

**MINERAÇÃO DE TEXTO NA INTERNET:
FERRAMENTAS E SEU USO**

Jonatha Martins Cardoso

2019

SUMÁRIO

1. REQUISITOS E SOFTWARES UTILIZADOS	5
1.1. REQUISITOS RECOMENDADOS	5
1.2. RELAÇÃO DE SOFTWARES.....	5
1.3. INSTALAÇÃO E CONFIGURAÇÃO	7
1.3.1. Google Chrome	7
1.3.2. Extensões do Google Chrome	9
1.3.3. Pacote de softwares.....	10
1.3.4. Facepager	13
1.3.5. Data Miner.....	16
1.3.6. Google CSE	20
2. USO.....	27
2.1. COLETA DOS DADOS.....	27
2.1.1. Uso do Facepager.....	27
2.1.1.1. <i>Databases</i> (bases de dados).....	27
2.1.1.2. <i>Nodes</i> (nós).....	28
2.1.1.3. <i>Presets</i> (predefinições).....	29
2.1.1.4. <i>Fetch data</i> (buscar dados).....	30
2.1.1.5. <i>Export data</i> (exportar dados).....	32
2.1.2. Facebook (páginas).....	33
2.1.3. Facebook (perfis e grupos).....	34
2.1.3.1. Rolagem de página	35

2.1.3.2. Busca e exportação.....	36
2.1.4. Facebook (comentários em páginas)	38
2.1.5. Facebook (comentários em perfis e grupos)	39
2.1.6. Twitter.....	40
2.1.7. YouTube.....	41
2.1.8. Jornais, revistas, blogs e <i>sites</i> que usam RSS.....	42
2.1.9. Demais <i>sites</i> e mídias tradicionais	45
2.2. CONVERSÃO DOS ARQUIVOS.....	47
2.2.1. Arquivos CSV (Facepager).....	47
2.2.2. Arquivos XLSX	50
2.3. PRÉ-PROCESSAMENTO	52
2.3.1. Uso do OpenRefine	53
2.3.1.1. <i>Projects</i> (Projetos)	53
2.3.1.2. <i>Text facet</i> (filtro por texto).....	54
2.3.1.3. <i>Choices</i> (opções)	55
2.3.1.4. <i>Numeric facet</i> (filtro por texto)	57
2.3.1.5. <i>Timeline facet</i> (filtro por data/tempo).....	58
2.3.1.6. <i>Outliers</i>	60
2.3.1.7. <i>Edit value</i> (editar valor).....	61
2.3.1.8. <i>Join choices</i> (juntar/agrupar opções).....	61
2.3.1.9. <i>Rename choices</i> (renomear opções).....	62
2.3.1.10. <i>Close filter</i> (fechar filtro).	62
2.3.1.11. <i>Remove rows</i> (Excluir linhas)	62
2.3.1.12. <i>Remove column</i> (remover coluna)	63
2.3.1.13. <i>Export</i> (exportação)	63
2.3.2. Limpeza inicial	65

2.3.3. Facebook com Facepager	65
2.3.4. Facebook com Data Miner	67
2.3.5. Twitter com Facepager	70
2.3.6. YouTube com Facepager	70
2.3.7. RSS com Facepager	71
2.3.8. CSE com Facepager	72
2.4. EXTRAÇÃO DAS INFORMAÇÕES	75
2.4.1. SOBEK	75
2.4.2. RAnalyzer	77

1. REQUISITOS E SOFTWARES UTILIZADOS

É importante colocar que todas as ferramentas utilizadas podem ser instaladas/configuradas para Windows, Mac e distribuições Linux. Porém, devido à sua predominância, o guia é baseado nos Windows 7 e 10.

Toda a estrutura foi desenvolvida entre agosto e setembro de 2019, e caso algo não funcione, entre em contato com o autor deste documento¹, para que sejam feitas as correções necessárias.

1.1. REQUISITOS RECOMENDADOS

- Microsoft Windows, preferencialmente na versão 10, sendo aceitável, no mínimo, a versão 7.
- Computador com memória RAM de, no mínimo, 4 GB. Recomenda-se 8 GB.
- Conexão à Internet com velocidade de, no mínimo, 5 Mbps. Recomenda-se superior a 10 Mbps.
- Redes com *proxy* podem enfrentar problemas – principalmente na coleta. Entre em contato com o administrador da rede e solicite a liberação temporária para a instalação. Para o uso, será necessário liberar o software Facepager, pois o mesmo não possui essa configuração de *proxy*.
- Computadores desatualizados e/ou com *drivers* desatualizados podem apresentar dificuldades.

