



IEC - Instituto de Educação Continuada  
Pós-Graduação em Ciência dos Dados e Big Data

**Recuperação da Informação na Web  
e em Redes Sociais**

**Análise exploratória de dados  
coletados da rede social Twitter com  
citações à CEMIG - Companhia  
Energética de Minas Gerais**

**Aluno:** Carlos Alberto Rocha Cardoso  
**Professor:** Zilton Cordeiro Jr.

Abril  
2019



IEC - Instituto de Educação Continuada  
Pós-Graduação em Ciência dos Dados e Big Data

## Projeto Final

# Análise exploratória de dados coletados da rede social Twitter com citações à CEMIG - Companhia Energética de Minas Gerais

Trabalho apresentado ao Instituto de Educação Continuada (IEC) da pós-graduação em Ciência dos Dados e Big Data da PUC Minas, como requisito parcial para a obtenção de créditos na disciplina de Recuperação da Informação na Web e em Redes Sociais.

**Aluno:** Carlos Alberto Rocha Cardoso

**Professor:** Zilton Cordeiro Jr.

Abril  
2019

# Conteúdo

<b>1</b>	<b>Resumo</b>	<b>1</b>
<b>2</b>	<b>Introdução</b>	<b>2</b>
<b>3</b>	<b>Descrição das Atividades</b>	<b>3</b>
3.1	Coleta . . . . .	3
3.2	Armazenamento . . . . .	3
3.3	Processamento e Análise . . . . .	5
3.3.1	Leitura e formatação dos dados . . . . .	6
3.3.2	Criação dos tokens . . . . .	6
3.3.3	Atribuição de polaridade . . . . .	6
3.3.4	Plotagem de gráficos . . . . .	6
<b>4</b>	<b>Análise dos Resultados</b>	<b>7</b>
4.1	Distribuição dos tweets por data de publicação . . . . .	7
4.2	Palavras mais frequentes . . . . .	8
4.2.1	Avaliando a ocorrência da palavra CMIG4 . . . . .	10
4.2.2	Avaliando a ocorrência da palavra CAMAROTE . . . . .	11
4.3	Frequência das palavras com polaridade Negativa . . . . .	11
4.3.1	Avaliando a ocorrência da palavra VENDA . . . . .	13
4.4	Frequência das palavras com polaridade Positiva . . . . .	16
4.5	Usuários mais influentes . . . . .	17
<b>5</b>	<b>Trabalhos Futuros</b>	<b>22</b>
	<b>Bibliografia</b>	<b>23</b>

# **1 Resumo**

As redes sociais tornaram-se um rico repositório de dados, alimentado por seu uso massivo e generalizado em nossa sociedade, aplicável aos mais diversos contextos. Possuindo dados nos mais diversos formatos, muitas vezes não estruturados, a coleta, seu tratamento e análise depende muitas vezes da aplicação de conceitos e técnicas específicas. O presente trabalho demonstra o uso de técnicas e conceitos presentes nas áreas da Recuperação da Informação e a Análise de Sentimentos a partir de dados oriundos da rede social *Twitter*, relacionados à empresa CEMIG - Companhia Energética do Estado de Minas Gerais. Como resultado foi possível identificar palavras negativas e positivas, usuários mais influentes e assuntos relevantes relacionados à empresa no período analisado. O resultado deste trabalho pode servir de base para o uso especializado de dados do *Twitter* em processos corporativos da empresa como: atendimento, manutenção, comunicação e relacionamento.

## **2 Introdução**

A adoção massiva de serviços hospedados na internet está intimamente ligada ao aumento no volume, velocidade e variedade com que dados são gerados e armazenados atualmente. As organizações que historicamente priorizaram a análise e tratamento de dados estruturados, gerados por seus sistemas de informação internos, começam a descobrir o valor dos dados não estruturados, hospedados em sites de compras, fóruns, redes sociais ou sites de vídeos. Por possuir características diferentes dos dados estruturados, os dados não estruturados exigem técnicas específicas que viabilizem sua coleta, armazenamento e processamento.

Nesse contexto, com o objetivo de demonstrar a aplicação de técnicas e conceitos presentes nas disciplinas de Recuperação da Informação e Análise de Sentimentos, foram coletados dados da rede social *Twitter* relacionados à empresa CEMIG - Companhia Energética de Minas Gerais, a fim de traçar uma visão geral das citações à empresa em conjunto de publicações.

A rede social *Twitter* permite a interação entre usuários por meio da publicação de textos livres (não estruturados) de no máximo 280 caracteres, chamados de *tweet*. Para este trabalho, foram coletados todos os *tweets* a partir do termo de pesquisa "cemig", no período de 16/03/2019 a 25/03/2019. Além dos *tweets*, foram coletados alguns metadados relacionados como usuário, data e horário da publicação.

Na seção 3 são descritas todas as atividades executadas para realização do trabalho, detalhando os procedimentos de coleta, armazenamento, processamento e análise dos dados. Na seção 4 é apresentado o resultado das análises realizadas. Finalmente na seção 5 são sugeridos trabalhos futuros.

## 3 Descrição das Atividades

Nesta seção são descritas as etapas de trabalho desempenhadas para alcance dos resultados, sendo elas: coleta de dados, armazenamento, processamento e análise.

### 3.1 Coleta

Possuir dados em quantidade e qualidade suficiente é requisito chave para execução de qualquer processo que os utilize como insumo para geração de valor, exige que o dado seja cuidado e tratado em todo seu ciclo de vida, desde a etapa de geração e coleta, com o emprego de técnicas e ferramentas adequadas. Nesse contexto, a fim de coletar os dados de interesse para o tema deste trabalho, foi utilizado o sistema *Apache NiFi*. Por meio dele, foi desenvolvido um fluxo de dados capaz de se conectar à API de *Streaming* da rede social *Twitter*, coletando em tempo próximo ao real todo *tweet* publicado contendo a palavra ”cemig” no corpo do texto, no período de 17 a 24 de março de 2019. Na figura 1 é possível observar o fluxo desenvolvido, composto pelos seguintes processos:

- GetTwitter: responsável pela coleta de dados via API de *Streaming* do *Twitter*;
- EvaluateJsonPath: responsável pela seleção de atributos contidos no arquivo JSON retornado pela API do *Twitter*;
- PutMongo: Armazenamento dos *Tweets* coletados no *MongoDB*;

Sobre o *Apache Nifi*, ele é um sistema que permite a criação de fluxos para coleta, processamento e transporte de dados, a partir de componentes pré definidos e configuráveis, em uma interface gráfica baseada em comandos do tipo arrastar e soltar ou *drag and drop*.

### 3.2 Armazenamento

Para armazenamento dos dados coletados foi utilizado o *MongoDB*, um banco de dados de documentos, da família *NoSQL*. Dentre as razões para uso do *MongoDB*, está o fato de ele utilizar o formato JSON para manipulação dos documentos, mesmo formato dos dados retornados pela API do *Twitter*. Essa característica permitiu o armazenamento e consulta dos dados sem a necessidade de conversão ou transformação de formato, o que ocorreria caso fosse utilizado um banco de dados relacional como o *MySQL* por exemplo.

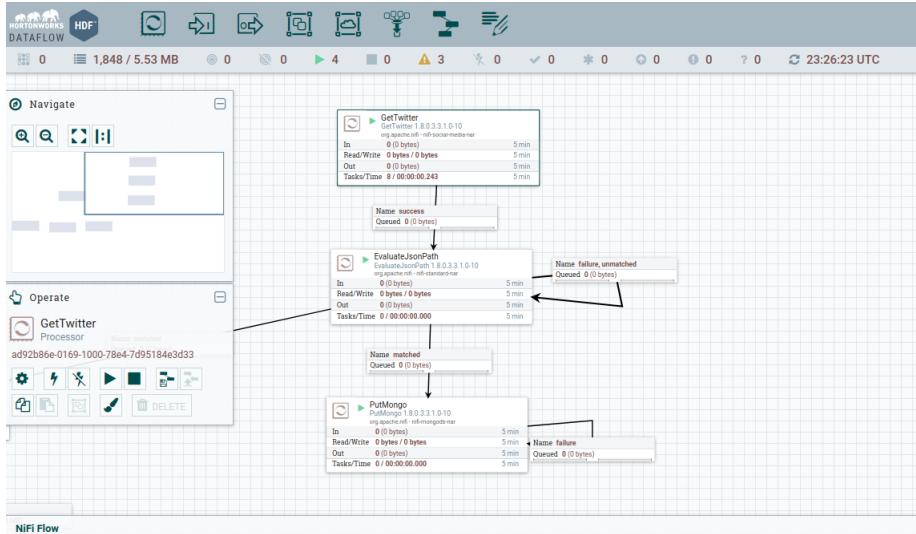


Figura 1: Fluxo de dados desenvolvido no Apache Nifi

Além disso, o *MongoDB* possui fácil integração com o *Apache NiFi* e com o *Python*, linguagem de programação utilizada para análise dos dados. Para visualizar e validar o correto armazenamento dos dados no *MongoDB*, foi utilizado o cliente *robo3t*, conforme apresentado na figura 2.

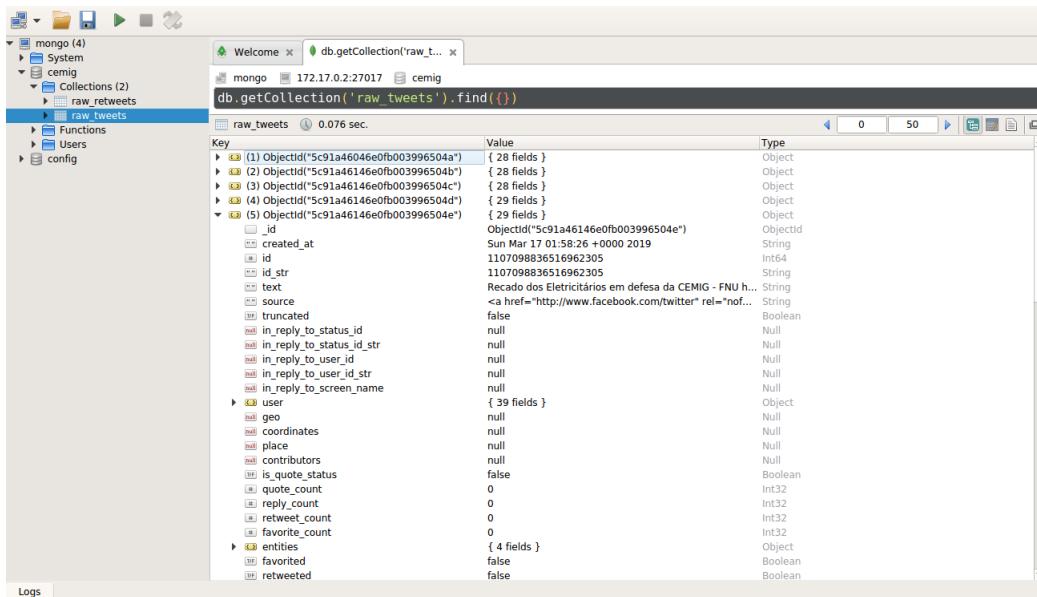


Figura 2: Interface do cliente MongoDB robo3t, listando os tweets armazenados

### 3.3 Processamento e Análise

Essa etapa teve como objetivo processar e analisar todos os tweets coletados e armazenados conforme descrição das etapas anteriores, a fim de responder às seguintes questões:

- qual a distribuição dos tweets ao longo de cada um dos dias de coleta;
- quais as palavras mais frequentes;
- quais as palavras negativas mais frequentes;
- quais as palavras positivas mais frequentes;
- qual o usuário mais influente, ou com mais tweets compartilhados (retweets);
- qual usuário com maior quantidade de tweets publicados no período;

Para suporte desta etapa, foi utilizado o software *Jupyter Notebook* integrado à linguagem de programação *Python*. O *Jupyter* combina funcionalidades de editor de texto e console para execução de códigos, permitindo o processamento e manipulação dos dados, plotagem de gráficos, descrição de análises e resultados em formato *markdown*.

