Comportamento das Hashtags Durante Grandes Eventos

Pedro Pinto
Universidade Federal do Rio de Janeiro
pedropinto@ufrj.br

Ingrhid Theodoro
Universidade Federal do Rio de Janeiro
indytheodoro@ufrj.br

Jonice Oliveira
Universidade Federal do Rio de Janeiro
jonice@dcc.ufrj.br

ABSTRACT
Hashtags are words or terms which are used to categorize publications in social media, indexing the content and making it available for retrieval. The *hashtags* creation is free process, i.e, there are no rules or restrictions to create them. Otherwise, there is a convergence of the terms used to explain and semantically enrich the content. This work aims to study the use of *hashtags* for the crowd to describe major events. The rise and acceptance of *hashtags* in major events related to politics, disasters, sports and entertainment were analyzed. Although the broad diversity, we identify some language patterns in the use of *hashtags* created by the population. Moreover, peculiarities were analyzed in each studied event.

RESUMO
As *hashtags* são palavras ou termos que servem para categorizar publicações em mídias sociais, indexando o conteúdo e tornando-o apto para a busca. As *hashtags* são de livre criação, ou seja, não há regras ou restrições para criá-las. Mesmo diante desta liberdade, nota-se uma convergência dos termos usados para explicar e trazer significado para conteúdos. Este trabalho visa estudar o uso das *hashtags* pela multidão para descrever grandes eventos. O surgimento e adesão das *hashtags* em grandes eventos relacionados à política, desastres, esporte e entretenimento foram analisados. Apesar da ampla diversidade, percebeu-se um padrão de linguagem característico no uso das *hashtags* criadas pela população. Além disso, também foram analisadas peculiaridades de cada um dos eventos estudados.

Author Keywords
Social Network Analysis, Zipf Law, Hashtag, Twitter.

INTRODUÇÃO

O Twitter é um *site* de *microblogging* onde os usuários podem publicar mensagens curtas chamadas *tweets*. Além de uma das mais utilizadas mídias sociais, foi o precursor do uso de *hashtags*, recurso que depois foi incorporado em outras mídias sociais.

Mas afinal, o que são *hashtags*? São termos ou expressões que tentam resumir uma mensagem. Em interações sociais baseadas principalmente em textos, as *hastags* tornaram-se – apesar de sua simplicidade - um poderoso recurso nas tecnologias de informação e comunicação *online*. As *hashtags* são usadas para:

- Indexação de conteúdo - *Hashtags* são compostas por termos antecedidos pelo símbolo de tralha (#), tornando-se *hiperlinks* indexáveis pelos mecanismos de busca. Os usuários podem busca-las, acessar todos os conteúdos associados a elas e as *hashtags* mais usadas no Twitter ficam agrupadas no menu *Trending Topics*.
- Sumarização de ideias – As mídias sociais são plataformas naturais para se disseminar ideias e as *hashtags* têm sido utilizadas para representá-las, permitindo o estudo de como as ideias são disseminadas e propagadas [4; 7]
- Expressão de sentimentos e opiniões – *Hashtags* podem estar associadas direta ou indiretamente a sentimentos como alegria, raiva, alívio, susto, medo, dentre outras, auxiliando também no entendimento do evento que gerou tal sentimento (Davidov, 2010). Opiniões (e suas polaridades) também são expressas por *hashtags* [6].
- Contextualização – As *hashtags* podem estar associadas a locais, eventos, sentimentos, ações, datas especiais e pessoas envolvidas, tornando-se um mecanismo de contextualização das mensagens.
- Definição de sarcasmo – A identificação da veracidade da informação é um grande desafio. Neste contexto, encontramos o sarcasmo. Algumas *hashtags* são utilizadas para especificar a veracidade ou grau de sarcasmo de seu conteúdo (i.e, #sqn)

Neste sentido, as *hashtags* são um importante mecanismo para as interações textuais, permitindo enriquecer semanticamente textos usualmente curtos e coloquiais.

Atualmente, tornou-se comum publicar *tweets* sobre acontecimentos e eventos quase em tempo real [3], pois com a utilização das *hashtags* nos *tweets* tornou-se possível resumir o assunto sobre o qual pretendia-se comentar, facilitando a transmissão de informações, cultura e contato com a sociedade. Percebe-se que as *hashtags* surgem pela livre espontânea vontade de cada indivíduo, e mesmo assim algumas são escolhidas pela multidão para descrever o evento. Como se ao longo do tempo, surgisse um vocabulário coletivo, descrito e representado pelas

WAIHCWS'16 was held as part of IHC'16, organized by the Brazilian Computing Society (SBC). October 04, 2016, São Paulo/SP, Brazil.

hashtags. A partir desta observação, tentamos analisar o comportamento do uso das *hashtags* em grandes eventos. Para efetuar este estudo, foi feita uma análise sobre o uso de *hashtags* nos *tweets*, verificando se existem padrões de comportamento ou forma de utilização que possa ser reproduzido para qualquer tipo de evento. Como as *hashtags* são usadas? Quão grande é a adesão da multidão (participantes ou comentaristas de grandes eventos), neste tipo de mídia? Existe um padrão na utilização de *hashtags*? Estas são as questões de pesquisa (QP) que analisaremos neste trabalho.