1.2. RELAÇÃO DE SOFTWARES

Todos os softwares listados a seguir serão utilizados no processo. Eles são gratuitos e, como dito anteriormente, compatíveis com os sistemas operacionais mencionados.

- **Google Chrome:** Importante, não apenas para acessar algumas informações necessárias para a coleta, mas também – e

¹ E-mail: ojonathacardoso@gmail.com

principalmente - por ser fundamental nas fases da mineração de texto (capítulo 2).

- **Data Miner (extensão do Chrome)**: Permite obter os dados por meio de *recipes*, que nada mais são do que configurações que permitem obter os dados de uma página da Internet, conforme o seu *layout*. No caso deste guia, o Data Miner será usado para coletar dados de perfis e grupos do Facebook.
- **Scroll It! (extensão do Chrome)**: Permite rolar uma página de forma automatizada. Em redes sociais, como o Facebook e o Instagram, para exibir mais conteúdo, é necessário rolar a página continuamente – essa extensão facilita a tarefa. Ela também auxiliará na coleta de dados de perfis e grupos do Facebook.
- **Java**: Necessário para algumas ferramentas, como o SOBEK e o OpenRefine.
- **Facepager**: *Software* que permite obter dados de diversas fontes por meio de uma tecnologia chamada *API*. Tal tecnologia permite, entre outras coisas, o acesso e a coleta de determinados dados a partir de *sites* ou outras fontes – como o Facebook e o Twitter, que possuem APIs. O *software* permite coletar, por meio de configurações pré-definidas (chamadas de *presets*), não apenas dados de mídias sociais, como também de outras fontes que usem/possuam API.
- **OpenRefine**: *Software* que permite efetuar a limpeza e preparação de dados, antes que passem por algum processo de extração de informação e conhecimento. Inclusive, possui ferramentas que já permitem a visualização de algumas informações a partir dos dados inseridos.
- **SOBEK**: *Software* para extrair informação e conhecimento a partir de texto. Ele exibe um grafo de palavras, mostrando a relação entre elas, bem como a sua frequência de aparição no texto. Existe uma versão *on-line*, mas algumas funcionalidades disponíveis nela são ausentes.
- **RAnalyzer**: Programa baseado na linguagem de programação R, que também permite extrair informação e conhecimento a partir de texto, gerando informações como gráficos, tabelas, entre outros.

- **R:** A linguagem R é uma poderosa ferramenta para, entre outras coisas, a mineração de texto, possuindo várias funcionalidades. Seu poder é aperfeiçoado com a instalação de pacotes adicionais.

1.3. INSTALAÇÃO E CONFIGURAÇÃO

Basicamente, o processo envolve dois passos:

1. instalação e configuração do Google Chrome e algumas extensões necessárias;
2. instalação do pacote de softwares necessários para o processo de mineração de texto.

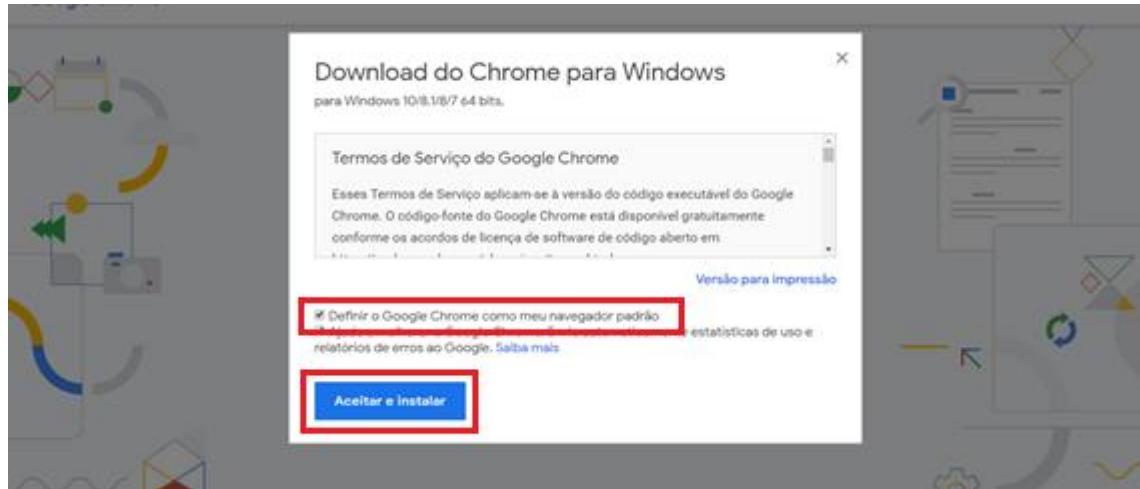
É importante salientar a necessidade de **permissões de administrador** para a instalação dos programas. Caso você não tenha credenciais com permissão (usuário e senha), contate o administrador do computador ou da rede. Entretanto, na maioria das vezes, em computadores pessoais, isso não será necessário.