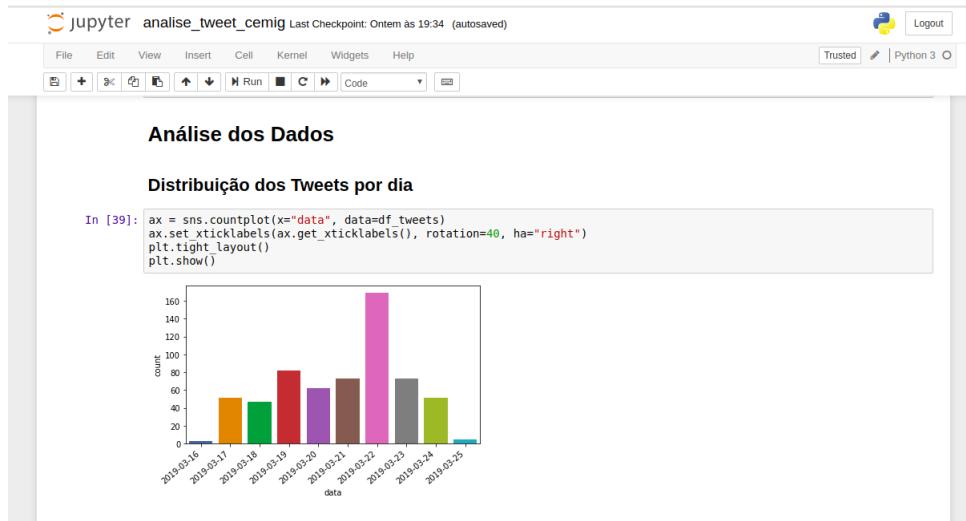


Figura 3: Interface do Jupyter Notebook

Visto que o objeto principal de análise, os *tweets*, são formados por pequenas sentenças ou frases em linguagem natural, fez-se necessária a aplicação de técnicas da recuperação da informação e processamento de linguagem natural para realização das análises pertinentes, conforme as estapas:

### **3.3.1 Leitura e formatação dos dados**

Utilizando a biblioteca *Python pymongo* foi realizada a leitura dos *tweets* armazenados no *MongoDB*. Uma vez lidos, esses registros foram carregados em uma estrutura otimizada para manipulação de dados chamada *dataframe*, presente na biblioteca *Python pandas*. Semelhante à estrutura de dados de uma planilha eletrônica, ela permite a execução de operações customizadas ou mesmo pré definidas, como agrupamentos e sumarizações. A partir desse processo, foi possível descrever de forma geral os dados relacionados aos *tweets* e avaliar algumas distribuições relacionadas à quantidade de *tweets* publicados para cada uma dos dias de coleta de dados.

### **3.3.2 Criação dos tokens**

Utilizando a biblioteca *Python nltk* e a estrutura *dataframe*, foi executada a criação de *tokens* a partir dos textos dos *tweets*. Nesse processo o texto é separado em palavras e as *stopwords* são descartadas. Chamamos de *stopwords* as palavras que não tem valor para a análise pretendida, como por exemplo preposições. A partir desse processo, foi possível realizar como exemplo a contagem de frequência das palavras nos *tweets*.

### **3.3.3 Atribuição de polaridade**

Utilizando um léxico de sentimentos para o português, chamado *SentiLex-PT*, foi feita a classificação de cada um dos *tokens*, ou palavras, quanto à sua polaridade: positiva, negativa ou neutra. Dessa forma foi possível analisar o sentimento das palavras presentes nos *tweets*, realizando por exemplo a análise de frequência das palavras negativas ou positivas.

### **3.3.4 Plotagem de gráficos**

Durante a análise foram utilizadas diferentes bibliotecas para plotagem de gráficos ou tabelas. Para apresentar a contagem de frequência de palavras por exemplo, foi utilizado tanto o *dataframe* da biblioteca *pandas* quanto o gráfico de nuvem de palavras utilizando a biblioteca *wordcloud*. Para apresentar a distribuição de dados como a quantidade de *tweets* por dia ou por usuário, foram utilizadas gráficos disponibilizados pelas bibliotecas *seaborn* e *matplotlib*.

## 4 Análise dos Resultados

A figura 4 apresenta os dados utilizados para análise e produção de resultados, após conclusão das etapas de coleta, armazenamento e processamento, sendo:

- **id**: o id do tweet;
- **tokens**: palavras relevantes para análise, extraídas do texto dos tweets;
- **polaridade**: classificação dos tokens, ou palavras, nas categorias negativas (-1), positivas (1) e neutras (0), com base no dicionário de sentimentos;
- **timestamp\_ms**: timestamp de publicação do tweet;
- **user.followers\_count**: quantidade de seguidores do usuário;
- **user.friends\_count**: quantidade de usuários seguidos;
- **user.location**: localização do usuário;
- **user.screen\_name**: identificação do usuário no Twitter;
- **data**: data de publicação do tweet;
- **user\_retweeted**: quando trata-se de um retweet, o usuário do tweet original;

Durante o desenvolvimento da análise dos dados, buscou-se responder às questões apresentadas na seção 3.3 deste documento, conforme as seguintes etapas:

### 4.1 Distribuição dos tweets por data de publicação

O objetivo desta etapa foi identificar qual a distribuição dos *tweets* ao longo dos dias em que a coleta de dados foi realizada, executando para isso a contagem por data de publicação. O resultado foi plotado em um gráfico de colunas conforme observado na figura 5. Nele pode-se observar que o dia 22/03/2019 se destaca frente aos demais, possuindo praticamente o dobro de *tweets* publicados quando comparado ao dia 19/03/2019, segundo dia com mais *tweets* no gráfico. Mais adiante, verificou-se que esse fato tem relação com dois eventos: a divulgação da uma compra de fatia da empresa Renova pela empresa *CEMIG* e a divulgação de uma notícia polêmica, relacionada

	tokens	polaridade	timestamp_ms	user.followers_count	user.friends_count	user.location	user.screen_name	data	user_retweeted
id									
1107042314097229824	ta	0	1552774430189	507	423	Minas Gerais, Brasil	MeloSahra	2019-03-16	
1107042314097229824	vacilona	0	1552774430189	507	423	Minas Gerais, Brasil	MeloSahra	2019-03-16	
1107042314097229824	hoje	0	1552774430189	507	423	Minas Gerais, Brasil	MeloSahra	2019-03-16	
1107042314097229824	assim	0	1552774430189	507	423	Minas Gerais, Brasil	MeloSahra	2019-03-16	
1107042314097229824	dá	0	1552774430189	507	423	Minas Gerais, Brasil	MeloSahra	2019-03-16	
1107052443580420096	ficar	0	1552776845246	152	344	Governador Valadares, Brasil	R_Freitas21	2019-03-16	
1107052443580420096	luz	0	1552776845246	152	344	Governador Valadares, Brasil	R_Freitas21	2019-03-16	
1107052443580420096	foda	0	1552776845246	152	344	Governador Valadares, Brasil	R_Freitas21	2019-03-16	
1107052443580420096	loja	0	1552776845246	152	344	Governador Valadares, Brasil	R_Freitas21	2019-03-16	
1107053694523228160	zuando	0	1552777143494	382	295	ciumentolandia	sou_geise	2019-03-16	

Figura 4: Dataframe com os dados para análise

ao aluguel de um camarote no estádio independência. Por fim, tem-se dias 16/03/2019 e 25/03/2019, cuja baixa quantidade de *tweets* pode ser justificada pelo curto período de tempo em que o sistema de coleta de dados esteve em funcionamento durante essas datas.

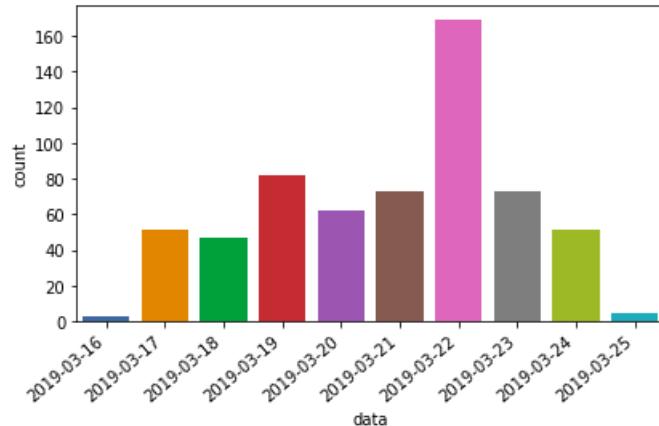


Figura 5: Distribuição dos tweets por data de publicação

## 4.2 Palavras mais frequentes

O objetivo desta etapa foi identificar quais as palavras mais frequentes nos *tweets* relacionados ao termo "cemig", executando para isso a contagem

de frequência das palavras contidas nos *tweets*, após processamento e seleção dos *tokens*. O resultado foi plotado em um ranking com a seleção das 10 palavras mais frequentes conforme figura 6, e no gráfico de nuvem de palavras conforme figura 7. A partir dessa seleção, realizou-se uma análise a fim de identificar a razão de ocorrência de algumas dessas palavras.

energia	61
luz	60
cmig4	54
cmig-nl	49
mil	39
renova	37
camarote	36
romeuzema	33
anos	32
independência	31

Figura 6: Ranking das 10 palavras mais frequentes

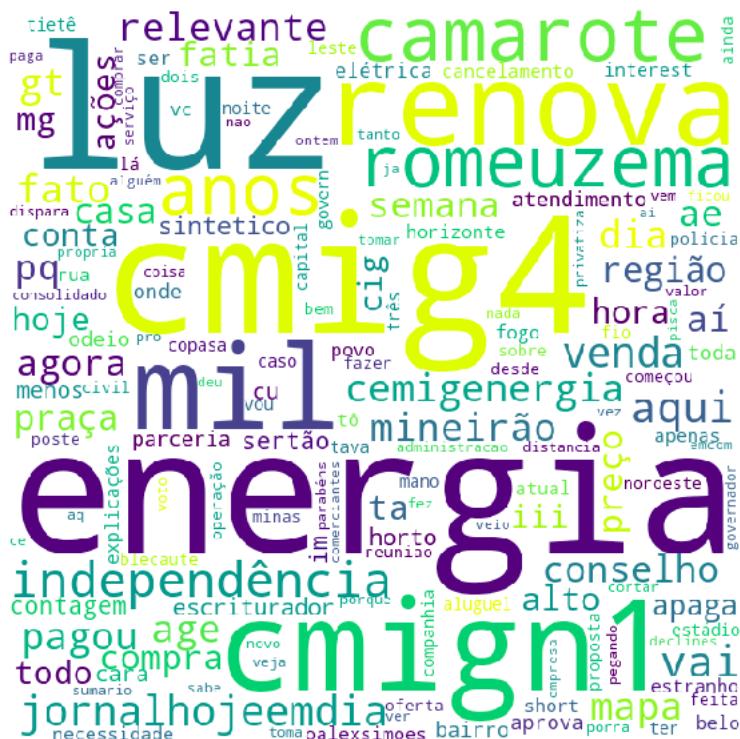


Figura 7: Nuvem das palavras mais frequentes

#### 4.2.1 Avaliando a ocorrência da palavra CMIG4

O termo *CMIG4* faz referência às ações da empresa CEMIG Holding S.A. negociadas na B3 - Bolsa de Valores de São Paulo. Sua alta ocorrência tem relação com o informe de fato relevante encaminhado pela CEMIG ao mercado de ações no dia 22/03/2019, conforme observado na lista de *tweets* relacionados, e no seu gráfico de ocorrências por data.

Figura 8: Amostra de tweets onde ocorre o tempo CMIG4

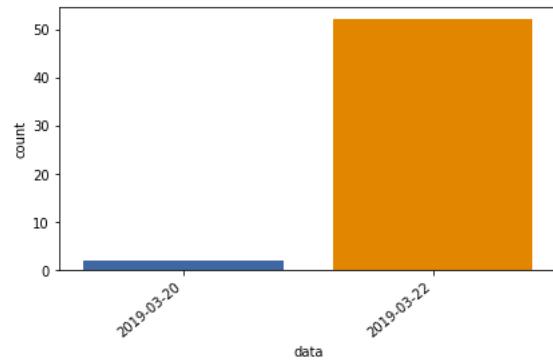


Figura 9: Ocorrência do termo CMIG4 nas datas de coleta

#### 4.2.2 Avaliando a ocorrência da palavra CAMAROTE

O termo tem relação com notícia veiculada pelo jornal Hoje em Dia, propagada entre os dias 22 e 24 de março. Segundo o jornal, a CEMIG teria pago o valor de 990 mil por 3 anos de aluguel de um camarote no estádio independência. Ainda segundo o jornal, o custo seria de 390 mil, caso o aluguel fosse contratado para o mesmo período no estádio Mineirão.