Analisamos a interação entre usuários em eventos nacionais e internacionais. Para analisar tais cenários, fez-se o uso da Lei de Zipf [9], uma distribuição de probabilidade que descreve boa parte das formas de comunicação escritas ou sonoras. Além disso, também foram analisadas peculiaridades de cada um dos eventos em estudo, como notícias e taxas de surgimento e adesão relacionadas à utilização de *hashtags*. Os resultados obtidos mostram que existe um padrão de linguagem característico na utilização *hashtags* pelos usuários do Twitter.

METODOLOGIA

A metodologia consiste em três etapas: i) coleta dos dados do *Twitter* e criação de uma base de dados brutos; ii) aplicação de filtros sobre a base de dados brutos; e iii) a análise sobre a base de dados já tratada.

Foram escolhidos quatro eventos para serem observados por causarem comoção e levantarem discussões nas redes sociais: Rock in Rio 2013, Eleições presidenciais 2014, Morte do candidato Eduardo Campos e a Estreia da quarta temporada de Game of Thrones. Tais eventos foram selecionados pela representativa e repercussão na época, bem como pela grande quantidade de mensagens trocadas

Coleta de Dados

Os dados coletados se deram por meio do uso de um *crawler*, que obtém os *tweets* acessando a API (*ApplicationProgramming Interface*) oficial de busca do Twitter[1] utilizando palavras pré-definidas. Foram escolhidos quatro eventos para serem abordados e foi criada uma base de dados para cada um.

Bases	*Strings* de busca	Período de Coleta	Quantidade de *Tweets*
Rock in Rio 2013	#rockinrio OR #rockinrio2013 OR #multishownorockinrio OR #multishownorockinrio2013	01/09/13 a 29/09/13	73.605
Eleições Presidenciais 2014	eduardo campos OR marina OR aecio OR dilma OR pastor everaldo OR Luciana Genro OR Levy Fidelix OR Mauro Iasi OR Eymael OR Everaldo Pereira OR Eduardo Jorge OR Jose Maria OR Rui Costa	20/08/14 a 03/11/14	746.794
Morte do candidato Eduardo Campos	"Eduardo Campos"	06/08/14 a 20/08/14	12.921
Game Of Thrones (Quarta temporada)	#got	12/05/14 a 08/08/14	6.652

Tabela 1. Lista de eventos e strings de busca.

Tratamento dos Dados

A base de dados brutos, formada com os dados obtidos no formato JSON, contém campos de dados (como *id*, data de publicação, *hashtags*, *tweet*, etc.). No passo seguinte foi realizado um processamento de todo o conteúdo dessa base. Um código em JavaScript com as funções de MapReduce, aplicado à base contida no MongoDB, foi utilizado para criar combinações entre campos para que fosse possível contabilizar a utilização dessas chaves. A aplicação desses filtros resultou em novas bases contendo apenas os dados que interessavam para as análises realizadas. Para efetuar as diferentes análises foram usadas a linguagem de programação Python e ferramentas de manipulação de planilhas, como tabela dinâmica.

ANÁLISES DOS RESULTADOS

Com os dados devidamente organizados, foi possível fazer as seguintes análises:

- Taxa de Surgimento - quantidade de novas *hashtags* que são criadas em cada dia;
- Taxa de adesão - quantidade de vezes que uma determinada *hashtag* foi usada em cada dia;
- Análise de notícias - relacionadas aos eventos em dias de picos de uso de *hashtags* sobre o assunto.

[1] https://dev.twitter.com/docs/api/1.1, acessado em 07/04/2014.

QP1: Como as *hashtags* se comportam?
Esta observação foi feita com base nas análises das notícias sobre os eventos. Serão utilizadas a Taxa de Surgimento e o Total de *Hashtags* do evento para entender os picos de *hashtags* nos acontecimentos, que justificam o comportamento da população em relação aos eventos. A seguir, nas Figuras 1 e 2, é possível observar tal comportamento em dois eventos: sobre entretenimento e no cenário político brasileiro. Além desses, também serão descritos os outros dois eventos analisados que possuem o mesmo tema.

