ESTUDO DE CASO SOBRE COLETA E ANÁLISE DE DADOS DO TWITTER

Resumo
Com o objetivo de demonstrar a importância de ferramentas computacionais no ambiente acadêmico, destacando sua contribuição para as várias áreas do conhecimento e da pesquisa científica, o presente artigo utiliza mecanismos de coleta, tratamento e análise de dados públicos da rede social Twitter, baseando-se num estudo de caso sobre as Eleições Municipais de 2016 ocorridas nas cidades de São Paulo e do Grande ABC paulista. A rápida popularização das Redes Sociais de Internet tem produzido importantes transformações nas relações sociais, criando novas formas de sociabilidade mediadas por computadores. Nesse novo espaço de interação, diversos temas são debatidos por seus usuários, que possuem um papel ativo na produção, transmissão e circulação de informações, tornando o processo comunicacional mais fluído e dinâmico. Para tanto, foram feitas análises de cunho estatístico com os dados obtidos, comparando-os com informações do Tribunal Superior Eleitoral (TSE) e destacando padrões de interação entre os usuários da rede, bem como atores mais influentes. O método proposto demonstrou possuir limitações, porém mostrou-se potencialmente auxiliador no que se diz respeito ao avanço de pesquisas e descobertas acadêmicas em diversas áreas do conhecimento.

Palavras-Chave: Ciência das Redes, Twitter, Redes Sociais de Internet, Eleições Municipais.

TWITTER DATA COLLECTION AND ANALYSIS CASE STUDY

Abstract
In order to demonstrate the importance of computational tools in academia, highlighting your contribution to the various areas of knowledge and of scientific research, this project uses mechanisms of collection, treatment and analysis of Twitter social network public data, based on a case study of the municipal elections in 2016 that occurred in the cities of São Paulo and "Grande ABC paulista". The rapid popularization of Internet Social Networks has produced important transformations in social relations, creating new forms of computer-mediated sociality. In this new interaction space, various topics are discussed by its users, who have an active role in the production, transmission and circulation of information, making the communicational process more fluid and dynamic. To this end, statistical measures analyses were made with the data obtained, comparing them with information from the Superior Electoral Court (TSE) and highlighting patterns of interaction between network users, as well as most influential actors. The proposed method showed limitations, but it proved to be potentially helpful with regard to the advancement of research and academic discoveries in various areas of knowledge.

Keywords: Internet Social Networks, Twitter, Network Science, Municipal Elections.

1 INTRODUÇÃO

A evolução da internet e a expansão de sua acessibilidade foram responsáveis pelo rápido crescimento e propagação das redes sociais virtuais. Em poucos anos, o ciberespaço tornou-se um ambiente de interação indispensável, sem o empecilho da distância; rápido, fácil e, muitas vezes, anônimo. As redes sociais de Internet ganharam espaço elementar na sociedade contemporânea, disseminando informações de maneira veloz e criando um ambiente virtual de debate e acesso aos mais diversos assuntos (RECUERO, 2016).

Nos últimos anos, as redes sociais virtuais, como o Twitter e o Facebook, tomaram um papel notável em diversos eventos sociopolíticos (SANTOS, 2014). A Primavera Árabe, por exemplo, conhecida como uma onda revolucionária de manifestações públicas ocorridas no Oriente Médio e no Norte da África entre 2010 e 2011, teve suas bases fomentadas pelas redes sociais (LOPES, 2013). Os ativistas se utilizaram do benefício da comunicação veloz e abrangente que a Internet proporciona para organizar os primeiros protestos, bem como espalhar seus ideais e reinvindicações mundo a fora. Outro exemplo, mais próximo à realidade nacional e mais recente, é o processo de Impeachment da presidente Dilma Rousseff, que se sucedeu por uma sequência de manifestações populares em 2015 e 2016: as redes sociais foram usadas como plataforma de disseminação de informações e opiniões, influenciando desde a eleição presidencial de 2014, na qual Rousseff foi reeleita, até o conflito daqueles que defendiam o impeachment *contra* aqueles que não o apoiavam (BERTOL et al, 2011; CARVALHO et al, 2016 e PENTEADO et al, 2014).

Desde as eleições presidenciais americanas onde Barack Obama foi eleito, a participação das redes sociais no contexto político tem tomado caráter decisivo no rumo e desfecho dos eventos (GOMES et al, 2009). No Brasil, o Twitter, objeto de estudo deste artigo, já é utilizado como veículo de disseminação para campanhas eleitorais (BACHINI, 2013).

Assim, nota-se que a rápida popularização das Redes Sociais de Internet tem produzido importantes transformações nas relações sociais, criando novas formas de sociabilidade mediadas por computadores. Nesse sentido, se faz necessário um estudo do comportamento dos usuários destas redes e de seus relacionamentos, tomando como base a rede social Twitter e sua ampla base de dados acessível, com auxílio das técnicas da Ciência da Computação (BENEVENUTO et al, 2011 e RECUERO, 2014). Isto é, o uso de ferramentas computacionais cria a possibilidade do desenvolvimento de novas formas de estudos, nas diferentes áreas de conhecimento social, sobre o comportamento dos usuários de redes sociais (VALIATI et al, 2013).

O Twitter é uma Rede Social de Internet baseada em *microblogs*, que consistem em atualizações pessoais feitas pelos usuários, nas quais se pode publicar informações e ideias livremente, porém com restrição de 140 caracteres por postagem. Essas mensagens são denominadas *tweets*. Os *tweets* são apresentados no perfil do usuário que os publica e também enviados para a *timeline* (linha do tempo) dos seus seguidores, estabelecendo, assim, uma relação interpessoal entre os participantes. As postagens podem ser rotuladas por meio de *hashtags*, que, de forma sintética, rotulam assuntos e opiniões contidas nestes *tweets*. (RECUERO, 2015 e VALIATI et al, 2013). De acordo com pesquisa realizada em 2015 pela GlobalWebIndex, que é um mecanismo de pesquisa digital de dados públicos para o mercado, o Twitter está entre as 10 redes sociais mais utilizadas no mundo todo (FARIA, 2015). Em agosto de 2017, estimava-se que havia cerca de 328 milhões de usuários ativos registrados (disponível em: https://about.twitter.com/pt/company). No Brasil, de acordo com pesquisa da comScore de 2015, cerca de 9 milhões de pessoas utilizam regularmente esta rede (BANKS, 2015). Portanto, conduzir a pesquisa tomando o Twitter como fonte de dados proporcionará uma parcela da população a partir da qual poderão ser efetuadas análises estatísticas

interessantes, identificando atores relevantes, termos e *hashtags* mais utilizados no espaço de tempo selecionado e padrões de interação (BRUNS; BURGUESS, 2012).

Desta forma, com a finalidade de demonstrar a importância de ferramentas computacionais no ambiente científico, explicitando sua notável contribuição para este meio de forma interdisciplinar, este estudo está estruturado na utilização de mecanismos de coleta, tratamento e análise de dados públicos do Twitter, baseando-se num estudo de caso sobre as Eleições Municipais de 2016 ocorridas nas cidades de São Paulo e do Grande ABC paulista. Um evento político como esse promove a produção de uma grande quantidade de dados públicos na rede. Tais dados podem ser analisados de diversas maneiras, de forma a abranger diversos interesses de estudo. Assim, os estudos efetuados podem fornecer respostas, apresentar atores virtuais influentes na rede, identificar tendências de preferência entre os candidatos e até mesmo prever acontecimentos ou o curso das eleições.

2 FERRAMENTAS E MECANISMOS

2.1 Programação básica e Python

A programação de computadores é, em termos gerais, um processo de criação de algoritmos. Define-se *script* (ou código de programação) como um algoritmo, que nada mais é do que um conjunto de instruções ordenadas, referente a uma tarefa automatizada, executado pelo computador. A partir da programação, é possível realizar operações complexas, que geralmente envolvem muitos cálculos e uma grande quantidade de dados, variáveis e possibilidades, sendo, desta forma, tarefas complicadas e demoradas para serem executadas sem o auxílio de máquinas. Assim, os *scripts*, em conjunto com outros elementos computacionais, tais como interface de usuário e banco de dados, auxiliam empresas, funcionários e demais entidades em diversos processos, visando facilitar e potencializar a produtividade em prol de um determinado objetivo (KOLIVER et al, 2004; SANTOS et al, 2005; SANTOS et al, 2006).