Também recomendamos que todos os links abaixo sejam abertos usando o botão direito, para evitar que o documento feche.

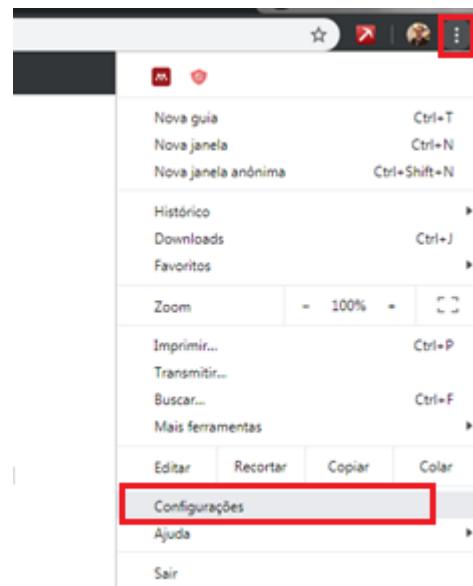
1.3.1. Google Chrome

Se ele já estiver instalado no computador, pule para o texto que está após a imagem. Caso não esteja, acesse o site google.com/intl/pt-BR/chrome e faça o download. Na tela, deixe marcada a opção de definir como navegador padrão.

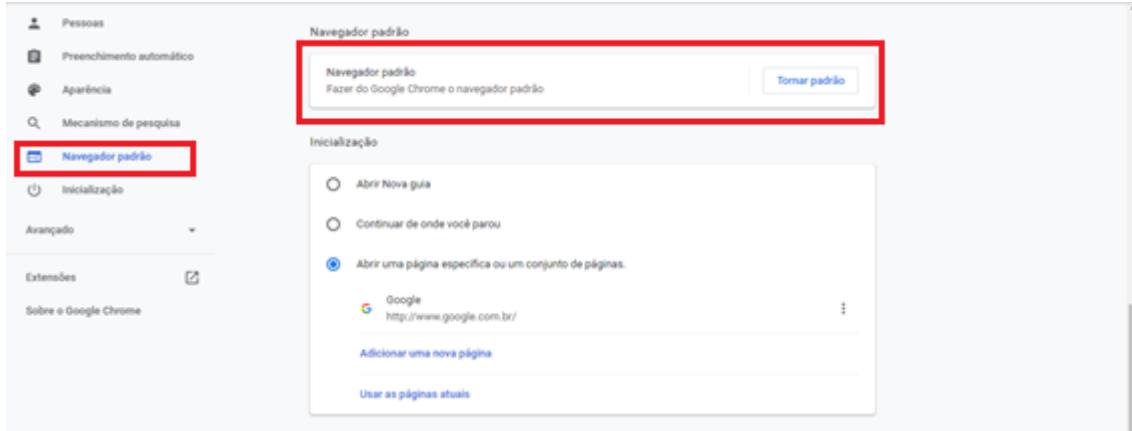
O instalador geralmente inicia automaticamente. Porém, se isso não ocorrer, é só abrir a pasta onde foi baixado – geralmente, a de Downloads – e dar dois cliques no arquivo “ChromeSetup”, para iniciar o processo de instalação.