As imagens abaixo apresentam os gráficos das taxas de surgimento do evento Rock in Rio 2013 e Morte do candidato Eduardo Campos.

Percebe-se que 57% de todas as *hashtags* referentes ao Rock in Rio 2013 surgiram na semana do dia 11 ao dia 17 de setembro, semana que precedeu ao show da cantora norte-americana Beyoncé, que foi dia 14 de setembro. Outros dois dias significativos foram os *shows* da banda Metallica e do cantor Bon Jovi que se apresentaram no palco principal do evento. Estes artistas foram mencionados entre as dez principais *hashtags* utilizadas do evento, representando aproximadamente 2% das de todas as *hashtags* publicadas pelos usuários durante todo o evento e aparecem entre as dez principais do evento, como iremos ver mais a frente detalhadamente análise da taxa de adesão. Esta análise é interessante para que se possa observar quais foram os dias o evento mais acompanhados e comentados pelos fãs nas mídias sociais.

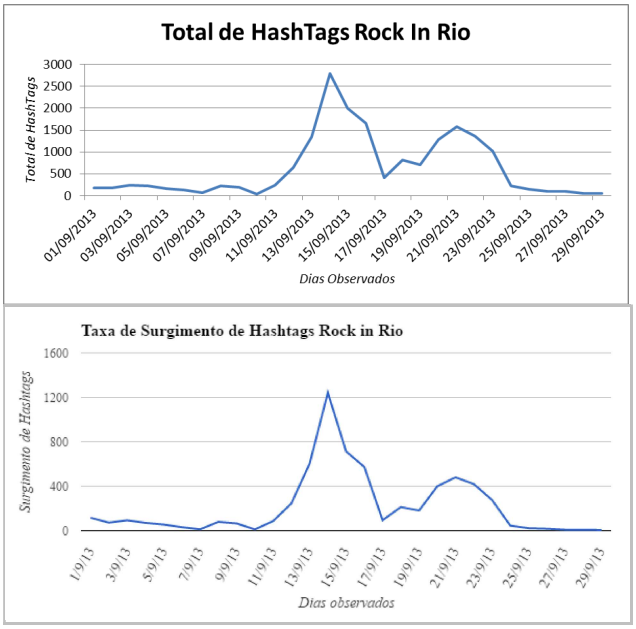

Figura 1. Análise de padrões entre Taxa de Surgimento e Total de Hashtags do Rock in Rio 2013.

Observa-se uma coleta prévia a data do acidente, sendo possível identificar o exato momento da explosão de *hashtags*, representado 72% de todas as *hashtags* referentes à coleta sobre a morte de Eduardo Campos surgiram entre os dias 12 e 15 de agosto. Tal fato deve-se a morte do candidato, ocorrida no dia 13 de agosto, por um desastre aéreo. Também houve uma leve subida no gráfico no dia 17 de agosto, data de seu sepultamento.

No pico do dia 13 de agosto mais de 2000 *Hashtags* relativas ao evento foram criadas. Isso é um indício de que a população acompanhou e se chocou com o ocorrido.

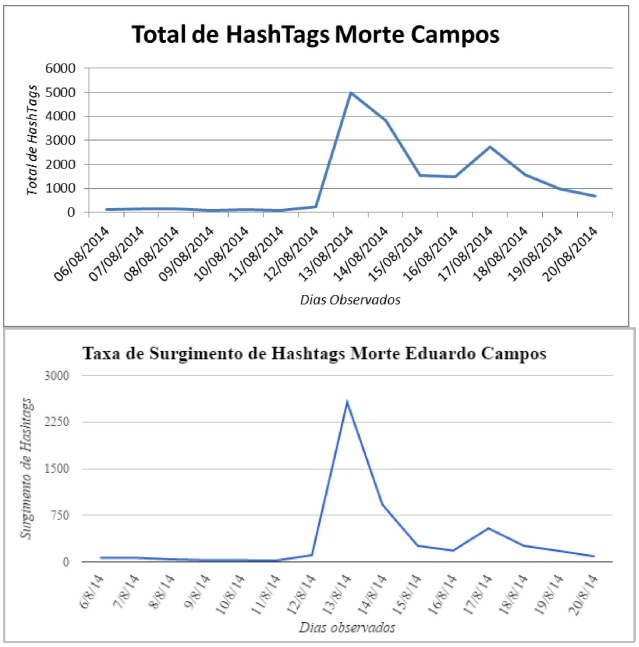

Figura 2. Análise de padrões entre Taxa de Surgimento e Total de Hashtags da morte de Eduardo Campos.

