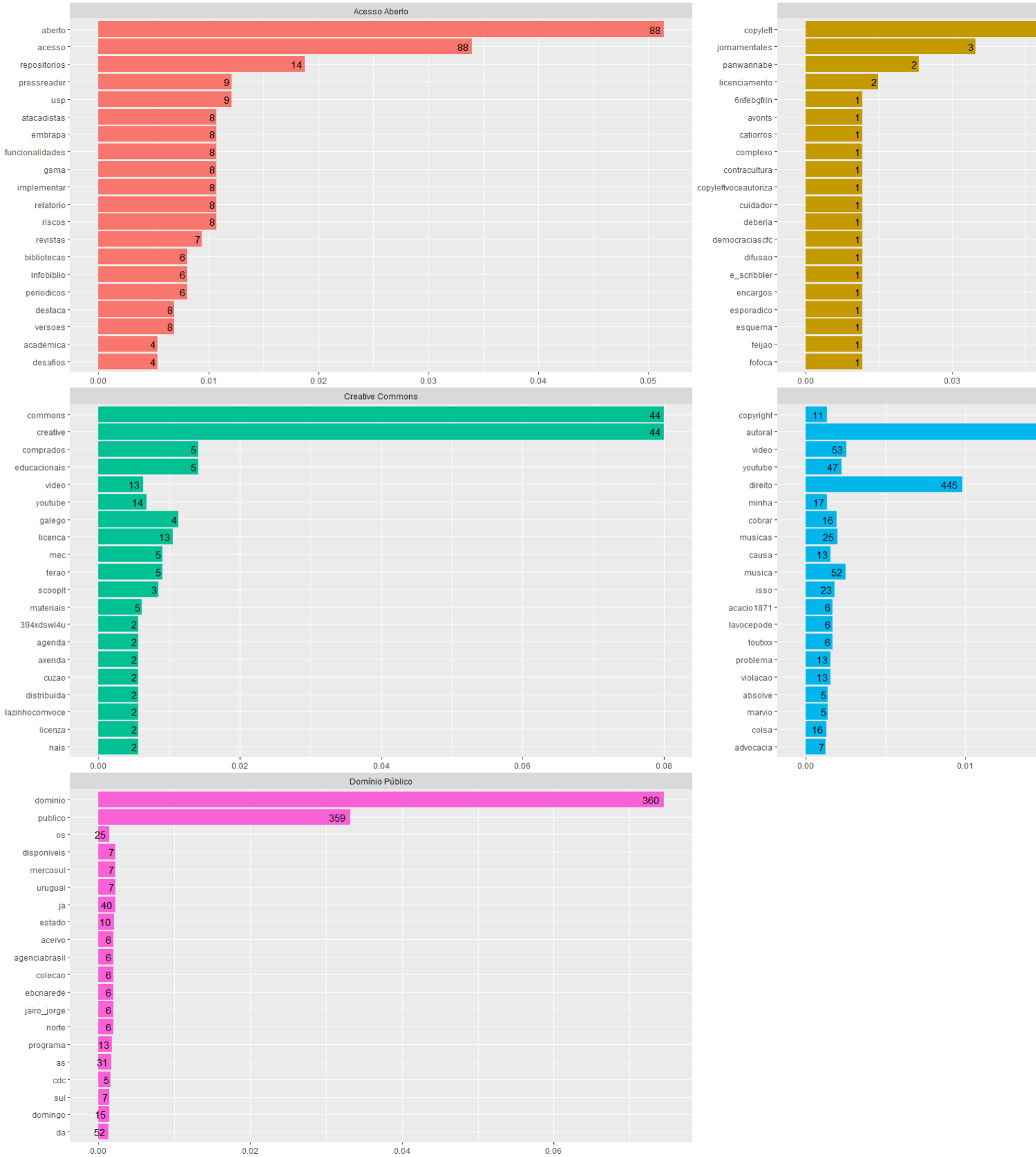
APÊNDICE C – FIGURA 10: IDF POR SEMANA





ncia Inversa no Documento

APÊNDICE D – FIGURA 11: IDF POR PALAVRA-CHAVE



Frequênci