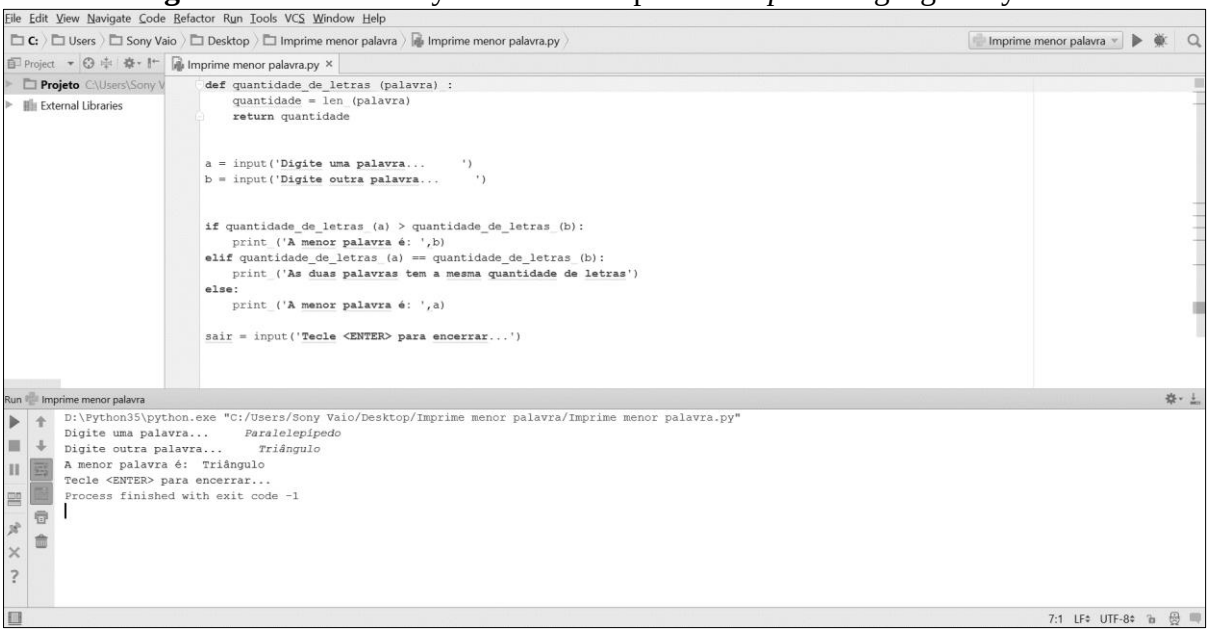
Para se construir um *script*, isto é, para se programar, é necessário fazer uso de uma linguagem de programação. Essas linguagens são a maneira padronizada para se fazer a comunicação entre pessoa e máquina, ou seja, para transmitir informação e instruções. Cada linguagem possui um conjunto de regras sintáticas diferente, que possibilita ao programador repassar ao computador as informações exclusivamente através de palavras predefinidas, números, expressões lógicas e expressões matemáticas (LAUER, 2008).

Neste projeto, os *scripts* foram criados através da linguagem de programação Python, que possui uma sintaxe clara e concisa. Tal linguagem vem crescendo no ambiente acadêmico, ganhando espaço na comunidade científica, já que é uma linguagem que gerencia diversas tarefas computacionais secundárias automaticamente, poupando o pesquisador científico de demais funções que não a criação de *scripts*. De modo geral, essa linguagem pode ser aplicada em desenvolvimento web e de jogos, bem como matemática computacional, bioinformática e muitas outras áreas, evidenciando seu abrangente campo de funcionalidade (COELHO, 2007). De acordo com o Índice TIOBE[1], que é um *ranking* das linguagens de programação mais utilizadas atualmente, Python ocupava, em agosto de 2017, a quinta posição. Grandes empresas de Tecnologia da Informação também utilizam essa linguagem para gerenciamento de suas funções, tais como Google e grande parte das distribuições do sistema operacional Linux.

Para auxílio do processo de programação, foi utilizado neste artigo o ambiente PyCharm Community: um software que proporciona ferramentas visuais e interativas para facilitar a utilização a programação de *scripts*.

[1] Disponível em: http://www.tiobe.com/tiobe-index/

Figura 1 - Ambiente PyCharm e exemplo de *script* na linguagem Python



Fonte: Elaborada pelos autores em 07 jan. 2017.

A figura 1 representa um exemplo de *script* elaborado com a linguagem Python utilizando o *software* PyCharm.

Além de seu uso facilitado e sintaxe clara, a linguagem Python oferece o recurso de importação de bibliotecas. Este recurso permite que o programador tenha acesso a uma série de funções predefinidas para determinados fins, utilizando-as de acordo com sua necessidade. Por exemplo, se o programador cria *scripts* para a área matemática, é recomendável o uso da biblioteca "math", que insere funções à linguagem, tais como raiz quadrada (sqrt), e logaritmo (log), tornando mais conveniente o processo de programação. Não obstante, Python é conhecida por disponibilizar uma extensa gama de bibliotecas-padrão e por possibilitar a instalação e utilização de bibliotecas desenvolvidas por terceiros, como é o caso da *Twython* e da *PyMongo*, bibliotecas cruciais para o desenvolvimento deste projeto, descritas nas próximas subseções.

2.2 API do Twitter e biblioteca Twython

Uma API (Application Programming Interface) é uma biblioteca que possibilita o acesso às funcionalidades de um determinado software ou servidor, para que possam ser utilizadas por aplicativos desenvolvidos por terceiros. Dentre essas funcionalidades, é possível acessar bancos de dados e outros recursos específicos.

A rede social Twitter, escolhida como fornecedora de dados para este projeto, oferece acesso gratuito a uma parcela de seu banco de dados através de sua API, de forma a disponibilizar um fluxo limitado a 1% da totalidade dos tweets publicados em toda a rede a cada momento, possibilitando a obtenção de dados públicos para análise (MAKICE, 2009).

Para utilizar a API do Twitter por meio da linguagem Python, é mais fácil fazer uso da biblioteca Twython, criada por Ryan Mcgrath. Essa biblioteca é direcionada especificamente para acessar o *database* da rede social e efetuar coleta de dados. Basicamente, existem dois tipos de acesso a essa API (MACGRATH[1], 2013):

1) Streaming API: funciona de maneira sincronizada com o servidor do Twitter, isto é, dadas especificas palavras-chave (*keywords*), este mecanismo irá coletar *tweets* logo após forem postados na rede e os retornará para o programador, garantindo a obtenção de dados em tempo real. Assim, este tipo de acesso à API configura uma

conexão constante, retornando *tweets* conforme são publicados por usuários. Não é possível obter dados publicados antes da execução do *script.*

2) API Rest: funciona de maneira a resgatar *tweets* postados em momentos anteriores à execução do *script*, isto é, ela permite acesso a dados de datas anteriores. Dadas especificas *keywords*, este mecanismo irá retornar *tweets* que já estão no banco de dados do Twitter, diferenciando-se da Streaming API por não configurar uma relação de sincronia e, portanto, não sendo possível resgatar *tweets* postados após a execução do algoritmo.

Neste artigo, como o objetivo foi efetuar um estudo de caso com base nas Eleições de 2016 em tempo real, a Streaming API é o modelo que melhor se encaixa como ferramenta no desenvolvimento do trabalho. No entanto, tal aplicação não significa inibição completa do segundo tipo de acesso à API, que foi eventualmente utilizada para resgatar pequenos espaços de tempo perdidos pela Streaming.

2.2.1 Obtendo Permissões e Chaves de Acesso

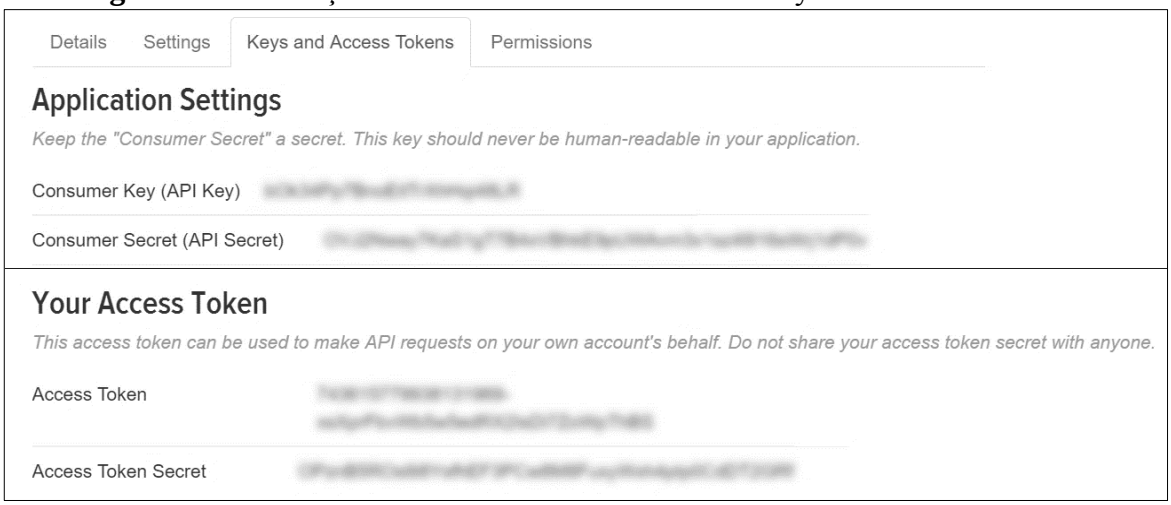
Para começar a utilizar a API, são necessários alguns pré-requisitos. Um deles é a permissão de acesso, concedida pelo próprio sistema do Twitter. Para obter esta permissão, é necessário gerar chaves de acesso através do próprio site da Rede Social, tais quais:

1) API Key: chave de identificação do aplicativo.
2) API Secret: código utilizado para confirmar a autenticidade da API Key.
3) Access Token: Chave de acesso que possibilita ao serviço da API verificar as permissões de acesso que o aplicativo possui. Esta chave age apenas depois da validação da API Key.
4) Token Secret: código utilizado para confirmar a autenticidade do Access Token.

Inicialmente, deve-se possuir uma conta na Rede Social, na qual se poderá registrar aplicativos no modo desenvolvedor (https://apps.twitter.com). Acessando o endereço citado, basta selecionar "Create new app", registrando, assim, um novo aplicativo. Após preenchimento de dados básicos, é necessário acessar a aba "Keys and Access Tokens", selecionando o botão "Create my access token" para gerar o Access Token e o Token Secret. A API Key e o API Secret são gerados automaticamente, e podem ser verificados na mesma aba (MACGRATH[2], 2013; TWITTER, 2017).

Tais dados de acesso são exclusivos para cada usuário e, portanto, privativos. Desta forma, nas figuras 2 e 3, apresentadas a seguir, estes dados foram camuflados.