Pode-se notar, comparando os gráficos das taxas de surgimento e total de *hashtags* de um mesmo evento, que eles seguem padrões semelhantes. Esse fenômeno ocorreu em todos os eventos observados. É interessante perceber que existe uma relação forte entre o total de *hashtags* usadas sobre determinado evento e o surgimento de novas *hashtags*.

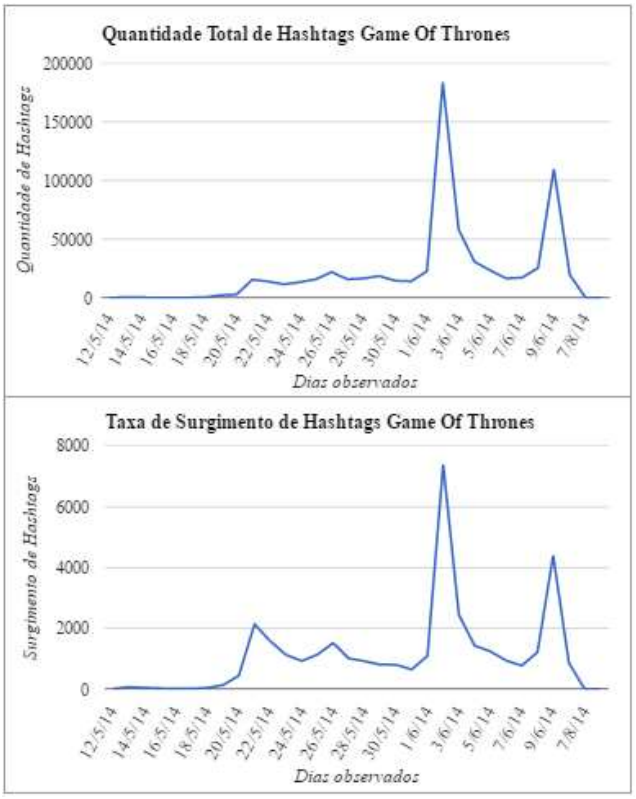

Figura 3. Análise de padrões entre Taxa de Surgimento e Total de Hashtags de Game of Thrones.

Na figura 3 observa-se que na quarta temporada de Game of Thrones há duas ocorrências de picos de publicações, bem como de surgimento de novas *hashtags.* Elas ocorretam nos dias seguintes aos últimos episódios da temporada. A grande comoção na rede social foi devido à morte de um personagem, entre as principais *hashtags* utilizadas continham o nome do personagem e a insatisfação das pessoas. 39% de todas as hashtags relacionadas a Game of Thrones surgiram entre o dia primeiro e o dia cinco de junho.

A seguir, observa-se o gráfico de outro cenário, este relacionado a política brasileira. Os dados sobre a opinião dos usuários sobre as Eleições Presidenciais Brasileiras de 2014. Apesar de todos os gráficos seguirem um padrão de surgimento similar ao de utilização de modo geral das *hashtags*, este possui algumas observações simples sobre determinados períodos.

Observa-se, neste acompanhamento em longo prazo, alguns picos na figura abaixo. Analisando o intervalo entre os dias 20 e 28 de agosto percebe-se a diferença entre os dois gráficos, pois no segundo caso há picos, enquanto no primeiro o desenvolvimento do gráfico está normal. Dado que, o dia 21 de agosto foi o início do horário eleitoral, houve pouca publicação, entretanto a variedade de *hashtags* criadas fez com que o gráfico da taxa de surgimento se destacasse.

Os picos do dia 27 de agosto se referem ao debate presidencial ocorrido em uma emissora de televisão, bem como os dias 3 e 14 de outubro. É possível verificar, observando apenas as dez primeiras *hashtags* utilizadas, em qual emissora foi o debate e qual era a opinião do público em relação aos candidatos.