Podemos observar que outros termos que aparecem na lista dos mais frequentes tem relação com esse mesmo fato: [mil, independência].

text	user.screen_name	data	user_retweeted
Cemig pagou R990milpor3anosdecamarotenoIndependência; noMineirão, preçoseriaR 390 mil\n	jornalhojeemdia	2019-03-22	
RT @jornalhojeemdia: Cemig pagou R 990milpor3anosdecamarotenoIndependência; noMineirão, preçoseriaR 390 mil\n	REGINALDOGALO10	2019-03-22	jornalhojeemdia
RT @jornalhojeemdia: Cemig pagou R 990milpor3anosdecamarotenoIndependência; noMineirão, preçoseriaR 390 mil\n	digaormf	2019-03-22	jornalhojeemdia
RT @jornalhojeemdia: Cemig pagou R 990milpor3anosdecamarotenoIndependência; noMineirão, preçoseriaR 390 mil\n	rafael_fr2	2019-03-22	jornalhojeemdia
RT @jornalhojeemdia: Cemig pagou R 990milpor3anosdecamarotenoIndependência; noMineirão, preçoseriaR 390 mil\n	sracansada	2019-03-22	jornalhojeemdia

Figura 10: Amostra de tweets onde ocorre o tempo CAMAROTE

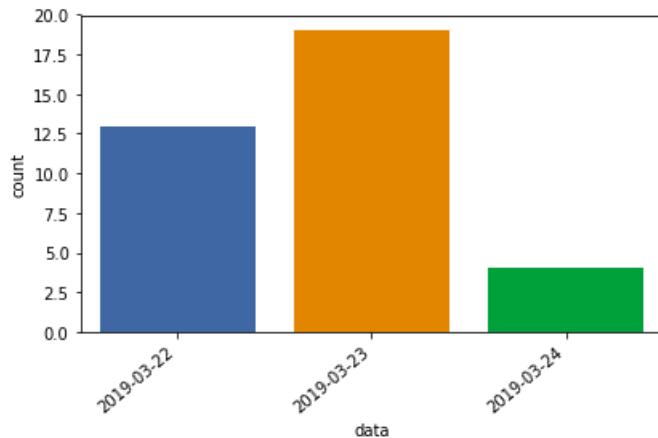


Figura 11: Ocorrência do termo CAMAROTE nas datas de coleta

#### 4.3 Frequência das palavras com polaridade Negativa

O objetivo desta etapa foi identificar quais as palavras de polaridade negativa são mais frequentes nos *tweets* relacionados ao termo "cemig". Para isso,

foi executada a contagem de frequência das palavras de polaridade negativa contidas nos textos dos *tweets*, a partir do processamento do texto, seleção dos *tokens* e atribuição de polaridade com base no dicionário de sentimentos. O resultado foi plotado em um ranking com a seleção das 10 palavras mais frequentes conforme figura 12, e no gráfico de nuvem de palavras conforme figura 13.

venda	23
apaga	14
odeio	12
estranho	10
pisca	6
apagava	5
caiu	5
derrotas	5
falta	5
vender	5

Figura 12: Ranking das 10 palavras negativas mais frequentes



Figura 13: Nuvem das palavras negativas mais frequentes

#### 4.3.1 Avaliando a ocorrência da palavra VENDA

Analizando a distribuição da palavra VENDA por data, conforme imagem 14, e os textos dos *tweets* relacionados, verificou-se os principais fatos relacionados à ocorrência da palavra. O primeiro é a informação propagada no dia 22/03/2019, a venda por parte da CEMIG do complexo eólico de Alto Sertão III, conforme imagens 15 e 16. Esse fato está relacionado ao comunicado enviado pela empresa ao mercado, no mesmo dia 22/03/2019, mesma razão de ocorrência da palavra *CMIG4*, avaliada em etapas anteriores.

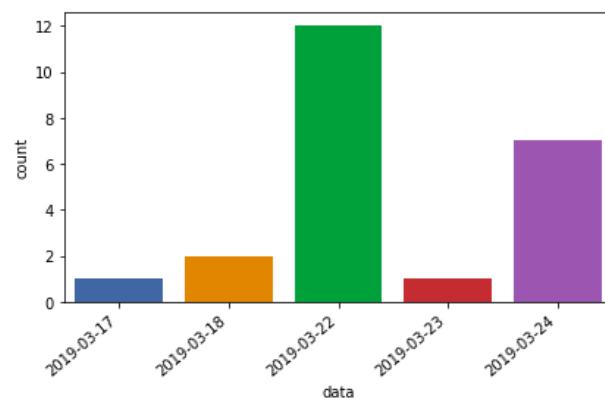


Figura 14: Ocorrência da palavra VENDA nas datas de coleta



Figura 15: Exemplo de um tweet sobre a compra de fatia da empresa Renova

1109061195003236354	Conselho da Cemig GT aprova compra de fatia na Renova e venda de Alto Sertão III -	jornaltijucas	2019-03-22
1109062989443973120	Conselho da Cemig GT aprova compra de fatia na Renova e venda de Alto Sertão III	INSTARBrazil	2019-03-22
1109063169673170944	Conselho da Cemig GT aprova compra de fatia na Renova e venda de...	bomfeli	2019-03-22
1109065486787112965	Conselho da Cemig GT aprova compra de fatia na Renova e venda de Alto Sertão III	jornalfloripa	2019-03-22

Figura 16: Tweets sobre a compra de fatia da empresa Renova

No mesmo dia 22/03/2019, foram propagados *tweets* com a notícia de que a CEMIG está colocando a venda alguns de seus imóveis, conforme a imagem 17.



Figura 17: Tweet sobre venda de imóveis da CEMIG

Por fim, no dia 24/03/2019, temos a reação de alguns usuários a um *tweet* do governador de Minas Gerais, Romeu Zema. Eles publicaram *tweets* solicitando ao governador que, aproveitando a ocasião de venda de uma aeronave do estado, vendesse (ou privatizasse) a CEMIG, conforme imagens 18 e 19.

 **Romeu Zema**  @RomeuZema · 24 de mar  
Pessoal, conforme compromisso assumido, acabamos de vender o primeiro avião da frota aérea que servia aos ex-governadores de Minas Gerais. Mais aeronaves serão vendidas. No meu governo não haverá espaço para esse tipo de mordomia, privilégios ou desperdício de dinheiro público.



 **AssisMarinhodf**  @AssisMarinhodf · Seguir · v  
Em resposta a @RomeuZema  
**Aproveita e venda a CEMIG!**  
21:12 - 23 de mar de 2019 de Jaraguá, Brasil

Figura 18: Exemplo de tweet solicitando ao Governador Romeu Zema a venda da CEMIG

1109203426230046721	Cemig divulga edital para venda de imóveis em Minas Gerais e Goiás. 	uberlandiapress	2019-03-22
1109564188907720709	Cemig divulga edital para venda de imóveis em Minas Gerais e Goiás	alo_überlandia	2019-03-23
1109608878986141702	@RomeuZema Aproveita e venda a CEMIG!	AssisMarinhodf	2019-03-24
1109636614404390914	Cemig divulga edital para venda de imóveis em MG e Goiás, incluindo terrenos Nepomuceno e Itutinga...	jlavras	2019-03-24
1109737432872824834	RT @AssisMarinhodf: @RomeuZema Aproveita e venda a CEMIG!	sergiaugusto AssisMarinhodf	2019-03-24
1109765863811571712	RT @AssisMarinhodf: @RomeuZema Aproveita e venda a CEMIG!	mordoceng AssisMarinhodf	2019-03-24
1109767045560877058	@RomeuZema Parabéns! Venda tudo! Aguardando a venda da Cemig.	alexhpf	2019-03-24

Figura 19: Tweets sobre venda de imóveis e solicitando venda da CEMIG

## 4.4 Frequênciadas palavras com polaridade Positiva

O objetivo desta etapa foi identificar as palavras com polaridade positiva mais frequentes, nos *tweets* relacionados ao termo de pesquisa "cemig". Para isso, foi executada a contagem de frequência das palavras de polaridade positiva contidas nos textos dos *tweets*, a partir do processamento do texto, seleção dos *tokens* e atribuição de polaridade com base no dicionário de sentimentos. O resultado foi plotado em um ranking com a seleção das 10 palavras mais frequentes conforme figura 20, e no gráfico de nuvem de palavras conforme figura 21.

energia	63
relevante	21
apaga	14
belo	10
bem	8
apagava	5
derrotas	5
boa	3
certíssimo	3
eficiente	3

Figura 20: Ranking das 10 palavras positivas mais frequentes



Figura 21: Nuvem das palavras positivas mais frequentes

Analizando de forma geral as palavras positivas mais frequentes, observa-se que a ocorrência da palavra energia é bem superior às demais, com presença constante em todos as datas de coleta. Avaliando os *tweets* de origem da palavra, observa-se que sua ocorrência está relacionada na grande maioria à reclamações de falta ou queda de energia, dessa forma, embora tenha uma polaridade positiva, no contexto CEMIG ela está geralmente relacionada a *tweets* que expressam sentimentos negativos.

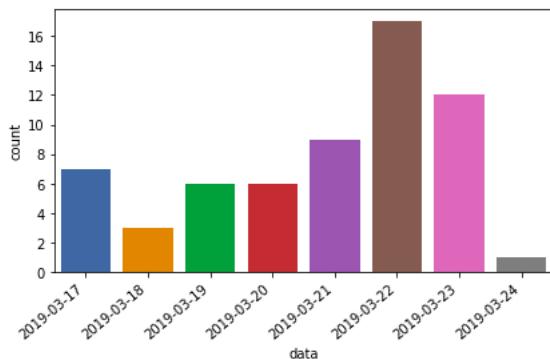


Figura 22: Ocorrência da palavra ENERGIA nas datas de coleta

text	data
CEMIG Serviços S.A., vocês conseguiram me recomendar alguma empresa de fornecimento de energia elétrica? A empresa...	2019-03-17
ta faltando energia aqui Cemig \nbunita	2019-03-17
caiu a energia e vou dormir no valor, valeu Cemig.	2019-03-17
que maravilha ein cemig sem estar chovendo to sem energia	2019-03-17
Eu tô pensando é como vamos trabalhar na segunda pós sunsetville, mas tudo bem \nCemig vai ter que me dar energia ☺	2019-03-17
Cemig eu te odeio pq se faltar energia vou surtar nem jantei ainda odeio comer no escuro	2019-03-17
RT @helena__carvihlo: Cemig eu te odeio pq se faltar energia vou surtar nem jantei ainda odeio comer no escuro	2019-03-17
Caralho o motivo de ter acabado a energia é q pego fogo no negócio da Cemig slc bicho q medo desses troço	2019-03-18
@cemig_energia RT @cemig_energia #Efficientia: A #empresa de #serviços de #energia da #Cemig, vai investir até R\$ 5...	2019-03-18
Cemig desgraçada. Mais de 6 horas sem energia. E sem aviso prévio	2019-03-18

Figura 23: Amostra de tweets contendo a palavra energia

## 4.5 Usuários mais influentes

Para identificação dos usuários mais influentes relacionados ao termo de pesquisa "cemig", foi feita a contagem do *retweets* recebidos por cada usuário.

O resultado foi plotado em um ranking com a seleção dos 10 mais *retweetados* conforme figura 24.

user_retweeted	
jornalhojeemdia	19
oalexsimoes	9
cemig_energia	6
em_com	6
AnáliseEnergia	5
GeraldodeMorais	3
gustavonolascoB	3
lopaugomes	3
lcerdz	3
AssisMarinhodf	2

Figura 24: Ranking dos 10 usuários mais retweetados ou mais influentes

Para visualizar os usuários que foram influenciados, ou que *retweetaram* os usuários influenciadores, foram plotados grafos conforme imagem 25.