Figura 2: Localização das chaves de acesso na aba "Keys and Access Tokens"



Fonte: Twitter, 2016. Disponível em: https://apps.twitter.com. Acesso em: 10 jan. 2017.

A figura 2 demonstra a localização das chaves citadas anteriormente, disponíveis no próprio endereço eletrônico da rede (especificado na fonte).

2.2.2 Autenticando o s*cript* à API do Twitter

Após autorização de acesso concedida pelo Twitter e possuindo as chaves de acesso, é possível fazer a autenticação do *script* à API da Rede Social através da biblioteca Twython.

Figura 3: Autenticação do *script* nos dois tipos de acesso à API

```
#--------Autenticação em Streaming API--------#

from twython import TwythonStreamer

apk = [illegible]
aps = [illegible]
oat = [illegible]
ots = [illegible]

stream = TwythonStreamer(apk, aps, oat, ots)

#--------Autenticação em API Rest--------#

from twython import Twython

apk = [illegible]
aps = [illegible]
oat = [illegible]
ots = [illegible]

twitter = Twython(apk, aps, oat, ots)
```

Fonte: Elaborada pelos autores em 10 jan. 2017.

A figura 3 mostra como é feita a autenticação nos dois tipos de acesso à API utilizando a biblioteca Twython. No algoritmo retratado, considera-se as chaves: "apk" para API Key; "aps" para API Secret; "oat" para Access Token e "ots" pra Token Secret. A autenticação é o primeiro passo na construção do *script* utilizado para coleta de dados da Rede Social neste projeto, sendo inteiramente baseada nas chaves de acesso, que devem ser obtidas corretamente.

2.3 Bases NoSQL, MongoDB e a biblioteca PyMongo

É imprescindível que haja uma forma de armazenamento que possibilite acesso, resgate e tratamento da grande massa de dados obtida com a coleta. Além disso, essa forma de armazenamento deve ser eficaz, de modo que o banco de dados gerado seja facilmente modificável e gerenciável. Para tanto, neste projeto, utilizou-se a tecnologia NoSQL[2], que é um tipo de banco de dados *não-relacional*, isto é, ele armazena as informações de modo que não há relação de dependência entre dados de mesmo tipo. Isso significa que os *tweets* armazenados são totalmente independentes um dos outros. Uma base NoSQL é mais eficaz, rápida e dinâmica

[2] Outro tipo de tecnologia de armazenamento muito utilizada é o SQL. Diferente do utilizado, essa tecnologia é do tipo relacional, ou seja, armazena os dados criando relações de hierarquia entre eles. É uma forma mais segura e protegida, mas com liberdade computacional inferior para manipulação e resgate de dados se comparado ao modelo NoSQL.

para organização e análise de grandes massas de dados, sendo, assim, um mecanismo condizente como ferramenta neste projeto.

Para efeito de auxílio, foi escolhida a plataforma MongoDB, um aplicativo de banco de dados do tipo NoSQL que utiliza padrões de organização do tipo JSON; padrões estes que correspondem a estruturas chamadas "dicionários" em linguagem Python. Assim, em combinação com essa linguagem, o MongoDB é uma ferramenta poderosa para se fazer análises de maneira facilitada, sem haver necessidade de converter os dados armazenados em outro tipo de organização (VIEIRA et al, 2012; GOMES et al, 2015).

Para introduzir o MongoDB ao *script*, é necessário fazer uso da biblioteca PyMongo, criada por Mike Dirolf e mantida por Bernie Hackett. Essa biblioteca tem a função de conceder ao algoritmo o acesso aos bancos de dados gerenciados pelo MongoDB, permitindo que o *script* adicione, altere e resgate dados armazenados nesta plataforma (MONGODB, 2015).

Além da biblioteca PyMongo, foi utilizado o RoboMongo[3], que é um software de interação com elementos visuais, desenvolvido para auxiliar nas consultas aos bancos de dados do MongoDB. Nele, é possível visualizar as collections (pastas onde são guardados os dados capturados) e as chaves (características) dos *tweets* rapidamente, sem digitar linhas de código no *console* ou criar *scripts* para localizar alguma informação. Essencialmente, neste projeto, o RoboMongo foi utilizado para monitorar o processo de coleta, verificando se o algoritmo estava inserindo corretamente os dados no banco e certificando-se de que não havia nenhum problema nesta fase do projeto.

Tendo conhecimento das ferramentas e de suas funções especificas, é possível construir os *scripts* e dar início ao desenvolvimento do método aplicado ao projeto.

3 MÉTODO

O método aplicado neste artigo se subdivide em três etapas:

1) Coleta dos dados: fase que contempla a conexão com o servidor do Twitter e a utilização de *scripts* cujas funções são as de capturar dados a partir da API e de armazená-los no banco de dados do MongoDB.
2) Tratamento dos dados: etapa que consiste em utilizar *scripts* cujas funções são as de acessar o banco de dados e extrair determinadas informações dos *tweets*, armazenando-os em arquivos do tipo txt.
3) Análise dos dados: fase final dos métodos aplicados, consistindo em efetuar análises estatísticas e representativas com base em valores numéricos associados às informações extraídas na etapa anterior.

Nas próximas subseções, cada etapa citada será apresentada de maneira mais explicativa, descrevendo precisamente como foi condicionada.

3.1 Coleta dos dados

O primeiro passo para o desenvolvimento do projeto foi a determinação das palavras-chave (*keywords*), que têm a função de resumir o assunto escolhido para o estudo de caso. De modo geral, as *keywords* devem ser escolhidas de maneira a obter o maior número possível de *tweets* sobre o assunto, além de inibir *tweets* fora de contexto (que contêm as *keywords*, mas que não se referem ao assunto esperado). Seja, por exemplo, o assunto do estudo de caso escolhido como "Eleições Presidenciais no Brasil". Caso as *keywords* sejam genéricas, como "eleições presidenciais" e "Brasil", o *script* poderá capturar muitos *tweets* que discursem sobre eleições presidenciais em outros países ou que citem o Brasil diante outros assuntos, retornando dados inconsistentes para análises estatísticas. Vale ressaltar que, para o Python, as *keywords*

[3] Software disponível gratuitamente em: https://www.mongodb.com/download-center

são *strings* (sequências de caracteres de texto), e, portanto, o *script* apenas captura *tweets* que contêm exatamente a mesma *string*. Para tanto, utilizou-se o parâmetro "u" para cada *keyword;* ele esclarece ao *script* que a busca deve ser feita utilizando as variações maiúscula e minúscula das letras. Dessa forma, a *keyword* "u'eleições'", por exemplo, retornaria tweets com "ELEIÇÕES", "Eleições" e "eleições".

A seleção das palavras-chave baseou-se nos nomes dos candidatos de maior popularidade para cada cidade estudada. A tabela 1 relaciona as *keywords* (separadas por ";") utilizadas com seus respectivos candidatos e cidades:

Tabela 1 – *Keywords* utilizadas para efetuar a coleta de dados

Cidade	Candidato	*Keywords*
São Paulo	Celso Russomanno	u'russomanno'; u'russomano'
	Fernando Haddad	u'haddad'; u'hadad'
	João Doria	u'doria'
	Levy Fidélix	u'fidelix'; u'fidélix'
	Luiza Erundina	u'erundina'
	Marta Suplicy	u'marta suplicy'
Santo André	Ailton Lima	u'ailton lima'
	Carlos Grana	u'carlos grana'
	Dr. Aidan	u'aidan'
	Paulo Serra	u'paulo serra'
	Rafael Daniel	u'rafael daniel'
	Ricardo Alvarez	u'ricardo alvarez'
São Bernardo do Campo	Alex Manente	u'alex manente'
	Aldo Santos	u'aldo santos'
	Cesar Raya	u'cesar raya'
	Orlando Morando	u'orlando morando'
	Tarcísio Secoli	u'secoli'
	Tunico Vieira	u'tunico vieira'
São Caetano do Sul	Auricchio	u'auricchio'
	Fabio Palacio	u'fabio palacio'; u'fábio palacio'
	Gilberto Costa	u'gilberto costa'
	Lucia Dal Mas	u'lucia dal mas'
	Paulo Pinheiro	u'paulo pinheiro'
Diadema	Amb. Virgílio	u'ambientalista virgilio'; u'ambientalista virgílio'
	Lauro Michels	u'lauro michels'
	Professor Ivanci	u'ivanci'
	Vaguinho	u'vaguinho'
Mauá	Atila Jacomussi	u'jacomussi'
	Clovis Volpi	u'volpi'
	Donisete Braga	u'donisete'
	Marcio Chaves	u'marcio chaves'
Ribeirão Pires	Carlos Sacomani	u'sacomani'
	Dedé da Folha	u'dede da folha'; u'dedé da folha'
	Grecco	u'grecco'
	Kiko	u'kiko'
	Rosana Figueiredo	u'rosana figueiredo'

Rio Grande da Serra	Claudinho da Geladeira	u'clauinho da geladeira'
	Cleson Alves	u'cleson alves'
	Edvaldo Guerra	u'edvaldo guerra'
	Gabriel Maranhão	u'gabriel maranhao'; u'gabriel maranhão'

Fonte: Elaborada pelos autores

O *script* destinado a fazer a coleta de dados faz uso das bibliotecas Twython, para conexão e autenticação à API do Twitter, e PyMongo, para acesso, edição e criação de bancos de dados através do MongoDB. Nesta etapa, é necessário fazer uso das chaves de acesso para autenticação do *script* (vide seção 3.2.1). A fase de coleta foi efetuada em tempo real, visando obter dados baseados no primeiro turno das eleições. O *script* permaneceu em execução durante 3 dias (de 31 de setembro de 2016 até 02 de outubro de 2016), sendo encerrado antes da apuração dos votos.