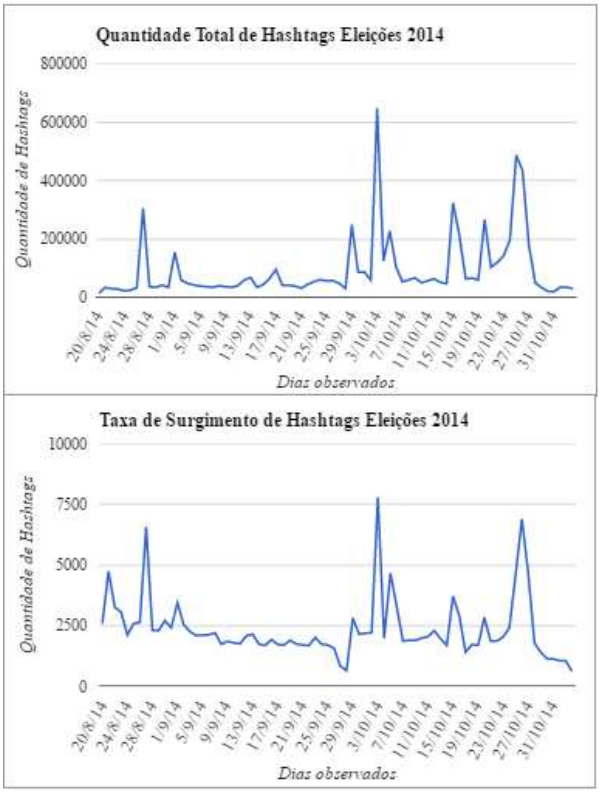

Figura 4. Análise de padrões entre Taxa de Surgimento e Total de Hashtags das Eleições Presidenciais Brasileiras de 2014.

Nos dias 5 e 26 de outubro, dias de votação, também houve grande participação popular. Entretanto, é possível verificar que há significativa participação nas mídias sociais em torno deste assunto um dia antes e um dia depois do pico.

QP2: Quão grande é a adesão da multidão?

Taxa de adesão é a quantidade de vezes que cada *hashtag* é utilizada em cada dia. Mostra o quanto os usuários aderiram aquela *hashtag.*

Na imagem a seguir, figura 5, está a distribuição da taxa de adesão das dez *hashtags* mais populares do evento Rock in Rio 2013. Para a análise das top 10, foi necessário ocultar do gráfico a *hashtag "#rockinrio"*, devido sua frequência ser aproximadamente dez vezes mais do que a segunda *hashtag* mais utilizada no evento, o que dificultaria a visualização deste comparativo.

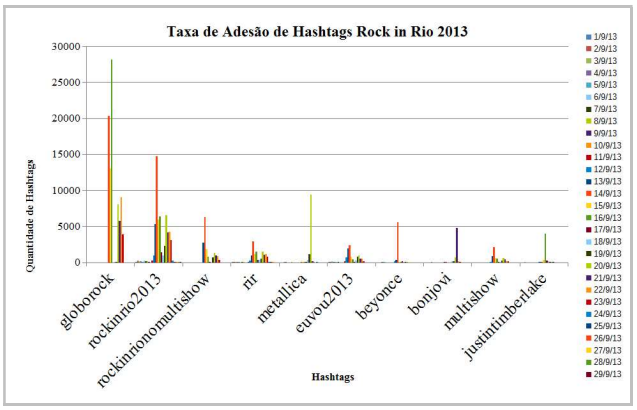

Figura 5. Taxa de adesão de hashtags sobre o Rock in Rio 2013.

Esta análise permite observar quais atrações foram as mais comentadas do evento e os dias em que elas ficaram em evidência, em geral, foram no dia do evento e no dia seguinte ao show dos artistas mais aguardados pelo público.

Além disso, as *hashtags* sugeridas pelos meios de comunicação, que transmitiram o evento ao vivo, também surgiram entre as mais citadas no evento.

A Figura 6 apresenta taxa de adesão das *hashtags* relacionadas à quarta temporada de Game of Thrones. Nota-se que a *hashtag* "themontaindandtheviper" se destaca no dia 2 de junho. Ela mostra a expectativa do público em relação a um evento prestes a acontecer na série durante aquela semana.

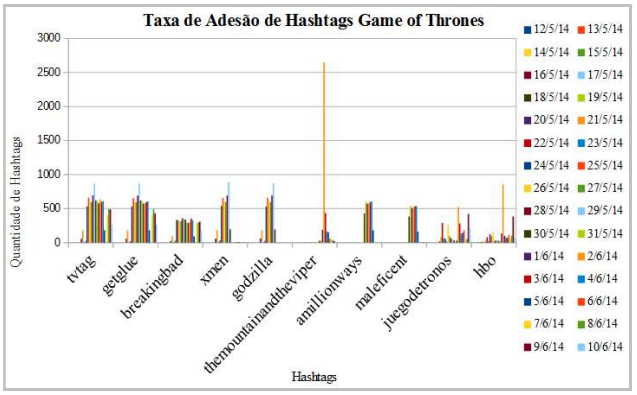

Figura 6. Taxa de adesão de hashtags sobre Game of Thrones.

Na Figura 7 Nota-se que a maioria das *hashtags* foi muito utilizada no dia 13 de agosto – data do falecimento do candidato - e decresceu ao longo do período analisado. Esse comportamento é típico de um acontecimento inesperado que choca a população e coloca o tema em foco instantaneamente.