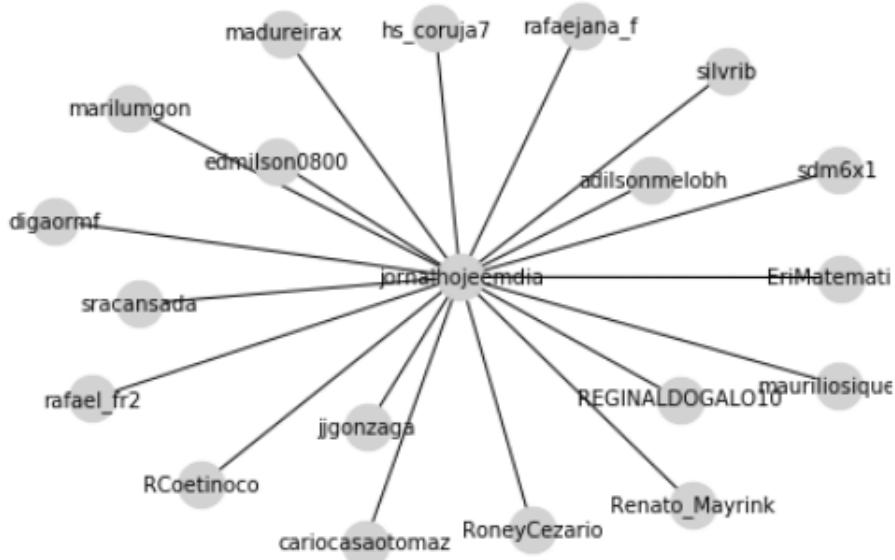


Figura 25: Grafo de usuários que retweetaram o usuário jornalhojeemdia

Para cada grafo de usuários influenciados e influenciador, foi feita a listagem de *tweets* relacionados a fim de identificar os assuntos relacionados à propagação. Para os dois primeiros usuários do ranking de influenciadores, *jornalhojeemdia* e *oalexsimoes*, o assunto foi restrito ao caso do camarote alugado pela CEMIG, já descrito em etapas anteriores de análise.

 **Jornal Hoje em Dia**   
@jornalhojeemdia 

Cemig pagou R\$ 990 mil por 3 anos de camarote no Independência; no Mineirão, preço seria R\$ 390 mil



Cemig pagou R\$ 990 mil por 3 anos de camarote no Independência; no Mineirão, preço seria R\$ 390 mil  
Pelo  
[hojeemdia.com.br](http://hojeemdia.com.br)

Figura 26: Tweet publicado pelo usuário jornalhojeemdia

 **Alexandre Simoes**  
@oalexsimoes 

O cancelamento, há apenas uma semana, não apaga a necessidade de explicações. Muito estranho!!!! Pelo menos o atual governo percebeu o desperdício de dinheiro público, pois o valor estava exagerado.



Cemig pagou R\$ 990 mil por 3 anos de camarote no Independência; no Mineirão, preço seria R\$ 390 mil  
Pelo  
[hojeemdia.com.br](http://hojeemdia.com.br)

Figura 27: Tweet publicado pelo usuário oalexsimoes

Para o usuário *em\_com*, o assunto tem relação com ação realizada pela Polícia Civil em parceria com a CEMIG, com objetivo de coibir o roubo de

cabos e equipamentos da empresa.

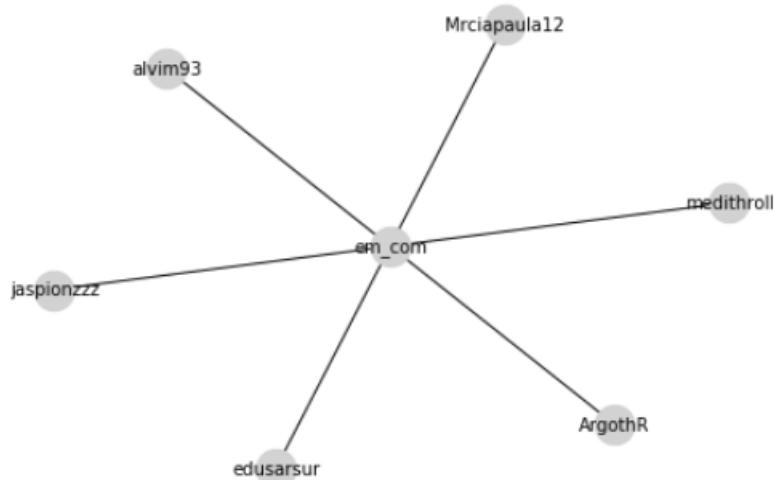


Figura 28: Grafo de usuários que retweetaram o usuário em\_com



Figura 29: Tweet publicado pelo usuário em\_com

Outra análise realizada para identificação de usuários influentes, foi a criação de um ranking dos 10 usuários com mais seguidores conforme imagem 30.

exame	2390173
em_com	497706
OficialBHTRANS	338156
otempo	276193
epocanegocios	209312
jornalhojeemdia	144154
portaluai	87909
CIObrasil	25533
cemig_energia	21620
SunoResearchcom	19466

Figura 30: Ranking dos 10 usuários com mais seguidores

Para esses usuários foram listados os *tweets* publicados, com objetivo de verificar os assuntos relacionados. Abaixo os principais:

- **exame:** Conselho da Cemig GT aprova compra de fatia na Renova;
- **em\_com:** Operação Blecaute, da Polícia Civil em parceria com a Cemig, foi feita na Região Noroeste da capital;
- **otempo:** Marco Antônio Lage assume comunicação da Cemig mirando gestão eficiente;
- **epocanegocios:** Antes dos anúncios desta semana, a Renova já havia recusado uma oferta da canadense Brookfield e da própria AES;
- **jornalhojeemdia:** Cemig pagou R990mil por 3anos de camarote no Independência;

## 5 Trabalhos Futuros

Em trabalhos futuros, sugere-se a aplicação de algoritmos de *machine learning* para aperfeiçoar a análise e atribuição de sentimentos aos textos coletados da rede social *Twitter*, utilizando uma base histórica de textos previamente classificados quanto à sua polaridade: negativo, positivo ou neutro. Outra possibilidade é a identificação de bigramas ou trigramas nos textos dos *tweets*, com o objetivo identificar expressões comuns e relevantes para o contexto de análise, exemplo: cortou luz, caiu energia, queimou equipamento. Os resultados apresentados neste trabalho podem servir como base para o uso especializado de dados oriundos da rede social twitter, aperfeiçoando processos de atendimento, manutenção, comunicação, reputação da marca, dentre outros.

## Bibliografia

DE AGUIAR, E. J.; FAIÇAL B. S.; UNEYAMA, J.; SILVA, G. C.; MENOLLINI, A. Análise de Sentimento em Redes Sociais para a Língua Portuguesa Utilizando Algoritmos de Classificação. In: SIMPÓSIO BRASILEIRO DE REDES DE COMPUTADORES E SISTEMAS DISTRIBUÍDOS, 36., 2018, Porto Alegre.

LIMA, V. R. Utilizando processamento de linguagem natural para criar um sumarização automática de textos. Disponível em: <<https://medium.com/@viniljf/utilizando-processamento-de-linguagem-natural-para-criar-um-sumariza%C3%A7%C3%A3o-autom%C3%A1tica-de-textos-775cb428c84e>>. Acesso em: 08 abr. 2018.

REVERT, F. Getting started with graph analysis in Python with pandas and networkx. Disponível em: <<https://towardsdatascience.com/getting-started-with-graph-analysis-in-python-with-pandas-and-networkx-5e2d2f82f18e>>. Acesso em: 08 abr. 2018.

CARVALHO, P.; SILVA, M. J. Léxico de sentimentos para o Português Sentilex-PT. Disponível em: <<https://b2share.eudat.eu/records/93ab120efdaa4662baec6adee8e7585f>>. Acesso em: 08 abr. 2018.

HORTONWORKS. Building a Sentiment Analysis Application Acquiring Twitter Data. Disponível em: <<https://br.hortonworks.com/tutorial/building-a-sentiment-analysis-application/section/3/>>. Acesso em: 08 abr. 2018.

SANTANA, R. Análise de Sentimentos de uma forma diferente. Disponível em: <<http://minerandodados.com.br/index.php/2018/05/15/analise-de-sentimentos-de-uma-forma-diferente/>>. Acesso em: 08 abr. 2018

# Anexo

Notebook contendo os códigos e análises executados

**IMPORTANTE:** O passo 1.1 contém os comandos utilizados para consultar e tratar os dados no MongoDB. Na ausência do banco de dados Mongo com os dados de origem, **pular o passo 1.1**. No passo 1.3 os dados tratados são lidos de uma planilha CSV, dispensando o uso do MongoDB.

In [1]:

```
import pandas as pd
import numpy as np

from pymongo import MongoClient

from bson import json_util, ObjectId
from pandas.io.json import json_normalize
import json

import seaborn as sns
import matplotlib.pyplot as plt

from wordcloud import WordCloud

import networkx as nx

pd.set_option('display.max_colwidth',300)
```

In [2]:

```
#Converter para coluna campos dos subdocumentos
def mongo_to_dataframe(mongo_data):
    sanitized = json.loads(json_util.dumps(mongo_data))
    normalized = json_normalize(sanitized)
    df = pd.DataFrame(normalized)

    return df

#Plotar gráfico de colunas
def draw_graph_count(df, col):
    ax = sns.countplot(x=col, data=df)
    ax.set_xticklabels(ax.get_xticklabels(), rotation=40, ha="right")
    plt.tight_layout()
    plt.show()

#Plotar nuvem de palavras
def draw_word_cloud(tokens):
    texto_cloud = ' '.join(tokens)
    wc = WordCloud(width = 600, height = 600, background_color ='white', min_font_size = 10, collocations=False).generate(texto_cloud)
    plt.figure(figsize = (8, 8), facecolor = None)
    plt.imshow(wc)
    plt.axis("off")
    plt.tight_layout(pad = 0)
    plt.show()

#Plotar gráfico de rede ou grafo
def draw_network(df, nodes, source, target):
    G = nx.from_pandas_edgelist(df=df, source=source, target=target)
    G.add_nodes_from(nodes_for_adding=nodes.tolist())
    nx.draw(G, with_labels = True, node_color='lightgrey', node_size = 500, font_size = 10)
```

## 1. Carga e preparação dos dados

### 1.1 Consulta e preparação dos Tweets salvos no Mongo

Conexão e consulta dos tweets salvos mongodb a partir da coleta via streaming utilizando o Apache NiFi.  
Carga dos dados no pandas dataframe.

In [ ]:

```
host = '172.17.0.2'
port = 27017
collection = 'cemig'
db_name = 'cemig'
collection = 'raw_tweets'
collection_rt = 'raw_retweets'

conn = MongoClient(host, port)
db = conn[db_name]
cursor = db[collection].find({}, {"_id":0,"id":1,"timestamp_ms":1,"text":1,'user.screen_name':1,'user.location':1,'user.followers_count':1,'user.friends_count':1})
df_tweets = mongo_to_dataframe(cursor)
```

Tratamento inicial dos dados como: converter de timestamp para data e hora, retirar links dos textos, coletar nome dos usuários que foram retweetados, retirar pontuações

In [ ]:

```
df_tweets["tempo"] = pd.to_datetime(df_tweets['timestamp_ms'], unit='ms').dt.tz_localize('America/Sao_Paulo')
df_tweets["data"] = df_tweets["tempo"].dt.date

df_tweets["text"] = df_tweets["text"].str.replace('https:\/\.\.*\s?', '')
df_tweets["text"] = df_tweets["text"].str.replace('\r|\n', '')

df_tweets["user_retweeted"] = df_tweets["text"].str.extract('^\RT \@([\w]+)\:', expand=False)
df_tweets["user_retweeted"] = df_tweets["user_retweeted"].fillna("")
```

In [ ]:

```
from string import punctuation
```

## Salvando csv dos tweets tratados

In [ ]:

```
df_tweets.columns = ['id','text','timestamp_ms','user_followers_count','user_friends_count','user_location','user_screen_name','tempo','data','user_retweeted']
df_tweets.to_csv(path or buf='tweets.csv', index=False)
```

## 1.2 Carga do Dicionário de Palavras e Polaridade

Carga e parse do arquivo contendo palavras e polaridades Criação de um dicionário contendo palavra e polaridade

In [3]:

```
df_arq = pd.read_csv('SentiLex-PT02/SentiLex-flex-PT02.txt', header=None, names=['col'], sep='#')
df_arq = df_arq[~df_arq.col.str.contains("IDIOM")]

df_sentilex = df_arq.col.str.extract(r'^(?P<palavra>.*),.*\.(?P<polaridade>.[0-9]);ANOT', expand=True)
df_sentilex = pd.Series(df_sentilex.polaridade.values, index=df_sentilex.palavra).to_dict()
```

## 1.3 Criando DataFrame com os tokens extraídos dos Tweets

In [4]:

```
df_tweets = pd.read_csv('tweets.csv')
df_tweets[['id', 'text', 'user_followers_count', 'user_friends_count', 'user_location', 'user_screen_name', 'data', 'user_retweeted']].head(5)
```

Out[4]:

	<b>id</b>	<b>text</b>	<b>user_followers_count</b>	<b>user_friends_count</b>
<b>0</b>	1107042314097229824	CEMIG ta vacilona hoje em assim não dá	507	423
<b>1</b>	1107052443580420096	Ficar sem luz e foda essa loja da Cemig	152	344
<b>2</b>	1107053694523228160	Cemig zuando os pratence hihihi	382	295
<b>3</b>	1107072220281688064	BETHEVITAS mxrcone CEMIG curtiu isso	1461	1466
<b>4</b>	1107098836516962305	Recado dos Eletricitários em defesa da CEMIG FNU	63	1697

In [5]:

```
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize

#Definindo as stopwords
df_tokens = pd.DataFrame()
w_custom = list(['cemig','rt','pra','r','é','tá'])

sw = set(stopwords.words('portuguese') + w_custom)

#Iterando sobre os tweets para coleta dos tokens
for index, row in df_tweets.iterrows():
    palavras = word_tokenize(str(row['text']).lower())

    df_token_tmp = pd.DataFrame()
    df_token_tmp["tokens"] = [palavra for palavra in palavras if palavra not in
sw]
    df_token_tmp["id"] = row['id']

    #atribuição da polaridade com base no dicionário
    df_token_tmp["polaridade"] = df_token_tmp["tokens"].apply(df_sentilex.get)

    df_tokens = df_tokens.append(df_token_tmp, ignore_index=True)

#Cria dataframe contendo tokens e demais informações dos tweets
df_tweets_tokens = df_tokens.set_index('id').join(df_tweets.set_index('id'))
df_tweets_tokens = df_tweets_tokens.fillna('0')
```

In [6]:

```
df_tweets_tokens[['tokens','polaridade','timestamp_ms','user_followers_count','user_friends_count','user_location','user_screen_name','data','user_retweeted']].head(10)
```

Out[6]:

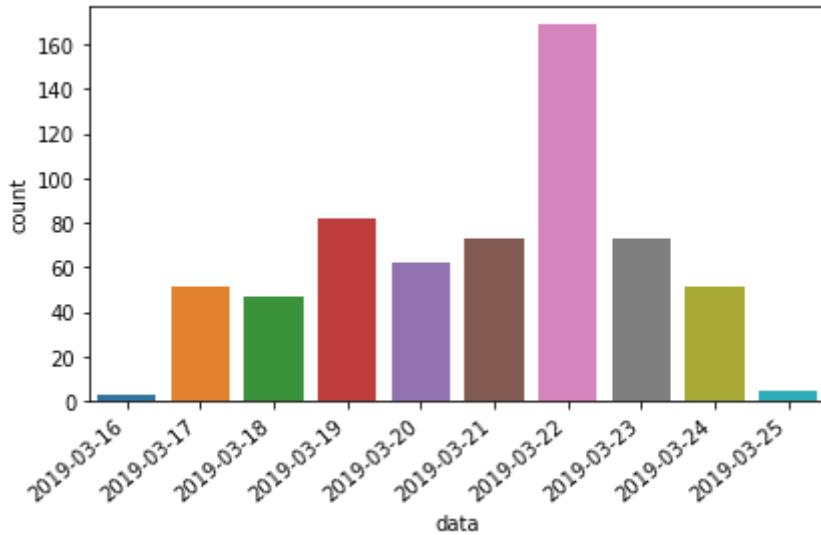
	tokens	polaridade	timestamp_ms	user_followers_count	user_friends_count
id					
1107042314097229824	ta	0	1552774430189	507	42
1107042314097229824	vacilona	0	1552774430189	507	42
1107042314097229824	hoje	0	1552774430189	507	42
1107042314097229824	assim	0	1552774430189	507	42
1107042314097229824	dá	0	1552774430189	507	42
1107052443580420096	ficar	0	1552776845246	152	34
1107052443580420096	luz	0	1552776845246	152	34
1107052443580420096	foda	0	1552776845246	152	34
1107052443580420096	loja	0	1552776845246	152	34
1107053694523228160	zuando	0	1552777143494	382	29

## Análise dos Dados

### Distribuição dos Tweets por dia

In [7]:

```
draw_graph_count(df_tweets, "data")
```



## Contagem de frequência das palavras

Listando as 10 mais frequentes

In [8]:

```
df_tweets_tokens.groupby('tokens').tokens.count().nlargest(10)
```

Out[8]:

```
tokens
energia      61
luz          60
cmig4        54
cmignl       49
mil          41
renova       38
camarote     36
anos          32
independênc 32
romeuzema    31
Name: tokens, dtype: int64
```

## Nuvem das palavras mais frequentes

In [9]:

```
draw_word_cloud(df_tweets_tokens)
```



## CMIG4: Avaliando a ocorrência da palavra

In [10]:

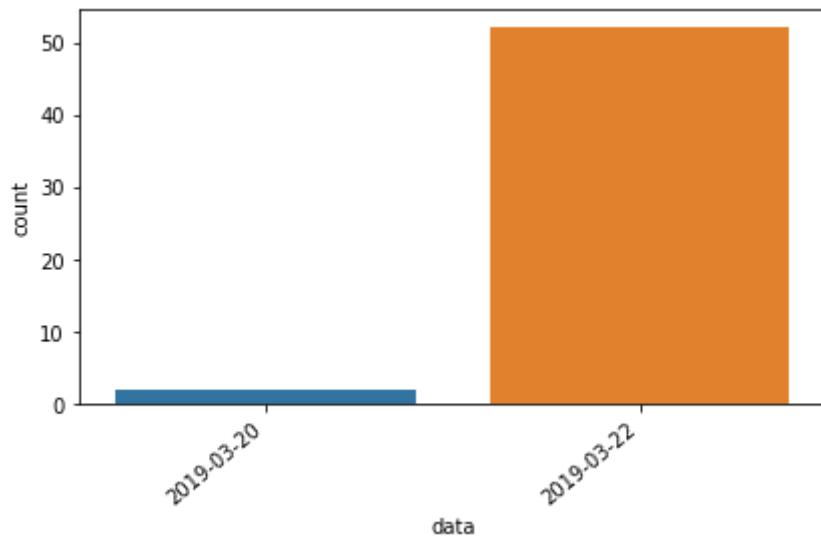
```
df_cmig4 = df_tweets_tokens[['text', 'data']].loc[df_tweets_tokens['tokens'] == 'cmig4']
df_cmig4.head(5)
```

Out[10]:

		text	data
id			
1108315694775492609	Cemig para Abril de 2019 gráfico de CMIG4		2019-03-20
1108317267865952256	CMIG4 Cemig para Abril de 2019 TradingView		2019-03-20
1108936597116903424	CMIG4 RNEW11 Celebração de Contrato para Aquisição de Participação na Renova Energia SARenova e Realizaç		2019-03-22
1109062517735731200	CMIG4 CEMIG CMIGN1 Fato Relevante 210319		2019-03-22
1109062519933554688	CMIG4 CEMIG CMIGN1 Fato Relevante 210319		2019-03-22

In [11]:

```
draw_graph_count(df_cmig4, "data")
```



O termo refere-se ao código do papel CEMIG negociado na bolsa de valores. Conforme observado na lista de tweets e data de propagação, tem relação com o informe de fato relevante encaminhado pela CEMIG ao mercado de ações.

## **CAMAROTE: Avaliando a ocorrência palavra**

In [12]:

```
df_camarote = df_tweets_tokens[['text','user_screen_name','data','user_retweete  
d']].loc[df_tweets_tokens['tokens'] == 'camarote']  
df_camarote.head(5)
```

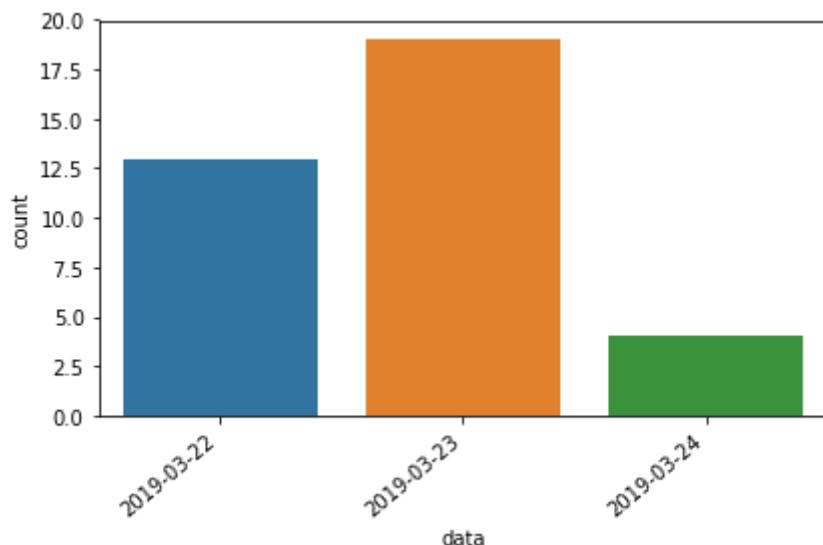
Out[12] :

	<b>text</b>	<b>user_screen_name</b>	<b>data</b>	<b>user_retweeted</b>
<b>id</b>				
<b>1109215677053837313</b>	Cemig pagou R 990 mil por 3 anos de camarote no Independência no Mineirão preço seria R 390 mil	jornalhojeemdia	2019-03-22	0
<b>1109215897464512515</b>	RT jornalhojeemdia Cemig pagou R 990 mil por 3 anos de camarote no Independência no Mineirão preço seria R 390 mil	REGINALDOGALO10	2019-03-22	jornalhojeemdia
<b>1109216983755698176</b>	RT jornalhojeemdia Cemig pagou R 990 mil por 3 anos de camarote no Independência no Mineirão preço seria R 390 mil	digaormf	2019-03-22	jornalhojeemdia
<b>1109217168451883009</b>	RT jornalhojeemdia Cemig pagou R 990 mil por 3 anos de camarote no Independência no Mineirão preço seria R 390 mil	rafael_fr2	2019-03-22	jornalhojeemdia

	text	user_screen_name	data	user_retweeted
id				
1109218880659079170	RT jornalhojeemdia Cemig pagou R 990 mil por 3 anos de camarote no Independência no Mineirão preço seria R 390 mil	sracansada	2019- 03-22	jornalhojeemdia

In [13]:

```
draw_graph_count(df_camarote, "data")
```



O termo tem relação com notícia veiculada pelo jornal Hoje em Dia, propagada entre os dias 22 e 24 de março. Segundo o jornal, a CEMIG teria pago o valor de 990 mil por 3 anos de camarote no estádio independência. Ainda segundo o jornal, o mesmo serviço poderia ter sido contratado no Minérião pelo valor de 390 mil.

Podemos observar que outros termos que aparecem na lista dos mais frequentes tem relação com esse mesmo fato: [mil, independência].