Como descrito brevemente nas seções anteriores, a Streaming API (método principal de acesso à API utilizada) é um tipo de acesso que fornece dados em tempo real e, por isso, estabelece uma conexão constante com o servidor do Twitter. Ela utiliza as bibliotecas PyMongo e Twython. De modo geral, o *script*, após sua autenticação, executará rotinas para que a API faça uma busca no banco de dados da rede social, verificando a todo instante se há um novo *tweet* publicado que contenha alguma *keyword*. Caso haja, os dados deste *tweet* são capturados e armazenados no banco de dados criado pelo MongoDB e, após isso, a busca continuará sendo feita. Esse processo apenas será interrompido pelo programador ou por eventuais desconexões com o servidor.

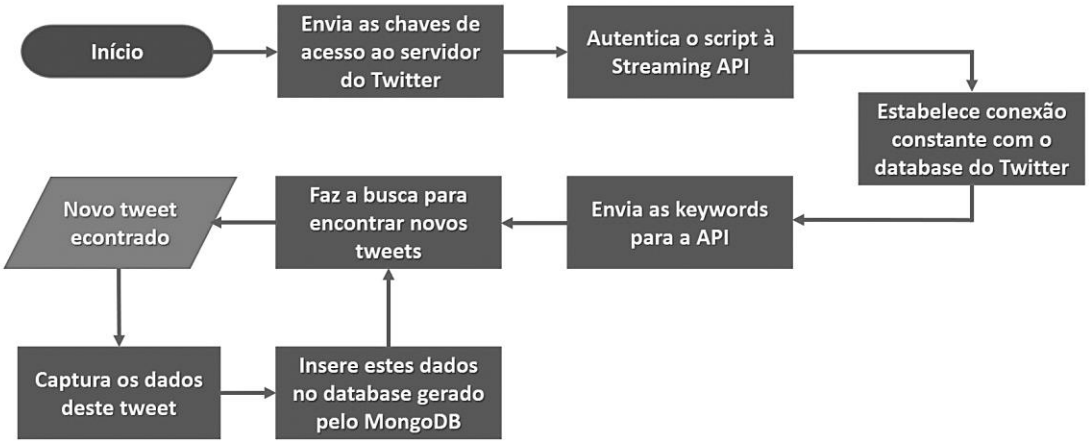
No *script*[4] utilizado para fazer a coleta de dados a partir da Streaming API, inicialmente se faz a importação das duas bibliotecas mencionadas. Logo após, é definida a classe "Coleta", responsável por praticamente fazer todo o processo de coleta. Nesta classe, há duas funções. A primeira faz o envio das chaves de autenticação para o servidor e a conexão com o banco de dados criado pelo MongoDB; a segunda executa o processo de inserção de dados no banco. Nas linhas abaixo, são definidas as chaves de acesso, identificando-as do mesmo modo mostrado na figura 3. O próximo comando define "stream" como um elemento da classe "Coleta" e insere neste elemento as strings correspondentes às chaves, criando assim uma entidade de acesso. Finalmente, os últimos comandos definem quais são as *keywords* e configuram a entidade "stream" para que o *script*, em conjunto com a API, execute o processo de busca com base nestas palavras-chave.

A figura 4 apresenta um fluxograma simplificado dos processos que o *script* descrito acima executa, evidenciando o fluxo de dados:

.

[4] Disponível em: https://gist.github.com/BMasunaga/e019fb1a435f4171bf368a25fba00d06

Figura 4: Fluxograma representativo do processo de coleta de dados a partir da Streaming API



Fonte: Elaborada pelos autores em 05 jun. 2017.

3.2 Tratamento dos dados

Cada *tweet* coletado pelos processos apresentados na seção anterior é armazenado no banco de dados contendo diversas informações:

Figura 5: Exemplo de informações sobre um *tweet* coletado (visíveis na plataforma RoboMongo)

place	null	Null
id	782025670171975680	Int64
truncated	false	Boolean
text	RT @PenseDifSP: . @mariagadu : "Para continuar a construir uma c...	String
lang	pt	String
retweet_count	0	Int32
entities	{ 4 fields }	Object
user_mentions	[2 elements]	Array
symbols	[0 elements]	Array
urls	[0 elements]	Array
hashtags	[3 elements]	Array
[0]	{ 2 fields }	Object
text	HoraH	String
indices	[2 elements]	Array
[1]	{ 2 fields }	Object
text	ViradaHaddad13	String
indices	[2 elements]	Array
[2]	{ 2 fields }	Object
text	Haddad13	String
indices	[2 elements]	Array
timestamp_ms	1475284423760	String
is_quote_status	false	Boolean

Fonte: Elaborada pelos autores em 08 jun. 2017.

A figura 5 demonstra que todas as informações armazenadas de um *tweet* podem ser acessadas e lidas. Tais informações encontram-se organizadas em diretórios específicos. As células destacadas representam, respectivamente, o texto do *tweet* e as *hashtags* presentes neste *tweet*. Muitas informações são importantes no processo de tratamento de dados, sendo algumas mais significativas do que outras dependendo do tipo de análise que se deseja fazer.

Neste artigo, o processo de tratamento de dados consiste, de modo geral, em um levantamento dos termos e das *hashtags* mais utilizadas; dos *users* que mais tweetaram, dos que mais foram citados e, por fim, do número de *tweets* referentes a cada candidato analisado (entende-se por "termo" um conjunto de caracteres isolado por espaços). Para tanto, foram utilizados cinco diferentes *scripts* nesta seção. Tais algoritmos geram como output (saída de dados) um arquivo do tipo txt, cujo conteúdo pode ser descrito como um conjunto de informações quantitativas, demonstrando uma relação numérica com cada termo/*hashtag*/*user* analisado (com exceção do último código, que retorna o texto bruto de cada *tweet* referente a cada candidato analisado). É importante ressaltar que os algoritmos utilizados nessa seção foram criados de forma a não efetuar a contagem das repetições de uma mesma palavra/*hashtag*/*user* mais de uma vez, visando reduzir o custo computacional exigido.

O primeiro algoritmo[5] utilizado nesta etapa do projeto foi versa sobre o levantamento dos termos mais utilizados e, portanto, visa recuperar apenas o "text" de todos os *tweets*. O número mínimo de repetições para se considerar um termo foi definido em 50, uma vez que há muitas palavras utilizadas poucas vezes em cada *tweet*. Como o número de dados capturados foi de ordem grande, o fator determinante para a escolha deste critério se deu pelo desejo de evitar custo computacional elevado, ignorando termos que aparecem sem grande frequência.

O segundo *script*[6] trata de um levantamento das *hashtags* mais utilizadas. O número mínimo de repetições para se considerar uma *hashtag* foi definido em 10, pelo mesmo motivo apresentado para o critério aplicado ao primeiro algoritmo. Este número é menor pois, maneira geral, *hashtags* são utilizadas com menor frequência do que palavras.

Os dois próximos algoritmos versam sobre o levantamento dos *users* mais citados[7] e dos *users* que mais tweetaram[8] no período estudado. O número mínimo de repetições para se considerar um *user* nos dois casos também foi definido em 10.

O último *script*[9] de tratamento elaborado para essa etapa do projeto é o mais simples; sua função é a de retornar o texto bruto de todos os *tweets* referentes a cada um dos candidatos (um por linha). Ao fim da execução do último *script*, foram gerados 40 arquivos do tipo txt, os quais foram analisados manualmente com ajuda do software Microsoft Office Excel, de modo a excluir eventuais *tweets* fora de contexto. Após isso, procedeu-se a contagem de linhas de cada arquivo txt utilizando-se do mesmo software, gerando, assim, números que representam uma aproximação da quantidade de *tweets* referentes a cada candidato.

3.3 Análise dos dados

3.3.1 As análises gerais

Foram denominadas "análises gerais" aquelas que comparam informações cujo significado é apenas quantitativo em relação ao projeto. Desta forma, trabalha somente com os números absolutos relacionados às informações. O objetivo dessas análises é traçar um perfil geral dos usuários da rede, identificando padrões na grande massa, bem como os atores políticos virtuais mais influenciadores.

Este grupo foca em três objetos de estudo diferentes: termos, *hashtags* e *users*, e contempla as seguintes análises:

1) Análise geral dos 200 termos mais utilizados: utiliza as informações produzidas a partir do uso do *script* de tratamento dos termos mais utilizados. Considera os 200

[5] Disponível em: https://gist.github.com/BMasunaga/062fb643510a77256781538db00b32c3
[6] Disponível em: https://gist.github.com/BMasunaga/c2815ae71a92f12d4837c2e83cfaa866
[7] Disponível em: https://gist.github.com/BMasunaga/e9660b33b15a58ce019571dc19e99d82
[8] Disponível em: https://gist.github.com/BMasunaga/38c94cf66abde382c4f4be240f46506d
[9] Disponível em: https://gist.github.com/BMasunaga/3fe0eee32ddd5c9b2afba5ca92068f89

termos mais utilizados, ordenados em ordem decrescente de frequência de utilizações;

2) Análise geral das 100 *hashtags* mais utilizadas: utiliza as informações produzidas a partir do uso do *script* de tratamento das *hashtags* mais utilizadas. Foram consideradas as 100 *hashtags* mais utilizadas, ordenadas em ordem decrescente de frequência de utilizações;
3) Análise geral dos 20 *users* mais citados e dos 20 que mais tweetaram: para os *users* mais citados, utiliza as informações produzidas a partir do uso do *script* de tratamento dos *users* mais citados, e para os *users* que mais tweetaram utiliza as informações do *script* de tratamento dos *users* que mais tweetaram. Nesta análise, foram considerados apenas os 20 *users* prevalecentes em cada caso, ordenados em ordem decrescente de citações ou de *tweets* produzidos.