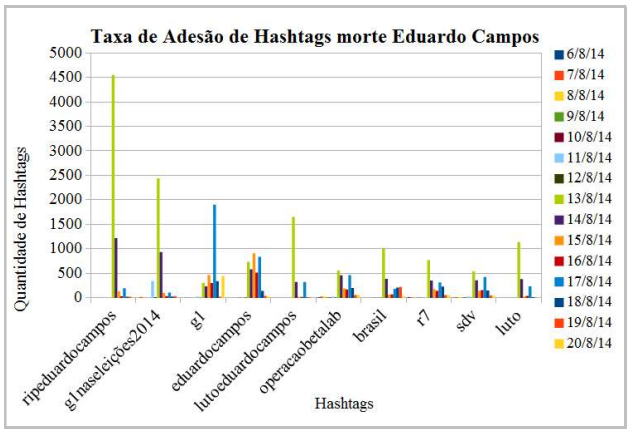

Figura 7. Taxa de adesão de hashtags sobre a morte de Eduardo Campos.

É possível perceber através das *hashtags* utilizadas no evento das eleições 2014 que a população acompanhou e participou, através das mídias sociais, dos debates pelas principais emissoras de televisão aberta, que são citadas entre as principais *hashtags* do evento.

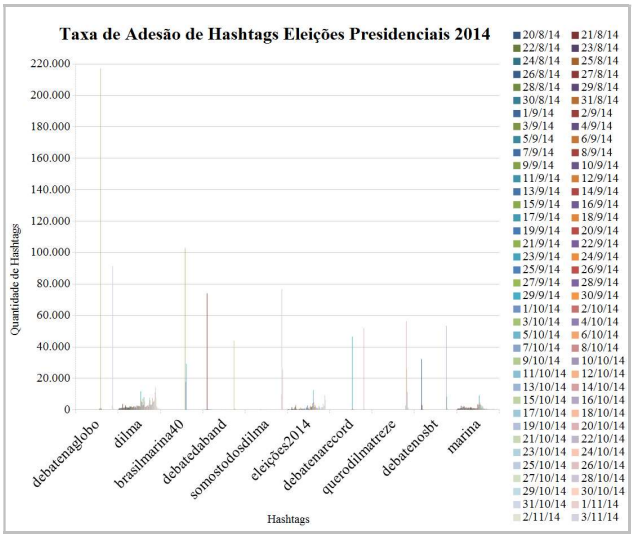

Figura 8. Taxa de adesão de hashtags sobre as Eleições Presidenciais de 2014.

Percebe-se que Dilma Rousseff e Marina Silva foram as únicas candidatas que apareceram no top 10 deste evento. Entretanto, a disputa no dia das votações não seguiu esta tendência, que teve como principais concorrentes a candidata, e futuramente eleita, Dilma Rousseff seguida pelo candidato Aécio Neves, que seguiram para votações em segundo turno.

QP3: Existe algum padrão na utilização de *hashtags*?

Para verificar se existem padrões na comunicação através de *hashtags* neste trabalho será utilizada a Lei de Zipf, que estuda a relação entre o posicionamento de uma palavra no *ranking* de frequência em um texto suficientemente grande

e a quantidade de vezes que ela aparece. A relação é dada pela lei de potência:

$$f(k)=C/k^a$$

Sendo 'k' a posição da palavra no ranking, 'c' a quantidade de vezes que a primeira palavra do ranking aparece e 'a' uma constante positiva. A análise abaixo mostra graficamente a distribuição das *hashtags* dos diferentes eventos.

Uma maneira visual de reconhecer a distribuição de Zipf é notar que os dados ordenados colocados na escala Log x Log se aproximam de uma reta cuja inclinação é igual '-a'.

O objetivo desta análise é verificar se a comunicação via *hashtags* segue o mesmo padrão dos textos escritos em português e em muitas outras línguas.

A seguir, observa-se a tabela que representa o cálculo, com base na Lei de Zipf, dos eventos analisados e suas respectivas leis de distribuição de *hashtags* fixando a mais citada de cada um dos temas e a constate positiva. Com base em dados parciais sobre uma tendência linear.

Evento	Distribuição aproximada
Rock in Rio 2013	f(k)=879.590/k^1,03
Quarta temporada de Game of Thrones	f(k)=371.426/k^1,12
Eleições Presidenciais 2014	f(k)=316.085/k^1,33
Morte de Eduardo Campos	f(k)=3.554/k^1,21

Tabela 2. Eventos e suas leis de distribuição de hashtags.

Abaixo, na figura 9, observa-se o resultado da análise referente ao evento Rock in Rio 2013. Pode-se perceber que a primeira *hashtag* se distancia das demais, pois a segunda *hashtag* com maior adesão representa apenas 10% da frequência da primeira mais utilizada "#rockinrio", definida pelo próprio evento e utilizada por usuários, artistas e emissoras de Rádio/TV.