## Contagem de frequência das palavras com polaridade Negativa

Listando as 10 mais frequentes

In [14]:

```
df_tweets_tokens_neg = df_tweets_tokens.loc[df_tweets_tokens['polaridade'] == '-1']
df_tweets_tokens_neg.groupby('tokens').tokens.count().nlargest(10)
```

Out[14]:

```
tokens
venda      23
apaga      14
odeio       12
estranho    10
pisca       6
apagava     5
cair        5
caiu        5
derrotas    5
falta       5
Name: tokens, dtype: int64
```

## Nuvem das palavras mais frequentes com polaridade negativa

In [15]:

```
draw_word_cloud(df_tweets_tokens_neg.tokens)
```

**venda**  
condenada nervoso incompetência vencido barato venceu retardado  
falência assusta responsabiliza lento triste obrigada  
assuta responsabiliza lento triste obrigada  
irrita escuro vende infarto inferno imposto corrida  
vendendo vendendo parada inútil  
imóveis vergonha chorei puta apobre remprado  
derrubou fria vacilo horível bateram  
cai falta arrumar corrupção  
vende vende legal ganha inimigo  
medo caralho fodida opriindo culpa  
abusiva vendo acabaço obrogado  
calu chora apagada  
caído queima ganho  
problema arrumei deida  
péssimo vão preso queimar diminuiu arruma  
estranho recusado folgado lixo susso  
odeio caiu clandestinas apagado  
corta derrubar  
merda queimarão  
apagou base

## **VENDA: Avaliando a ocorrência palavra**

Tweets

In [16]:

```
df_venda = df_tweets_tokens_neg[['text', 'user_screen_name', 'user_retweeted', 'date']].loc[df_tweets_tokens_neg['tokens'] == 'venda']
df_venda.head(5)
```

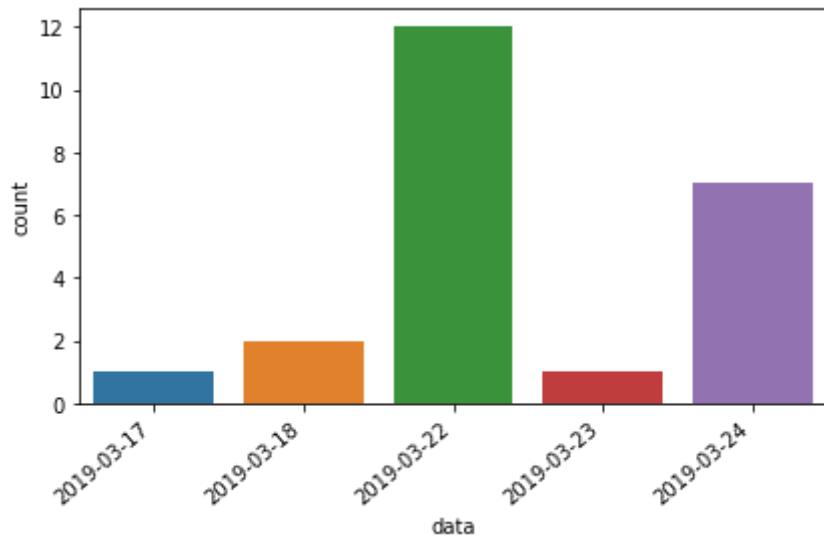
Out[16] :

	<b>text</b>	<b>user_screen_name</b>	<b>user_retweeted</b>	<b>data</b>
<b>id</b>				
<b>1107123548622123008</b>	joaoamoedonovo até agora nada do novo ter anunciado a venda das 2 estatais CEMIG E COPASA que estão oprimindo o povo	Desmio	0	2019-03-17
<b>1107656184926998528</b>	Vejam o péssimo atendimento da Cemig 20 minutos para pegar uma senha de atendimento na Agência de Venda Nova	renato_artur	0	2019-03-18
<b>1107657455771758592</b>	RT renatoartur Vejam o péssimo atendimento da Cemig 20 minutos para pegar uma senha de atendimento na Agência de Venda NovaCEMIG ht	helberthadm	renato_artur	2019-03-18
<b>1109060826328166400</b>	Conselho da Cemig GT aprova compra de fatia na Renova e venda de Alto Sertão III	jornaltijucas	0	2019-03-22
<b>1109061194986459136</b>	Conselho da Cemig GT aprova compra de fatia na Renova e venda de Alto Sertão III	jornaltijucas	0	2019-03-22

## Distribuição

In [17] :

```
draw_graph_count(df_venda, "data")
```



A alta ocorrência do termo tem relação com dois fatos. O primeiro é a informação propagada no dia 22/03, de compra por parte da CEMIG de fatia da empresa Renova, e **venda** do complexo eólico de Alto Sertão III. Esse fato está relacionado ao comunicado enviado pela empresa ao mercado, no mesmo dia 22/03, razão da grande ocorrência do termo CMIG4, avaliado anteriormente.

Outra razão para alta ocorrência do termo tem origem nos tweets propagados no dia 24/03, relacionados à notícia de que a CEMIG está vendendo alguns de seus imóveis, e a reação de usuários à essa notícia, pedindo que o governador Romeu Zema aproveite a ocasião para vender (ou privatizar) a empresa.

## APAGA e ESTRANHO: Avaliando a ocorrência das palavras

Tweets: APAGA

In [18]:

```
df_apaga = df_tweets_tokens_neg[['text', 'user_screen_name', 'user_retweeted', 'date']].loc[df_tweets_tokens_neg['tokens'] == 'apaga']
df_apaga.head(5)
```

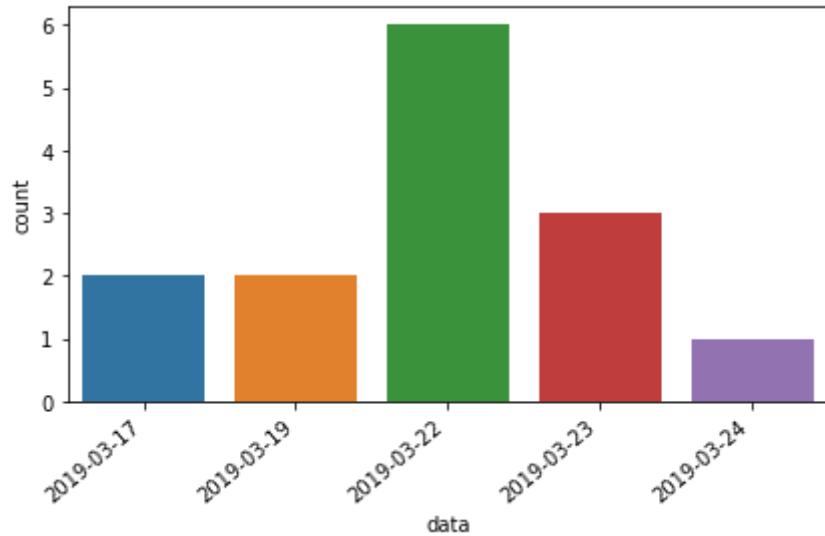
Out[18] :

	<b>text</b>	<b>user_screen_name</b>	<b>user_retweeted</b>	<b>data</b>
<b>id</b>				
<b>1107133154068971521</b>	BellaMoreira8 e a minha que tá o dia todo apaga a luz e deixa o tempo passar BABY VOCÊ NÃO PERDE POR ESPERAR APAG	JujubaMS	0	2019-03-17
<b>1107417138879438849</b>	Cemig se for pra piscar a luz de 10 em 10s apaga essa porra logo caralhoPelo menos não queima o resto do meu PC	danielpontello	0	2019-03-17
<b>1107828522687565824</b>	Luz Aq de casa tá acende apaga acende apaga Decide logo se vai ter apagão ou não cemig	PinheiroRodson	0	2019-03-19
<b>1107828522687565824</b>	Luz Aq de casa tá acende apaga acende apaga Decide logo se vai ter apagão ou não cemig	PinheiroRodson	0	2019-03-19
<b>1109210554269384705</b>	O cancelamento há apenas uma semana não apaga a necessidade de explicações Muito estranho Pelo menos o atual	oalexsimoes	0	2019-03-22

## Distribuição: APAGA

In [19]:

```
draw_graph_count(df_apaga, "data")
```



A maior ocorrência da palavra apaga está relacionada ao caso do camarote alugado pela CEMIG no independência, divulgado em um tweet do usuário @oalexsimoes

In [20]:

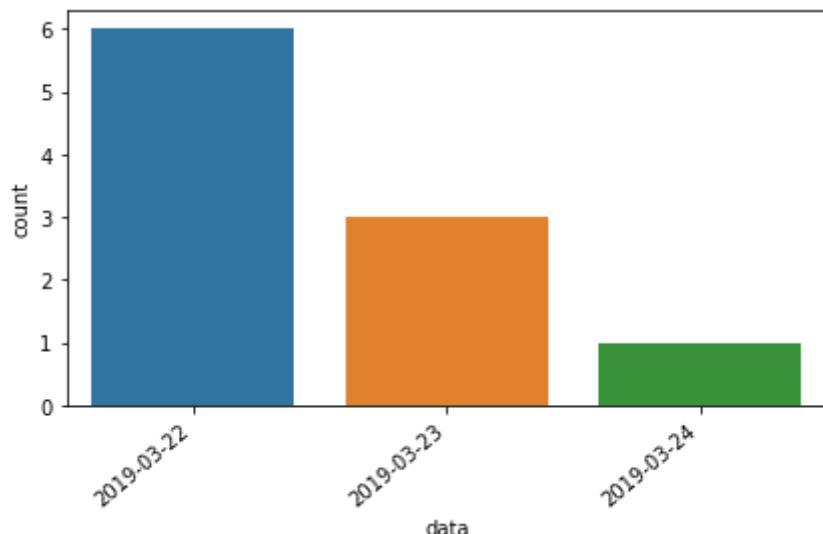
```
df_estrano = df_tweets_tokens_neg[['text','data']].loc[df_tweets_tokens_neg['tokens'] == 'estrano']
df_estrano.head(5)
```

Out[20]:

		text	data
	id		
	<b>1109210554269384705</b>	O cancelamento há apenas uma semana não apaga a necessidade de explicações Muito estranho Pelo menos o atual	2019-03-22
	<b>1109210645096996867</b>	RT oaleximoes O cancelamento há apenas uma semana não apaga a necessidade de explicações Muito estranho Pelo menos o atual govern	2019-03-22
	<b>1109211591860514816</b>	RT oaleximoes O cancelamento há apenas uma semana não apaga a necessidade de explicações Muito estranho Pelo menos o atual govern	2019-03-22
	<b>1109212636061810688</b>	RT oaleximoes O cancelamento há apenas uma semana não apaga a necessidade de explicações Muito estranho Pelo menos o atual govern	2019-03-22
	<b>1109215638982180865</b>	RT oaleximoes O cancelamento há apenas uma semana não apaga a necessidade de explicações Muito estranho Pelo menos o atual govern	2019-03-22

In [21]:

```
draw_graph_count(df_estrano, "data")
```



A palavra estranho também está relacionada ao mesmo tweet do usuário @oalexsimoes. A distribuição dos tweets que contém a palavra estranho, a partir do dia 22/03, é a mesma da palavra apaga a partir da mesma data.

## Contagem de frequência das palavras com polaridade Positiva

Listando as 10 mais frequentes

In [22]:

```
df_tweets_tokens_pos = df_tweets_tokens.loc[df_tweets_tokens['polaridade'] ==  
'1']  
df_tweets_tokens_pos.groupby('tokens').tokens.count().nlargest(10)
```

Out[22]:

```
tokens  
energia      61  
relevante    21  
belo         10  
voto          7  
boa           3  
cedo          3  
certíssimo    3  
eficiente    3  
abençoada    2  
aguenta       2  
Name: tokens, dtype: int64
```

In [23]:

```
draw_word_cloud(df_tweets_tokens_pos.tokens)
```



In [24]:

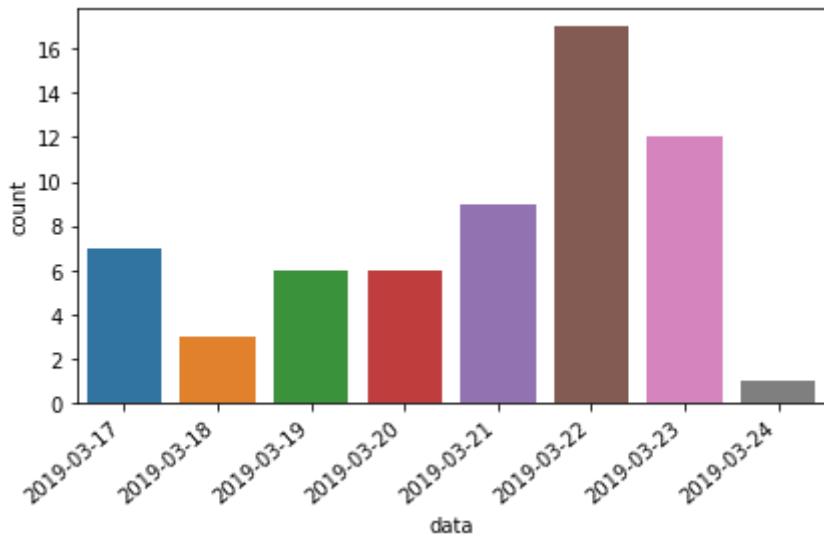
```
df_energia = df_tweets_tokens_pos[['text', 'data', 'user_location']].loc[df_tweets_tokens_pos['tokens'] == 'energia']
df_energia.head(10)
```