Os procedimentos adotados para efeito de cada uma dessas análises consistem em comparar a representação quantitativa das informações (no caso, os números relacionados aos termos, *hashtags* ou *users*). Para fins de apresentação, utilizou-se nuvens de palavras para os termos e *hashtags* e tabelas para os *users*.

3.3.2 As análises comparativas

As análises comparativas focaram em desenvolver as porcentagens envolvidas em cada contexto de análise, além de ter como objeto de estudo a quantidade de *tweets* relacionados a cada candidato pesquisado. Desta forma, buscam trabalhar os dados gerados pelo *script* de tratamento dos *tweets* referentes a cada candidato.

De maneira similar às análises gerais, o procedimento básico aplicado é o de comparação das representações quantitativas dos dados. No entanto, as análises comparativas se diferem no quesito de significação para o projeto, uma vez que visam efetuar estudos mais profundos e de caráter relativo, explicitando padrões de comportamento ou de perfil de usuários da rede com base nas cidades e/ou regiões na qual vivem.

De maneira geral, as tarefas deste grupo de análise se dividem em:

1) Análise comparativa entre São Paulo e o Grande ABC: trata-se da divisão percentual da totalidade de *tweets* capturados. Para tanto, tem como base os números de *tweets* referentes aos candidatos de cada região;
2) Análise comparativa entre as cidades do Grande ABC: trata-se da divisão percentual da parcela de *tweets* capturados da região do Grande ABC. Para tanto, tem como base os números de *tweets* referentes aos candidatos de cada cidade desta região;
3) Análise comparativa entre os candidatos de cada cidade: trata-se da divisão percentual de *tweets* capturados para cada cidade pesquisada. Para tanto, tem como base os números de *tweets* referentes a cada candidato de cada cidade pesquisada.

Não obstante, os resultados obtidos pelas análises descritas acima serão comparados com os dados reais fornecidos pelo Tribunal Superior Eleitoral (TSE)[10] sobre o 1° turno das eleições municipais de 2016, a fim de efetuar uma última comparação entre o estudo de caso conduzido por esta pesquisa e o processo eleitoral.

4 RESULTADOS

4.1 Das análises gerais

[10] Disponíveis em: http://divulga.tse.jus.br/oficial/index.html

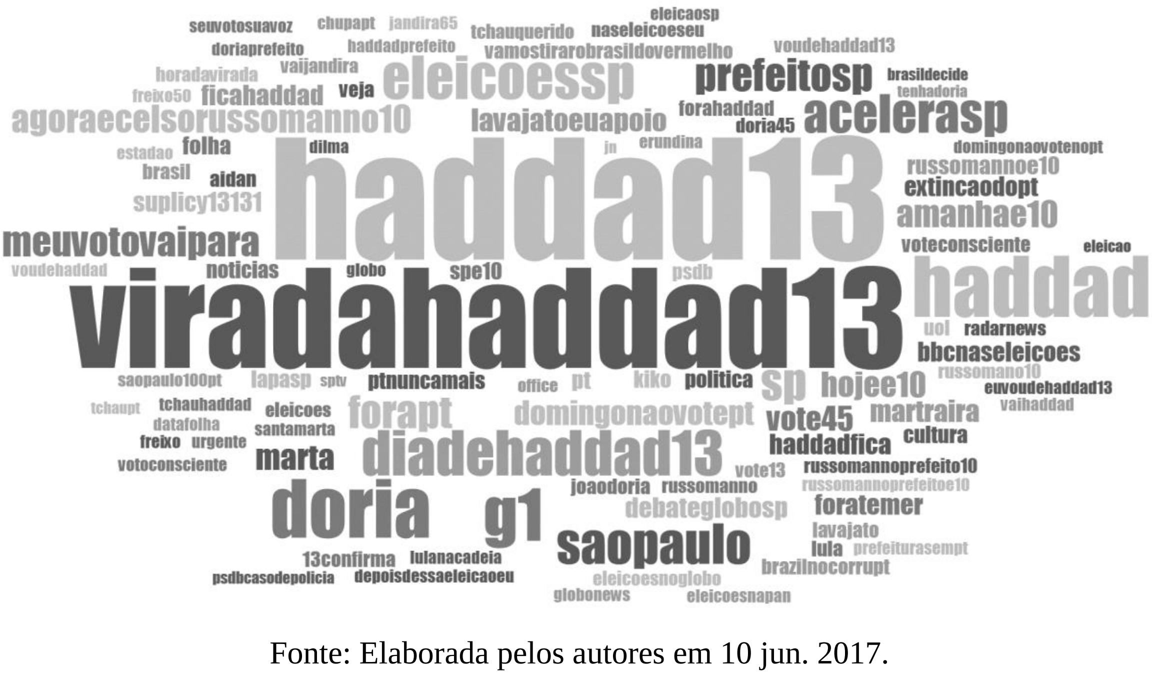
Para os termos mais utilizados, verificou-se “doria” como sendo o primeiro colocado, com 137.256 aparições e “fracasso”, como o último, com 1.471 aparições. A figura 6 apresenta a nuvem de palavras[11] gerada, que abrange todos os 200 termos, sendo o tamanho da fonte de cada um diretamente proporcional ao seu número de aparições.

Figura 6: Nuvem de palavras representativa dos 200 termos mais utilizados

Fonte: Elaborada pelos autores em 10 jun. 2017.

Para as *hashtags* mais utilizadas, verificou-se “#haddad13” como sendo a primeira colocada, com 2.892 aparições e “#foradoria”, como a última, com 20 aparições. A figura 7 apresenta a nuvem de palavras[11] gerada, que abrange todas as 100 *hashtags* (sem “#”).

Figura 7: Nuvem de palavras representativa das 100 *hashtags* mais utilizadas



Fonte: Elaborada pelos autores em 10 jun. 2017.

---

[11] Ferramenta utilizada no endereço eletrônico: http://jasondavies.com/wordcloud.

Por sua vez, as análises baseadas nos *users* seguiu uma apresentação um pouco distinta. A tabela 2 sumariza os resultados, indicando os atores virtuais mais influentes em cada caso:

Tabela 2 – As análises gerais relativas aos *users* mais citados e os que mais tweetaram

Os 20 *users* mais citados		**Os 20 *users* que mais tweetaram**	
User	**Citações registradas**	***User***	***Tweets* produzidos**
@estadao	2770	@prof_fabio666	1184
@daniduncan	2441	@lobaoeletrico	355
@ecantanhede	2389	@trajanojorge	353
@xicosa	2325	@ricardo201611	327
@sensacionalista	2077	@odio_nao	245
@cartamaior	1964	@rafaelalvesilva	233
@blogdopim	1964	@oconsciente	231
@jpcuenca	1877	@espinoza19661	230
@berriel	1790	@dionianjos	205
@folha	1637	@midiacrucis	201
@chris_branco	1525	@juliocesaramor	201
@chicobarney	1511	@ikmkoji	198
@g1	1438	@araujosergio	178
@taviao	1403	@iva_ivani	175
@rodrigodasilva	1333	@annapscorrea	173
@uolnoticias	1245	@alexsandraponce	172
@_abeautifullie	1028	@fernandocesar77	172
@j_livres	1008	@nancyafc	171
@o_antagonista	995	@robertostarckle	171
@blogdojefferson	959	@spfc_92_93_005	170
@gabesimas	957	@thecomiranda	167
@brasil247	954	@gumartinslive	166
@haddad_femando	940	@liaaraujo19	164

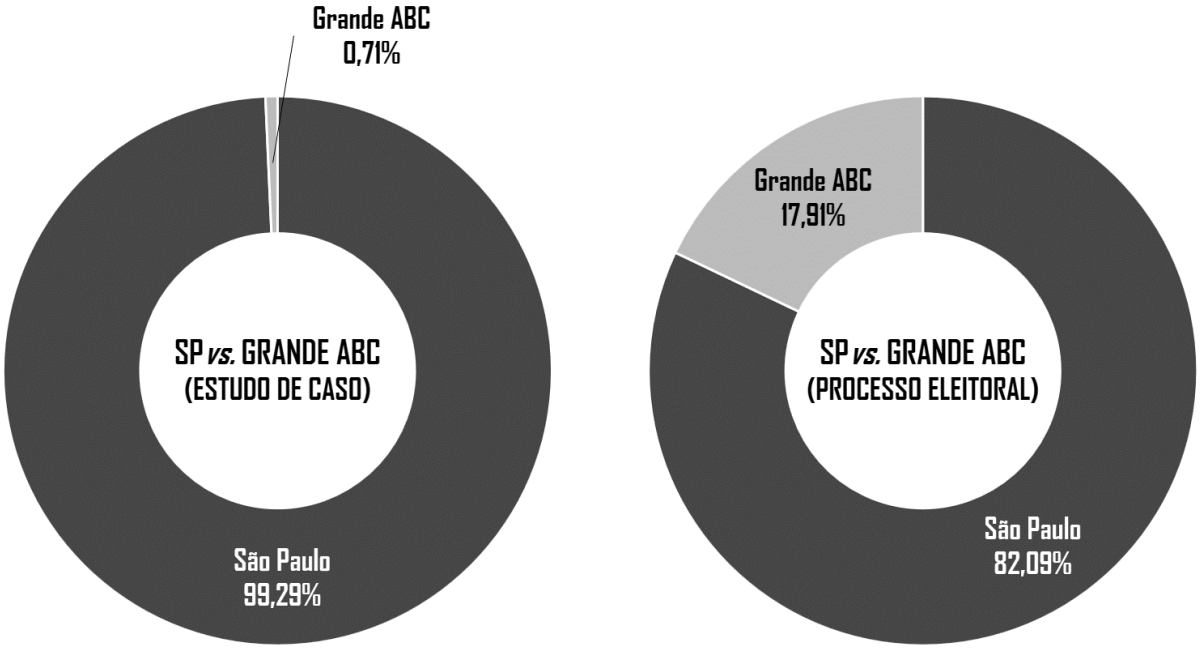
Fonte: Elaborada pelos autores em 10 jun. 2017.