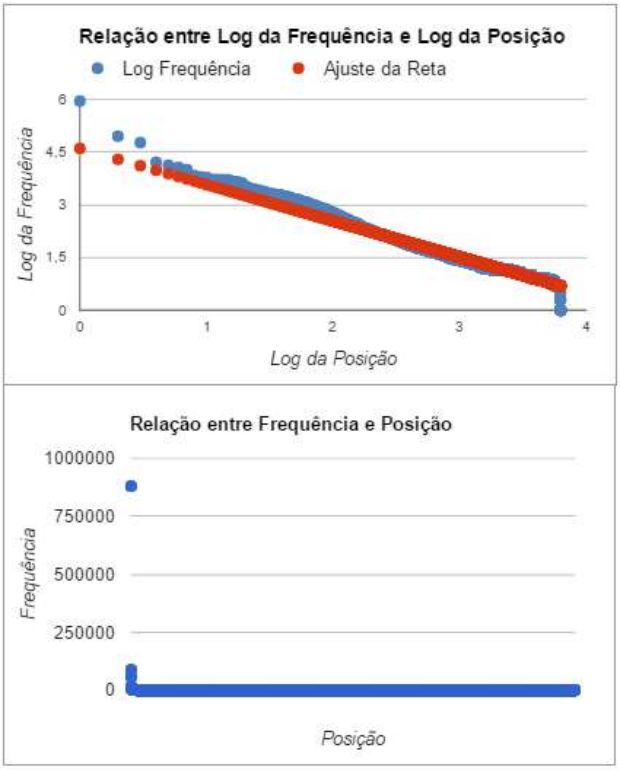

Figura 9. Análises da Lei de Zipf sobre o Rock in Rio 2013.

O ajuste da reta, representado em vermelho nos gráficos, foi calculado através de dados parciais de uma equação linear. Onde o primeiro elemento é dado a partir dos resultados do Log da Posição e a constante '-a' e o segundo é o resultado da subtração do erro dados pelo ajuste linear e o primeiro elemento.

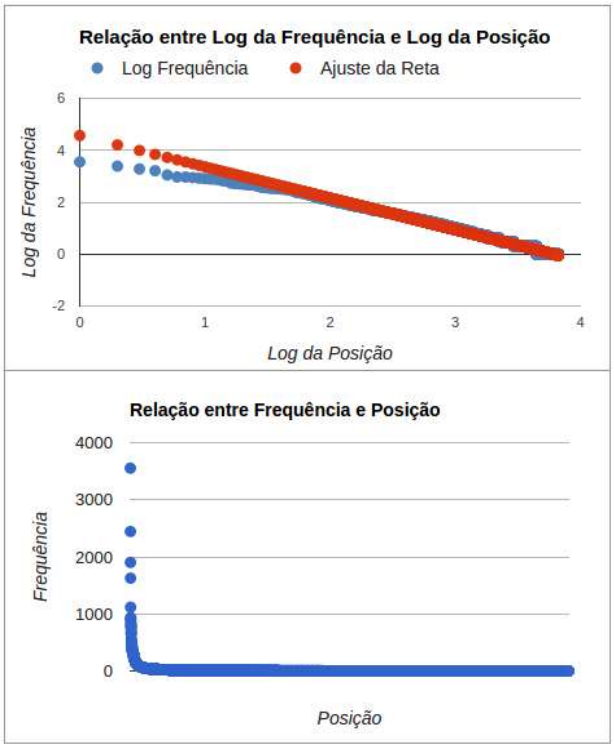

Figura 10. Análises da Lei de Zipf sobre a Morte de Eduardo Campos.

A figura 10 mostra a análise das *Hashtags* sobre a morte de Eduardo Campos. Percebe-se que as *hashtags* localizadas nas primeiras posições do *ranking* destoam um pouco do ajuste linear no gráfico LogxLog. Essa diferença já é conhecida no experimento de Zipf. Espera-se que as primeiras e últimas palavras analisadas fujam do padrão. Entretanto, a partir de um determinado ponto a lei de formação é seguida a risca.

TRABALHOS RELACIONADOS

As leis de potência são encontradas em áreas bem diversas: economia e finanças, educação, demografia, geologia, história, climatologia, bibliometria e informetria, terrorismo e guerras, corrupção, turismo, esportes, artes, agronomia, ecologia, biologia, linguística, ciência da computação, ciências cognitivas, ciências sociais, astronomia, mecânica dos sólidos, física e química [1]. A lei de potência usada nas linguagens escritas e sonoras é conhecida como Lei de Zipf, que foi usada neste trabalho para analisar as tags relacionadas a eventos populares.