Out[24]:

		text	data	user_location
	id			
	<b>1107125367767207936</b>	CEMIG Serviços SA vocês conseguiram me recomendar alguma empresa de fornecimento de energia elétrica A empresa	2019-03-17	0
	<b>1107138321044647938</b>	ta faltando energia aqui Cemig bunita	2019-03-17	Timothy City
	<b>1107145095013433344</b>	caiu a energia e vou dormir no valor valeu Cemig	2019-03-17	Ipatinga, Minas Gerais
	<b>1107338090975236097</b>	que maravilha ein cemig sem estar chovendo to sem energia	2019-03-17	0
	<b>1107380419018797064</b>	Eu tô pensando é como vamos trabalhar na segunda pós sunsetville mas tudo bem Cemig vai ter que me dar energia	2019-03-17	Belo Horizonte, Minas Gerais
	<b>1107417158668247040</b>	Cemig eu te odeio pq se faltar energia vou surtar nem jantei ainda odeio comer no escuro	2019-03-17	+035
	<b>1107420107767275520</b>	RT helenacarv1ho Cemig eu te odeio pq se faltar energia vou surtar nem jantei ainda odeio comer no escuro	2019-03-17	Na puta que pariu
	<b>1107482804529565701</b>	Caralho o motivo de ter acabado a energia é q pego fogo no negócio da Cemig slc bicho q medo desses troço	2019-03-18	Itajubá, Brasil
	<b>1107599066307969025</b>	cemigenergia RT cemigenergia Efficientia A empresa de serviços de energia da Cemig vai investir até R 5	2019-03-18	Franca
	<b>1107703467446099969</b>	Cemig desgraçada Mais de 6 horas sem energia E sem aviso prévio	2019-03-18	Belo Horizonte/ Sete Lagoas MG

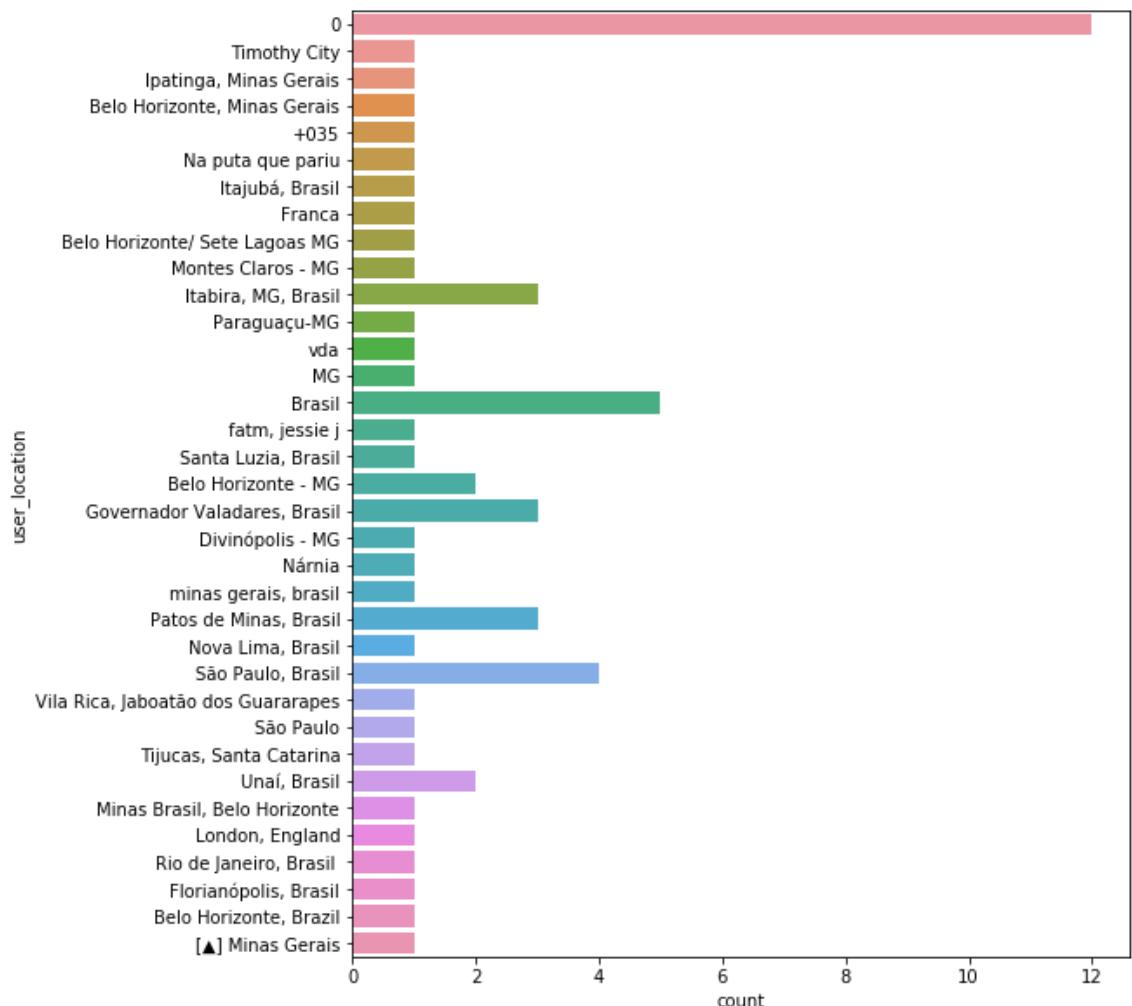
In [25]:

```
draw_graph_count(df_energia, "data")
```



In [26]:

```
plt.figure(figsize=(8,10))
ax = sns.countplot(y="user_location", data=df_energia)
plt.show()
```



In [27]:

```
df_relevante = df_tweets_tokens_pos[['text','data']].loc[df_tweets_tokens_pos['tokens'] == 'relevante']
df_relevante.head(5)
```

Out[27]:

		text	data
	id		
1109062517735731200	CMIG4 CEMIG CMIGN1 Fato Relevante 210319	2019-03-22	
1109062519933554688	CMIG4 CEMIG CMIGN1 Fato Relevante 210319	2019-03-22	
1109062601764495361	CMIG4 CEMIG CMIGN1 Fato Relevante 210319	2019-03-22	
1109062604159356928	CMIG4 CEMIG CMIGN1 Fato Relevante 210319	2019-03-22	
1109062646630887427	CMIG4 CEMIG CMIGN1 Fato Relevante 210319	2019-03-22	

## Usuários mais influentes

### Usuários mais retweetados no período

In [28]:

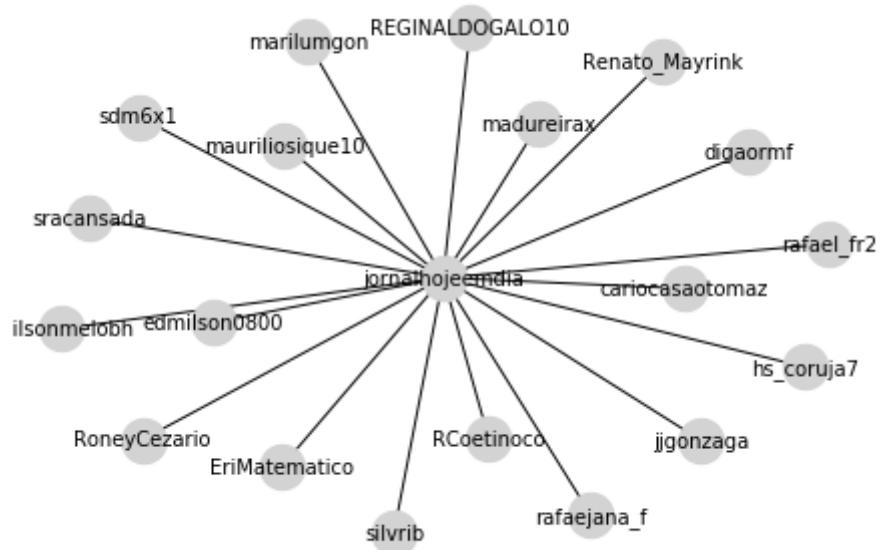
```
df_retweets = df_tweets.loc[~(df_tweets['user_retweeted'] == "")]
df_retweets.groupby('user_retweeted').user_retweeted.count().nlargest(10)
```

Out[28]:

```
user_retweeted
jornalhojeemdia    19
oalexsimoes         9
cemig_energia        6
em_com               6
AnaliseEnergia       5
GeraldodeMorais      3
gustavonolascoB     3
lopaugomes          3
lxcerdz              3
AssisMarinhodf       2
Name: user_retweeted, dtype: int64
```

In [29]:

```
df_retweets_1 = df_tweets[['user_screen_name', 'user_retweeted', 'text']].loc[~(df_tweets['user_retweeted']=="jornalhojeemdia")]
draw_network(df_retweets_1, df_retweets_1.user_screen_name, 'user_screen_name',
'user_retweeted')
```



In [30]:

```
df_retweets_1
```

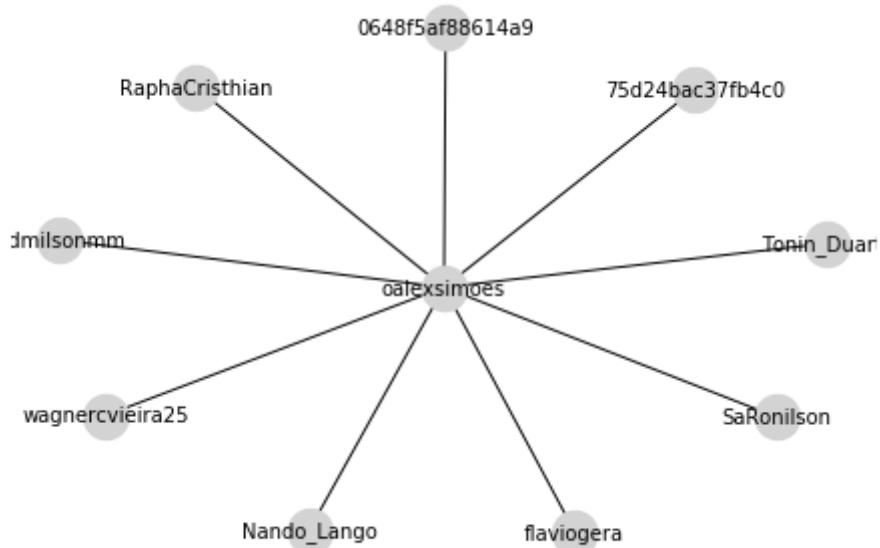
Out[30] :

	<b>user_screen_name</b>	<b>user_retweeted</b>	<b>text</b>
<b>465</b>	REGINALDOGALO10	jornalhojeemdia	RT jornalhojeemdia Cemig pagou R 990 mil por 3 anos de camarote no Independência no Mineirão preço seria R 390 mil
<b>467</b>	digaormf	jornalhojeemdia	RT jornalhojeemdia Cemig pagou R 990 mil por 3 anos de camarote no Independência no Mineirão preço seria R 390 mil
<b>468</b>	rafael_fr2	jornalhojeemdia	RT jornalhojeemdia Cemig pagou R 990 mil por 3 anos de camarote no Independência no Mineirão preço seria R 390 mil
<b>469</b>	sracansada	jornalhojeemdia	RT jornalhojeemdia Cemig pagou R 990 mil por 3 anos de camarote no Independência no Mineirão preço seria R 390 mil
<b>471</b>	hs_coruja7	jornalhojeemdia	RT jornalhojeemdia Cemig pagou R 990 mil por 3 anos de camarote no Independência no Mineirão preço seria R 390 mil
<b>473</b>	silvrib	jornalhojeemdia	RT jornalhojeemdia Cemig pagou R 990 mil por 3 anos de camarote no Independência no Mineirão preço seria R 390 mil
<b>477</b>	RCoetinoco	jornalhojeemdia	RT jornalhojeemdia Cemig pagou R 990 mil por 3 anos de camarote no Independência no Mineirão preço seria R 390 mil
<b>480</b>	Renato_Mayrink	jornalhojeemdia	RT jornalhojeemdia Cemig pagou R 990 mil por 3 anos de camarote no Independência no Mineirão preço seria R 390 mil
<b>482</b>	edmilson0800	jornalhojeemdia	RT jornalhojeemdia Cemig pagou R 990 mil por 3 anos de camarote no Independência no Mineirão preço seria R 390 mil
<b>511</b>	RoneyCezario	jornalhojeemdia	RT jornalhojeemdia Cemig pagou R 990 mil por 3 anos de camarote no Independência no Mineirão preço seria R 390 mil
<b>524</b>	EriMatematico	jornalhojeemdia	RT jornalhojeemdia Cemig pagou R 990 mil por 3 anos de camarote no Independência no Mineirão preço seria R 390 mil
<b>540</b>	mauriliosique10	jornalhojeemdia	RT jornalhojeemdia Pelo aluguel por três anos de um camarote do Estádio Independência no Horto na região Leste de Belo Horizonte a C
<b>543</b>	sdm6x1	jornalhojeemdia	RT jornalhojeemdia Pelo aluguel por três anos de um camarote do Estádio Independência no Horto na região Leste de Belo Horizonte a C