4.2 Das análises comparativas

Como especificado anteriormente, as análises comparativas foram aplicadas tanto aos dados obtidos pelo estudo de caso (número de *tweets*) quanto aos dados retirados do TSE sobre as eleições municipais (número de votos).

A figura 8 sintetiza os resultados da análise comparativa entre a cidade de São Paulo e a região do Grande ABC, destacando suas respectivas porcentagens conforme o conjunto de dados analisado:

Figura 8: Conjunto representativo da análise comparativa entre São Paulo e o Grande ABC

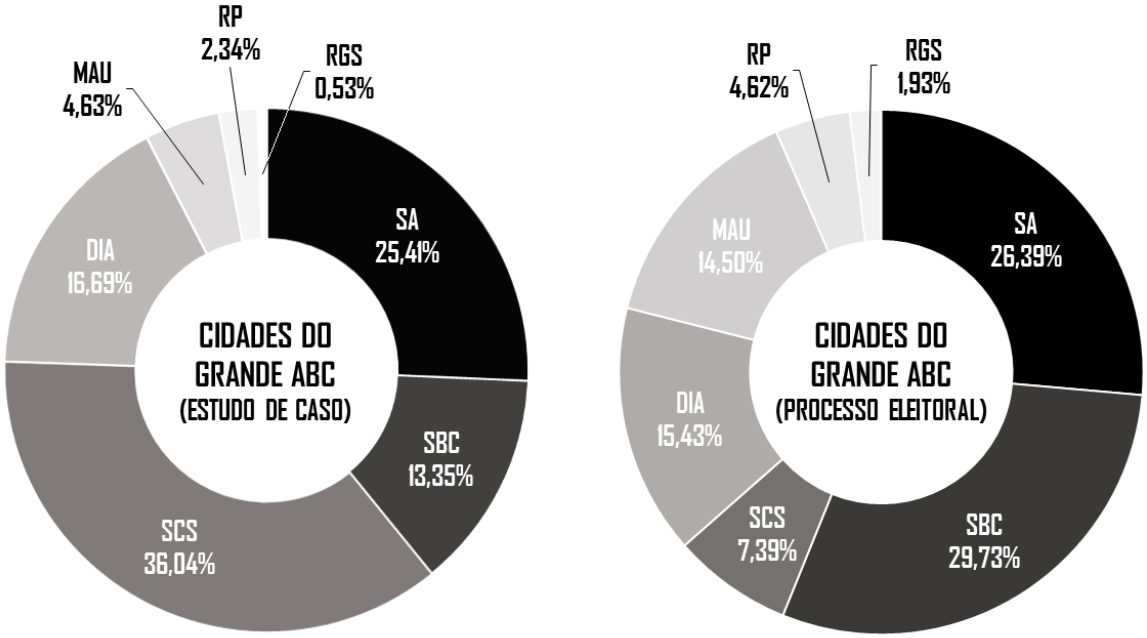


Fonte: Elaborada pelos autores

Segundo dados do TSE, o número total de votos válidos para essas duas regiões (incluindo os votos para candidatos não abrangidos neste estudo) foi de 7.053.282. No estudo de caso, o número total de *tweets* capturados que fazem referência a essas regiões foi de 265.961.

A análise comparativa entre as cidades do Grande ABC foi conduzida de mesma maneira que a análise anterior, apenas diferenciando-se por considerar apenas os *tweets* ou votos correspondentes à Região do ABC. A figura 9 apresenta seus resultados:

Figura 9: Conjunto representativo da análise comparativa entre as cidades do Grande ABC



Fonte: Elaborada pelos autores

Segundo dados do TSE, o número total de votos registrados da região do Grande ABC foi de 1.263.391 (incluindo os votos para candidatos não abrangidos neste estudo). No estudo de caso, no entanto, o número total de *tweets* capturados que fazem referência a essa região foi de 1.881.

A terceira análise deste grupo, que versa sobre a comparação entre os candidatos de cada cidade pesquisada, foi aplicada, no caso dos dados do TSE, apenas aos candidatos estudados nesta pesquisa e, portanto, a soma percentual pode não resultar em 100%. A tabela 3 apresenta cada candidato com seus respectivos números de *tweets* e de votos.

Tabela 3 – Relação entre os candidatos estudados e seus respectivos n° de *tweets*

Cidade	Candidato (Partido)	N° de *tweets* (%)	N° de votos (%)
São Paulo	João Doria (PSDB)	113.365 (42,93%)	3.085.187 (53,29%)
	Fernando Haddad (PT)	108.179 (40,96%)	967.190 (16,70%)
	Celso Russumanno (PRB)	18.995 (7,19%)	789.986 (13,64%)
	Marta (PMDB)	14.653 (5,55%)	587.220 (10,14%)
	Luiza Erundina (PSOL)	8.572 (3,25%)	184.000 (3,18%)
	Levy Fidélix (PRTB)	316 (0,12%)	21.705 (0,37%)
Santo André	Paulo Serra (PSDB)	223 (46,65%)	119.540 (35,85%)
	Carlos Grana (PT)	174 (36,4%)	67.628 (20,28%)
	Dr. Aidan (PSB)	80 (16,74%)	56.804 (17,04%)
	Ricardo Alvarez (PSOL)	1 (0,21%)	11.099 (3,33%)
	Ailton Lima (SD)	0 (0%)	49.959 (14,98%)
	Rafael Daniel (PMDB)	0 (0%)	10.716 (3,21%)
São Bernardo do Campo	Orlando Morando (PSDB)	129 (47,78%)	169.310 (45,07%)
	Alex Manente (PPS)	92 (34,07%)	106.726 (28,41%)
	Tarcísio Secoli (PT)	45 (16,67%)	84.768 (22,57%)
	Aldo Santos (PSOL)	3 (1,11%)	6.972 (1,86%)
	Tunico Vieira (PMDB)	1 (0,37%)	7.046 (1,88%)
	Cesar Raya (PSTU)	0 (0%)	815 (0,22%)
São Caetano do Sul	Auricchio (PSDB)	561 (82,74%)	32.047 (34,34%)
	Paulo Pinheiro (PMDB)	89 (13,13%)	28.674 (30,71%)
	Gilberto Costa (PEN)	14 (2,06%)	5.740 (6,15%)
	Fabio Palacio (PR)	13 (1,92%)	19.291 (20,66%)
	Lucia Dal Mas (PRTB)	1 (0,15%)	5.102 (5,46%)
Diadema	Lauro Michels (PV)	247 (78,66%)	93.772 (48,10%)
	Vaguinho (PRB)	66 (21,02%)	42.596 (21,85%)
	Amb. Virgílio (REDE)	1 (0,32%)	834 (0,43%)
	Professor Ivanci (PSTU)	0 (0%)	537 (0,28%)
Mauá	Atila Jacomussi (PSB)	50 (57,47%)	85.615 (46,73%)
	Donisete Braga (PT)	35 (40,23%)	41.958 (22,90%)
	Clovis Volpi (PSDB)	1 (1,15%)	37.065 (20,23%)
	Marcio Chaves (PSD)	1 (1,15%)	7.691 (4,20%)

Ribeirão Pires	Kiko (PSB)	38 (86,36%)	17.703 (30,31%)
	Grecco (PRB)	6 (13,64%)	13.942 (23,87%)
	Carlos Sacomani (PSL)	0 (0%)	784 (1,34%)
	Dedé da Folha (PPS)	0 (0%)	15.385 (26,34%)
	Rosana Figueiredo (REDE)	0 (0%)	443 (0,76%)
Rio Grande da Serra	Gabriel Maranhão (PSDB)	6 (60%)	11.080 (45,47%)
	Claudinho da Geladeira (PT)	3 (30%)	8.166 (33,51%)
	Edvaldo Guerra (PMDB)	1 (10%)	2.528 (10,37%)
	Cleson Alves (PMB)	0 (0%)	2.595 (10,65%)

Fonte: Elaborada pelos autores

5 DISCUSSÃO

A partir dos resultados apresentados na seção anterior, diversas observações e discussões podem ser tomadas, de modo a alcançar os objetivos propostos neste projeto.

No caso dos resultados obtidos nas análises gerais, era esperado que nomes de candidatos aparecessem entre os termos e/ou *hashtags* mais utilizadas, uma vez que as *keywords* utilizadas na fase de coleta foram variantes textuais dos nomes de cada candidato analisado (vide tabela 1). De maneira geral, nota-se visualmente que os nomes presentes nas nuvens de palavras são predominantemente de candidatos pertencentes a São Paulo, refletindo que a presença desta cidade na massa de dados capturada é a maior (e tal fato pode ser confirmado pela análise comparativa representada pela figura 6).