Na lei de Zipf observa-se que, num texto suficientemente longo, existe uma relação entre a frequência que uma dada palavra ocorria e sua posição na lista de palavras ordenadas segundo sua frequência de ocorrência [2]. Uma lista é confeccionada, levando-se em conta a frequência decrescente de ocorrências. À posição nesta lista dá-se o nome de ordem de série (rank) como foi utilizado para as análises deste trabalho. Assim, a palavra de maior frequência de ocorrência tem ordem de série 1, a de segunda maior frequência de ocorrência, ordem de série 2 e, assim, sucessivamente.

CONSIDERAÇÕES FINAIS

Este estudo analisou o comportamento das *hashtags* criadas em 5 grandes eventos. Neste estudo, analisamos se as *hashtags* possuíam comportamentos linguísticos como o surgimento de novos termos mediante novos eventos (taxa de surgimento), além da fixação e popularidade de termos (taxa de adesão). A partir da análise destes dois critérios, percebemos o comportamento de inovação linguística, que pode ser descrito como qualquer mudança em um sistema linguístico existente [8].

Outro comportamento linguístico analisado foi a obediência ao padrão de Zipf. A lei de Zipf é uma teoria com base na matemática e linguística que analisa e quantifica como as palavras são distribuídas dentro de um determinado texto. Zipf analisou a comunicação escrita em vários textos em inglês e verificou que as palavras seguem uma lei de potência. Tal estudo foi replicado por vários estudiosos, em diferentes línguas, em textos de diferentes conotações (músicas, textos científicos, poesias e romances) e chegou-se à mesma conclusão: a comunicação humana segue o princípio de Zipf. Maia [10] realizou uma das primeiras análises para o português-brasileiro. Por este motivo, escolhemos a Lei de Zipf para verificar se *hashtags* poderiam se comportar como um vocabulário, o que foi mostrado que sim.

Até a presente data, não ocorreram estudos que analisassem o comportamento linguístico das *hashtags* em grandes eventos. Este artigo conduziu e descreveu um estudo inédito para verificar se, mesmo sendo criadas individualmente, o conjunto de *hashtags* usadas em grandes eventos teriam comportamentos linguísticos, seguindo um padrão semelhante ao dos idiomas humanos.

Este é um estudo preliminar. Como trabalhos futuros, pretendemos analisar uma quantidade maior de eventos e modelar o comportamento de surgimento, uso e desatualização das *hashtags*.

AGRADECIMENTOS

Agradecimento ao CNPQ, FAPERJ, Fabrício Firmino, Fabio Rangel que contribuíram para a evolução deste trabalho.

REFERÊNCIAS BIBLIOGRÁFICAS

1. Bortolossi, H. J., Queiroz, J. J. D. B., & da Silva, M. M., (2012), A Lei de Zipf e Outras Leis de Potência em Dados Empíricos.
2. Guedes, V. L. and Borschiver, S. (2013), Bibliometria: Uma Ferramenta Estatística Para a Gestão da Informação e do Conhecimento, Em Sistemas De Informação, de Comunicação e de Avaliação Científica e Tecnológica

3. Naaman, C.-H. L. Mor., and Boase, J. (2010), "Is it all About Me? User Content in Social Awareness Streams", ACM Conference on Computer Supported Cooperative Work, 2010.
4. Tsur, Oren, and Ari Rappoport (2012). "What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities." Proceedings of the fifth ACM international conference on Web search and data mining. ACM.
5. Davidov, Dmitry, Oren Tsur, and Ari Rappoport (2010). "Enhanced sentiment learning using twitter hashtags and smileys." Proceedings of the 23rd international conference on computational linguistics: posters. Association for Computational Linguistics, 2010.
6. Barbosa, Glívia AR, et al. (2013)"Caracterização do uso de hashtags do Twitter para mensurar o sentimento da população online: Um estudo de caso nas Eleições Presidenciais dos EUA em 2012." Simpósio Brasileiro de Banco de Dados: Recife.
7. Wang, Yazhe, and Baihua Zheng (2014). "On macro and micro exploration of hashtag diffusion in Twitter." Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on. IEEE.
8. Breivik, L.E., and Jahr, E.H. (Eds.) 1989. Language change: Contributions to the study of its causes. Berlin/New York: Mouton de Gruyter.
9. Zipf, George Kingsley. Human behavior and the principle of least effort: An introduction to human ecology. Ravenio Books, 2016 (reprinted).
10. MAIA, E. de L. Q S. Comportamento bibliométrico da língua portuguesa como veículo de representação da informação. Ciência da Informação, Rio de Janeiro, 2(2):99-138, 1973. (Dissertação de Mestrado de 1973).