	<b>user_screen_name</b>	<b>user_retweeted</b>	<b>text</b>
<b>545</b>	rafaejana_f	jornalhojeemdia	RT jornalhojeemdia Pelo aluguel por três anos de um camarote do Estádio Independência no Horto na região Leste de Belo Horizonte a C
<b>549</b>	madureirax	jornalhojeemdia	RT jornalhojeemdia Pelo aluguel por três anos de um camarote do Estádio Independência no Horto na região Leste de Belo Horizonte a C
<b>551</b>	cariocasaotomaz	jornalhojeemdia	RT jornalhojeemdia Pelo aluguel por três anos de um camarote do Estádio Independência no Horto na região Leste de Belo Horizonte a C
<b>556</b>	adilsonmelobh	jornalhojeemdia	RT jornalhojeemdia Pelo aluguel por três anos de um camarote do Estádio Independência no Horto na região Leste de Belo Horizonte a C
<b>557</b>	jjgonzaga	jornalhojeemdia	RT jornalhojeemdia Pelo aluguel por três anos de um camarote do Estádio Independência no Horto na região Leste de Belo Horizonte a C
<b>564</b>	marilumgon	jornalhojeemdia	RT jornalhojeemdia Pelo aluguel por três anos de um camarote do Estádio Independência no Horto na região Leste de Belo Horizonte a C

In [31]:

```
df_retweets_2 = df_tweets[['user_screen_name', 'user_retweeted', 'text']].loc[~(df_tweets['user_retweeted']=="oalexsimoes")]
draw_network(df_retweets_2, df_retweets_2.user_screen_name, 'user_screen_name',
'user_retweeted')
```



In [32]:

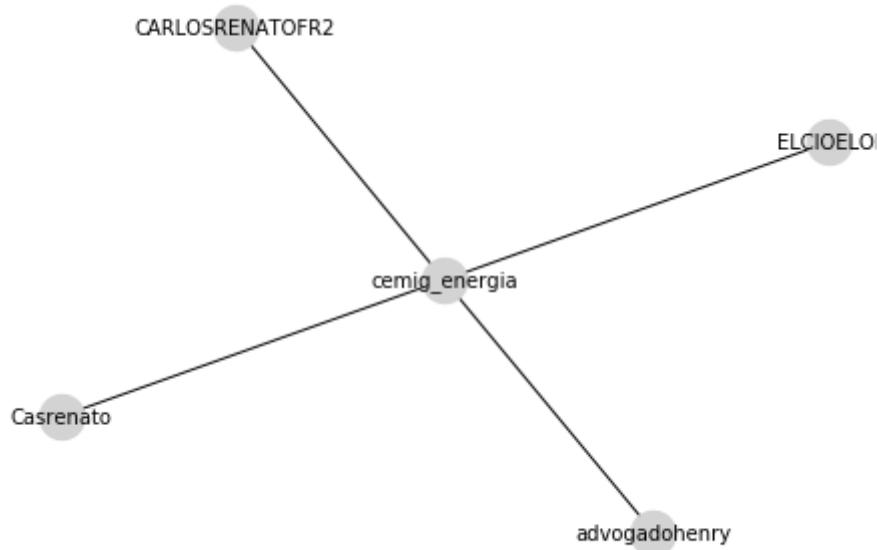
```
df_retweets_2
```

Out[32]:

	<b>user_screen_name</b>	<b>user_retweeted</b>	<b>text</b>
<b>458</b>	Tonin_Duarte	oalexsimoes	RT oalexsimoes O cancelamento há apenas uma semana não apaga a necessidade de explicações Muito estranho Pelo menos o atual govern
<b>459</b>	RaphaCristhian	oalexsimoes	RT oalexsimoes O cancelamento há apenas uma semana não apaga a necessidade de explicações Muito estranho Pelo menos o atual govern
<b>461</b>	Nando_Lango	oalexsimoes	RT oalexsimoes O cancelamento há apenas uma semana não apaga a necessidade de explicações Muito estranho Pelo menos o atual govern
<b>463</b>	0648f5af88614a9	oalexsimoes	RT oalexsimoes O cancelamento há apenas uma semana não apaga a necessidade de explicações Muito estranho Pelo menos o atual govern
<b>483</b>	Edmilsonmm	oalexsimoes	RT oalexsimoes O cancelamento há apenas uma semana não apaga a necessidade de explicações Muito estranho Pelo menos o atual govern
<b>496</b>	wagnercvieira25	oalexsimoes	RT oalexsimoes O cancelamento há apenas uma semana não apaga a necessidade de explicações Muito estranho Pelo menos o atual govern
<b>519</b>	SaRonilson	oalexsimoes	RT oalexsimoes O cancelamento há apenas uma semana não apaga a necessidade de explicações Muito estranho Pelo menos o atual govern
<b>532</b>	flaviogera	oalexsimoes	RT oalexsimoes O cancelamento há apenas uma semana não apaga a necessidade de explicações Muito estranho Pelo menos o atual govern
<b>565</b>	75d24bac37fb4c0	oalexsimoes	RT oalexsimoes O cancelamento há apenas uma semana não apaga a necessidade de explicações Muito estranho Pelo menos o atual govern

In [33]:

```
df_retweets_3 = df_tweets[['user_screen_name', 'user_retweeted', 'text']].loc[~(df_tweets['user_retweeted']=="cemig_energia")]
draw_network(df_retweets_3, df_retweets_3.user_screen_name, 'user_screen_name',
'user_retweeted')
```



In [34]:

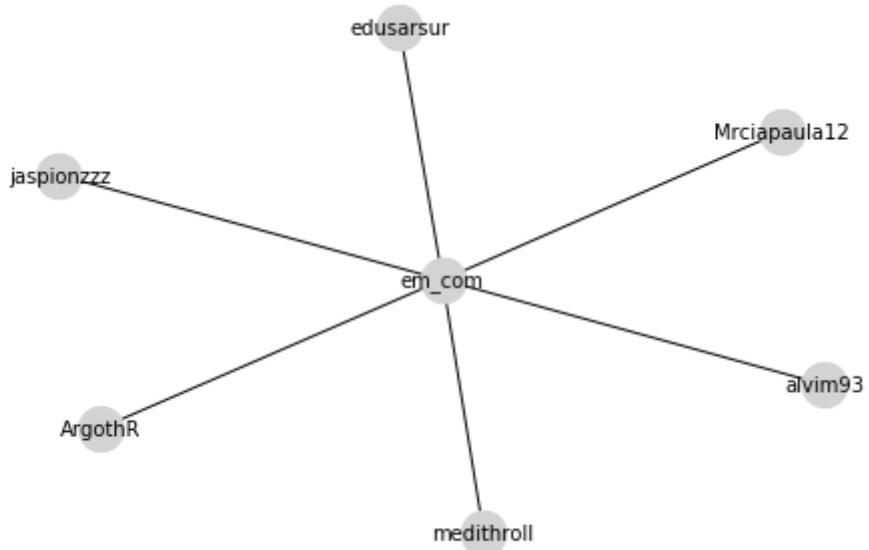
df\_retweets\_3

Out[34]:

	<b>user_screen_name</b>	<b>user_retweeted</b>	<b>text</b>
<b>69</b>	advogadohenry	cemig_energia	RT cemigenergia A Cemig Geração Distribuída Cemig GD inaugurou em Janaúba no Norte de Minas a primeira usina de minigeração destinad
<b>147</b>	Casrenato	cemig_energia	RT cemigenergia As ligações clandestinas na rede elétrica os chamados gatos são uma prática perigosa e ilegal Fique atento Chame a
<b>537</b>	Casrenato	cemig_energia	RT cemigenergia Caso veja fios elétricos caídos sobre um veículo não se aproxime se estiver dentro dele evite sairCaso seja necessá
<b>542</b>	ELCIOELOI	cemig_energia	RT cemigenergia Caso veja fios elétricos caídos sobre um veículo não se aproxime se estiver dentro dele evite sairCaso seja necessá
<b>604</b>	Casrenato	cemig_energia	RT cemigenergia Mantenha sua geladeira ligada Desligar a geladeira durante a noite reduz a durabilidade de seus alimentos além de não
<b>606</b>	CARLOSRENATOFFR2	cemig_energia	RT cemigenergia Mantenha sua geladeira ligada Desligar a geladeira durante a noite reduz a durabilidade de seus alimentos além de não

In [35]:

```
df_retweets_4 = df_tweets[['user_screen_name', 'user_retweeted', 'text']].loc[~(df_tweets['user_retweeted']=="em_com")]
draw_network(df_retweets_4, df_retweets_4.user_screen_name, 'user_screen_name',
'user_retweeted')
```



In [36]:

```
df_retweets_4
```

Out[36]:

	<b>user_screen_name</b>	<b>user_retweeted</b>	<b>text</b>
<b>178</b>	medithroll	em_com	RT emcom Operação Blecaute da Polícia Civil em parceria com a Cemig foi feita na Região Noroeste da capital Dois comerciantes foram p
<b>180</b>	edusarsur	em_com	RT emcom Operação Blecaute da Polícia Civil em parceria com a Cemig foi feita na Região Noroeste da capital Dois comerciantes foram p
<b>187</b>	alvim93	em_com	RT emcom Operação Blecaute da Polícia Civil em parceria com a Cemig foi feita na Região Noroeste da capital Dois comerciantes foram p
<b>188</b>	ArgothR	em_com	RT emcom Operação Blecaute da Polícia Civil em parceria com a Cemig foi feita na Região Noroeste da capital Dois comerciantes foram p
<b>189</b>	jaspionzzz	em_com	RT emcom Operação Blecaute da Polícia Civil em parceria com a Cemig foi feita na Região Noroeste da capital Dois comerciantes foram p
<b>197</b>	Mrciapaula12	em_com	RT emcom Operação Blecaute da Polícia Civil em parceria com a Cemig foi feita na Região Noroeste da capital Dois comerciantes foram p

## Usuários com mais seguidores

In [37]:

```
df_tweets.groupby('user_screen_name').user_followers_count.max().nlargest(10)
```

Out[37]:

```
user_screen_name
exame           2390173
em_com          497706
OficialBHTRANS 338156
otempo          276193
epocanegocios   209312
jornalhojeemdia 144154
portaluai        87909
CIObrasil        25533
cemig_energia     21620
SunoResearchcom   19466
Name: user_followers_count, dtype: int64
```

In [38]:

```
df_exame = df_tweets[['text','data']].loc[df_tweets['user_screen_name'] == 'exame']
df_exame.head(5)
```

Out[38]:

		text	data
336	Conselho da Cemig GT aprova compra de fatia na Renova		2019-03-22

In [39]:

```
df_em_com = df_tweets[['text','data']].loc[df_tweets['user_screen_name'] == 'em_com']
df_em_com.head(5)
```

Out[39]:

		text	data
177	Operação Blecaute da Polícia Civil em parceria com a Cemig foi feita na Região Noroeste da capital Dois comercia		2019-03-19

In [40]:

```
df_OficialBHTRANS = df_tweets[['text','data']].loc[df_tweets['user_screen_name'] == 'OficialBHTRANS']
df_OficialBHTRANS.head(5)
```

Out[40]:

		text	data
315	19h45 AV AMAZONAS Sentido bairro Trânsito com trechos intensos mas fluindo entre Av do Contorno Expominas e		2019-03-21

In [41]:

```
df_otempo = df_tweets[['text','data']].loc[df_tweets['user_screen_name'] == 'otempo']
df_otempo.head(5)
```

Out[41]:

		text	data
65	Marco Antônio Lage assume comunicação da Cemig mirando gestão eficiente		2019-03-18

In [42]:

```
df_epocanegocios = df_tweets[['text','data']].loc[df_tweets['user_screen_name'] == 'epocanegocios']
df_epocanegocios.head(5)
```

Out[42]:

		text	data
407	Antes dos anúncios desta semana a Renova já havia recusado uma oferta da canadense Brookfield e da própria AES Tie		2019-03-22

In [43]:

```
df_jornalhojeemdia = df_tweets[['text','data']].loc[df_tweets['user_screen_name'] == 'jornalhojeemdia']
df_jornalhojeemdia.head(5)
```

Out[43]:

		text	data
464	Cemig pagou R 990 mil por 3 anos de camarote no Independência no Mineirão preço seria R 390 mil		2019-03-22
538	Pelo aluguel por três anos de um camarote do Estádio Independência no Horto na região Leste de Belo Horizonte		2019-03-23