Apesar de, segundo análise comparativa ilustrada na tabela 3, o candidato Fernando Haddad (de São Paulo) não possuir o maior número de *tweets*, seu nome predomina parcialmente nos termos mais utilizados e demonstra possuir grande espaço também na análise das *hashtags*. Isso pode indicar, por exemplo, que o perfil geral dos usuários que comentaram sobre este candidato em específico possui a característica de utilizar *hashtags* mais do que os usuários que comentaram sobre os demais candidatos. Outro destaque, neste mesmo contexto, é o candidato Celso Russomanno (de São Paulo) que, embora tenha obtido um número baixo de *tweets* em relação aos dois primeiros colocados de sua cidade, demonstrou notável influência na análise das *hashtags*, pois visualmente é possível perceber várias delas referentes ao número 10, que é o número representativo do partido político do qual faz parte.

Acerca das análises gerais baseadas nos números referentes a usuários, verifica-se que aparecem, como *users* mais citados, páginas jornalísticas virtuais (como @estadao, @folha, @g1 etc), enquanto na análise de *users* que mais tweetaram apareceram, quase em sua totalidade, páginas de pessoas físicas. Tais fatos podem demonstrar que usuários que representam jornais virtuais atuam no ambiente do Twitter disseminando pouca informação, mas detêm grande influência no espaço virtual, dado que são frequentemente citadas. Desta forma, são atores políticos influentes no meio. Em contraposição, as pessoas físicas, de modo geral, produzem mais *tweets* do que são citadas por outros usuários, fazendo com que sejam atores políticos producentes no meio, isto é, que contribuem para a expressão quantitativa dos dados públicos (no caso, o número de *tweets* capturados).

Por sua vez, os resultados das análises comparativas forneceram informações que podem ser interpretadas de diversas maneiras. De modo geral, observa-se, a primeiro momento, que os resultados das análises aplicadas tanto aos dados do estudo de caso quanto aos dados do processo eleitoral apresentam semelhanças, como, por exemplo, nas razões de proporções entre cidades e entre candidatos de cidades. Apesar de existir tal concordância, no entanto, a

porcentagem esperada (isto é, a porcentagem que foi obtida através do estudo de caso) diverge consideravelmente dos resultados eleitorais em alguns casos. Desta forma, para quantificar tal relação, pode-se fazer uma análise estatística simples, definindo o Índice de Concordância (I_C). Este índice busca inferir percentualmente a precisão média que descreve a relação de concordância entre a porcentagem de *tweets* obtida por um candidato e sua porcentagem de votos obtida no processo eleitoral. É calculada através da fórmula abaixo:

$$I_C = 1 - \left(\frac{1}{N} * \sum_{i=1}^{N} |PT_i - PV_i| \right)$$

Onde:
N é o número de objetos considerados na análise;
PT é a porcentagem relativa de *tweets* referentes ao objeto *i;*
PV é a porcentagem relativa de votos referentes ao objeto *i*.

Entende-se por "objeto" o foco analítico de cada análise comparativa efetuada. Por exemplo, para a primeira análise comparativa, define-se como "objeto" cada região comparada (cidade de São Paulo e aglomerado de cidades do Grande ABC). No caso da segunda análise, o "objeto" passa a ser cada cidade do Grande ABC estudada.

Visto que o processo de coleta de dados foi encerrado antes da apuração dos votos, com essa ferramenta, é possível estimar numericamente quão fiéis foram os resultados obtidos pelo estudo de caso em relação aos resultados do processo eleitoral real, de modo a verificar se o método utilizado nesta pesquisa poderia ser utilizado, por exemplo, para efetuar projeções de resultados reais antes do acontecimento do evento.

Para demonstrar o funcionamento desta fórmula, considerou-se a primeira análise comparativa. Para o objeto "São Paulo", o termo "|PT – PV|" é (|99,29% – 82,09%| = |17,2%| = 17,2%). Para o objeto "Grande ABC", o termo "|PT – PV|" é (|0,71% – 17,91%| = |-17,2%| = 17,2%). Desta forma, o resultado do somatório é (17,2% + 17,2% = 34,4%). Se N é o número de objetos, então N é igual a 2. Logo, dividindo o somatório por N, obtém-se 17,2%. Finalmente, subtraindo de 1, tem-se I_C = 83,8%. Assim, pode-se dizer que a confiabilidade dos resultados obtidos a partir do estudo de caso é de 83,8% nesta análise.

Aplicando-se esse método estatístico para o restante das análises, obtém-se:

- Para a análise comparativa entre as cidades do Grande ABC, I_C = 89,42%;
- Para a análise comparativa entre os candidatos de cada cidade estudada, I_C = 89,72%;

Finalmente, tomando a média aritmética simples destes índices, infere-se que a confiabilidade média dos resultados das análises comparativas é de 87,65%, com um desvio padrão de 3,34%.

No entanto, não se limita a esse tipo estatístico de análise as possibilidades de extração de índices de concordância a partir dos resultados fornecidos pelas análises comparativas. Por exemplo, utilizando-se as informações da tabela 3 e do TSE, a tabela 4 relaciona a posição dos candidatos dentro de suas respectivas cidades:

Tabela 4 – Comparação entre os dados do Estudo de Caso e os dados do Processo Eleitoral

Cidade	Posição	Estudo de Caso	Processo Eleitoral*
São Paulo	**Eleito**	João Doria (42,93%)	João Doria (53,29%)
	2°	Fernando Haddad (40,96%)	Fernando Haddad (16,70%)
	3°	Celso Russomanno (7,19%)	Celso Russomanno (13,64%)
Santo André	**Eleito**	Paulo Serra (46,65%)	Paulo Serra (35,85%)
	2°	Carlos Grana (36,40%)	Carlos Grana (20,28%)
	3°	Dr. Aidan (16,74%)	Dr. Aidan (17,04%)

São Bernardo do Campo	**Eleito**	Orlando Morando (47,78%)	Orlando Morando (45,07%)
	2°	Alex Manente (34,07%)	Alex Manente (28,41%)
	3°	Tarcísio Secoli (16,67%)	Tarcísio Secoli (22,57%)
São Caetano do Sul	**Eleito**	Auricchio (82,74%)	Auricchio (34,34%)
	2°	Paulo Pinheiro (13,13%)	Paulo Pinheiro (30,71%)
	3°	Gilberto Costa (2,06%)	Fabio Palacio (20,66%)
Diadema	**Eleito**	Lauro Michels (78,66%)	Lauro Michels (48,10%)
	2°	Vaguinho (21,02%)	Vaguinho (21,85%)
	3°	Ambientalista Virgílio (0,32%)	Maninho (16,37%)
Mauá	**Eleito**	Atila Jacomussi (57,47%)	Atila Jacomussi (46,73%)
	2°	Donisete Braga (40,23%)	Donisete Braga (22,90%)
	3°	C. Volpi/M. Chaves (1,15%)	Clovis Volpi (20,23%)
Ribeirão Pires	**Eleito**	Kiko (86,36%)	Kiko (30,31%)
	2°	Grecco (13,64%)	Dedé da Folha (26,34%)
	3°	-	Grecco (23,87%)
Rio Grande da Serra	**Eleito**	Gabriel Maranhão (60%)	Gabriel Maranhão (45,47%)
	2°	Claudinho da Geladeira (30%)	Claudinho da Geladeira (33,51%)
	3°	Edvaldo Guerra (10%)	Cleson Alves (10,65%)

Fonte: Elaborada pelos autores

De acordo com tais dados, em todas as cidades estudadas, o candidato que mais foi citado no estudo de caso também foi aquele que mais recebeu votos no processo eleitoral. Também se nota que muitos candidatos descritos nas 2° e 3° colocações coincidiram nas mesmas posições nos dois casos. Esse tipo de comparação entre os dados do estudo de caso poderia ser utilizado para efetuar projeções sobre os futuros candidatos eleitos ou possíveis segundos turnos, bem como os candidatos envolvidos na ocasião. Descrevendo quantitativamente, tem-se que de 24 candidatos, 19 assumiram a mesma posição tanto no estudo de caso quanto no processo eleitoral. Logo, considerando-se a razão simples entre os dois números, infere-se que é de 79,17% a confiabilidade deste tipo de projeção (valor próximo ao da confiabilidade média das análises comparativas).

Apesar dos altos números que representam a confiabilidade, não é de todo correto correlacionar positivamente o número de *tweets* obtido por um candidato com o seu número de votos. Tal impasse se dá pelo fato de que o voto significa apoio a determinado candidato, e, no entanto, uma citação num *tweet* pode ter caráter positivo ou negativo, implicando que um usuário que cite um candidato não necessariamente o apoie. Embora forneça um panorama geral (que, visto pelo índice médio de concordância, é coincidente com a realidade neste estudo de caso), a análise aplicada no projeto não é capaz de diferenciar as opiniões públicas dos atores virtuais. Essa impossibilidade pode explicar, por exemplo, o fato de que as *hashtags* que citam o candidato Fernando Haddad (de São Paulo) são as que apresentam maior impacto, mesmo o candidato não possuindo o maior número de *tweets*. Visto que tais *hashtags* apresentam caráter positivo ao candidato do partido PT (vide figura 7), uma possibilidade é a do candidato da oposição, João Doria (PSDB), que teve o maior número de citações em *tweets*, ter sido muito citado também pelos usuários que apoiam Haddad, mas de maneira negativa. Há complicação no intuito de tratar os dados de modo a diferenciar os *tweets* conforme o caráter opinativo que carregam. Para efeito, seria necessária a aplicação de técnicas elaboradas, como a análise de

sentimentos – atualmente utilizada também em pesquisas que utilizam o Twitter como fonte de dados (FRANÇA; OLIVEIRA, 2014) e também a identificação dos usuários mais influentes na rede (LOPES; VIEIRA, 2016), de modo a aprofundar a pesquisa e aprimorar o método, possibilitando o estudo e produção de novas descobertas acadêmicas.

6 CONCLUSÃO

Os processos aplicados neste projeto, que abrangem coleta, tratamento e análise de dados públicos do Twitter, foram satisfatórios para alcançar os objetivos propostos. Foi possível traçar um perfil superficial dos atores políticos mais influentes no estudo de caso, bem como explicitar padrões textuais (termos e/ou *hashtags*), destacando possíveis comportamentos dos atores virtuais. Embora haja impasse promovido pela incapacidade de diferenciação do caráter opinativo dos *tweets*, isto é, se o *tweet* que cita um candidato de fato o apoia ou não o apoia, as análises estatísticas aplicadas demonstraram um elevado índice de concordância quantitativa entre os dados obtidos por este estudo e os dados reais disponibilizados pelo Tribunal Superior Eleitoral. Tal fato implica que em determinado ponto de vista, os processos aplicados condizem bem com a realidade, denunciando que o meio virtual é capaz de representar o mundo real, fornecendo, assim, a possibilidade de encontrar tendências para um determinado espaço temporal futuro, de efetuar projeções e de coletar dados afim de efetuar um panorama geral de um determinado caso (igualmente como foi conduzido este projeto).

Ainda que o evento escolhido tenha envolvido a área de Ciências Sociais Aplicadas, o método apresentado poderia ser aplicado a outros temas, envolvendo outras áreas do conhecimento. Poder-se-ia ainda utilizar-se do método afim de desenvolver uma pesquisa de dados a partir de outras bases de dados públicos, como do Google (e, neste caso, seria necessária a utilização de outras bibliotecas nos *scripts* feitos em Python). Visto tais fatos, o presente projeto carrega uma notável importância no que se diz respeito à contribuição científica em demais áreas, destacando o seu caráter interdisciplinar.

Não obstante, há possibilidades de implementação ao sistema desenvolvido, de modo a otimizar o desempenho dos processos. O desenvolvimento de outras funcionalidades, que, por exemplo, resolveriam o impasse do caráter opinativo dos *tweets*, bem como a criação de sistemas automatizados e de auxílio ao usuário, como UI (*user interface*) e banco de dados, promoveriam a criação de um software específico orientado a este fim, de modo a servir como ferramenta de pesquisa mais prática.

Assim, embora o método proposto neste artigo possua limitações, como o acesso restrito a 1% dos *tweets* publicados na rede e a necessidade de análise do sentimento das postagens, ele permite novas abordagens, as quais podem auxiliar o avanço de pesquisas em diversas áreas do conhecimento.

7 REFERÊNCIAS

BANKS, A. 2015 Brazil Digital Future in Focus. comScore, Inc, 2015. Disponível em: <https://www.comscore.com/por/layout/set/popup/content/download/29805/1525207/version/6/fi le/2015+Brasil+Digital+Future+in+Focus+PORBR.pdf>. Acessado em: 27 jun. 2016.

BENEVENUTO, F.; ALMEIDA, J. M.; SILVA, A. S. Explorando redes sociais online: Da coleta e análise de grandes bases de dados às aplicações. Minicursos do Simpósio Brasileiro de Redes de Computadores (SBRC), 2011.

BERTOL, S. R. S.; BACALTCHUCK, B.; MEZZAROBA, M. P. Campanha Eleitoral na Internet: Uma Análise do Twitter dos Candidatos à Presidência Dilma Roussef e José Serra. Revista Democracia Digital e Governo Eletrônico (ISSN 2175-9391), no. 5, 2011, pp. 172-185.

BACHINI, N. As cibercampanhas no Brasil: uma análise dos Twitters de Dilma, Serra e Marina em 2010. Revista ponto-e-vírgula, vol. 12, 2013, pp. 135-164.

BRUNS, A.; BURGESS, J. Researching news discussion on Twitter: New methodologies. Journalism Studies, 13.5-6, 2012, pp. 801-814.

CARVALHO, C.S.; GOYA, D. H.; PENTEADO, C. L. C. The people have spoken: Conflicting Brazilian Protests on Twitter. 49th Hawaii International Conference on System Sciences, 2016, pp. 1986–1995.

COELHO, F. C. (2007) "Computação Científica com Python: Uma introdução à programação para cientistas". ISBN: 9 78-85-907346-0-4. Petrópolis, RJ, Brasil.

FARIA, A. Quais são as Redes Sociais mais usadas no Brasil, 2015. Blog Análise Digital, 2015. Disponível em: <http://analise.digital/blog/informacao/quais-sao-as-redes-sociais-mais-usadasno-brasil/>. Acesso em: 19 jun. 2016.

FRANÇA, T.C.; OLIVEIRA, J. Análise de Sentimento de Tweets Relacionados aos Protestos que ocorreram no Brasil entre Junho e Agosto de 2013. BraSNAM - III Brazillian Workshop on Social Networks Analysis and Mining, 2014, pp. 128-139.

GOMES, B. C. K., BASSO, C. A. M. (2015) "Desempenho de Banco de Dados Não Relacionais com Big Data". CONTECSI - International Conference on Information Systems and Technology Management ISSN 1041-2448. São Paulo, SP, Brasil.

GOMES, W., et al. "Politics 2.0": a campanha online de Barack Obama em 2008. Rev. Sociol. Polit., 2009, vol.17 no.34, pp. 29-43.

KOLIVER, C., DORNELES, R. V., CASA, M. E. (2004) "Das (muitas) dúvidas e (poucas) certezas do ensino de algoritmos". XII Workshop de Educação em Computação (WEI'2004). Salvador, BA, Brasil.

LAUER, D. (2008) "Comparação entre Linguagens de Programação". Universidade Federal do Paraná. Curitiba, PR, Brasil.

LOPES, G.C. As redes sociais e os novos fluxos de agendamento: uma análise da cobertura da Al Jazeera sobre a Primavera Árabe. Revista Palabra Clave, vol.16, no. 3, 2013, pp. 789–811.

LOPES, R.; VIEIRA, R. Identificação de Usuários Influentes no Twitter. Universidade Federal de São João del-Rei, 2016. Disponível em: <http://www.dcomp.ufsj.edu.br/~ronanlopes/trabalhos/acmsac14.pdf/>. Acesso em: 20 jun. 2017.

MAKICE, K. Twitter API: Up and running – Learn how to build applications with the Twitter API. O'Reilly Media, 2009.

MCGRATH[1], R. Twython 3.4.0 documentation. Endereço eletrônico Twython.readthedocs, 2013. Disponível em: <https://twython.readthedocs.io/en/latest/>. Acesso em: 22 jan. 2017.

MCGRATH[2], R. Twython API Documentation. Endereço eletrônico Twython.readthedocs, 2013. Disponível em: <https://twython.readthedocs.io/en/latest/>. Acesso em: 22 jan. 2017.

MONGODB, Inc. PyMongo 3.4.0 documentation. Endereço eletrônico Api.mongodb, 2008-2015. Disponível em: <https://api.mongodb.com/python/current/>. Acesso em: 02 fev 2017.

PENTEADO, C. L. C.; GOYA, D. H.; FRANÇA, F. O. O debate político no twitter nas eleições presidenciais de 2014 no Brasil. Revista Em Debate, vol. 6, no. 6, 2014, pp. 47–54.

RECUERO, R. Contribuições da Análise de Redes Sociais para o estudo das redes sociais na Internet: o caso da hashtag #Tamojuntodilma e #CalaabocaDilma. Revista Fronteiras - Estudos Midiáticos, vol. 16, no. 2, 2014, pp. 61-77.

RECUERO, R. O twitter como esfera pública: como foram descritos os candidatos durante os debates presidenciais do 2º turno de 2014? Revista Brasileira de Linguística Aplicada, vol. 16, no. 1, 2016, pp. 157-180.

SANTOS, C. Análise da viralidade de eventos acadêmicos através das redes sociais. Universidade Federal da Bahia - Instituto de Matemática, 2014. Disponível em: <http://homes.dcc.ufba.br/~dclaro/tcc/monografia_Final_Camila.pdf>. Acesso em: 27 jun. 2016.

SANTOS, R. P., COSTA, H. A. X. (2005) "TBC-AED e TBC-AED/WEB: Um Desafio no Ensino de Algoritmos, Estruturas de Dados e Programação". Workshop de Educação em Computação e Informática do Estado de Minas Gerais (WEIMIG'2005). Lavras, MG, Brasil.

SANTOS, R. P., COSTA, H. A. X., ZAMBALDE, A. L. (2006) "Avaliação de Interfaces de Ferramentas Computacionais para o Ensino de Estruturas de Dados e Algoritmos em Grafos: Heurísticas de Usabilidade". Workshop de Educação em Computação e Informática do Estado de Minasa Gerais (WEIMIG'2006). Lavras, MG, Brasil.

TWITTER, Inc. Twitter Developer Documentation. Endereço eletrônico Twitter, 2017. Disponível em: <https://dev.twitter.com/docs>. Acesso em: 24 jan 2017.

VALIATI, H., et al. Uma estratégia baseada em difusão de informação para determinação de conteúdos relevantes e usuários influentes em redes sociais. Revista de Informática Teórica e Aplicada, vol. 20, no. 3, 2013, pp. 184-208.

VIEIRA, M.R., et al. (2012) "Bancos de Dados NoSQL: Conceitos, Ferramentas, Linguagens e Estudos de Casos no Contexto de Big Data". Simpósio Brasileiro de Bancos de Dados - SBBD 2012. Brasil.