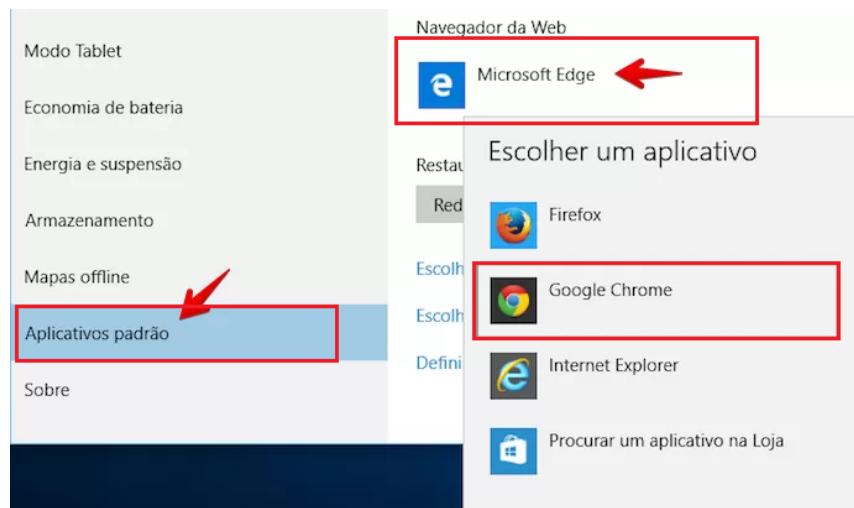
Tanto faz se você já possui o Chrome ou acabou de instalar, é necessário que o Chrome seja definido como o navegador padrão. Para isto, clique nos três pontinhos, abaixo do “X” que fecha a janela, e, então, clique em Configurações.



À esquerda, clique em “Navegador padrão”. Se ele ainda não for padrão, vai aparecer um botão escrito “Tornar padrão”. Do contrário, aparecerá a frase “O Google Chrome é seu navegador padrão”.



No Windows 10, ao clicar em “Tornar padrão”, será aberta uma janela para defini-lo como padrão. Basta clicar no grupo “Navegador” e alterar a opção para “Google Chrome”. Se ele perguntar sobre a troca, basta clicar em “Alternar mesmo assim”.



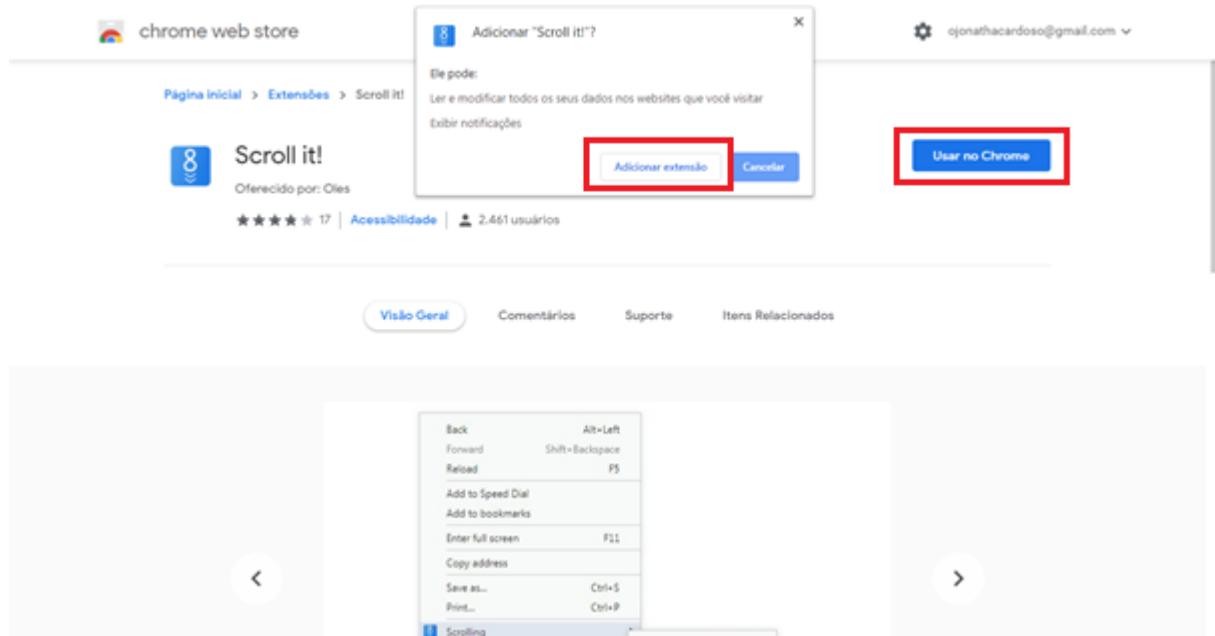
É importante salientar que a versão recomendada é a na linguagem “Português (Brasil)”. Logo, os comandos que envolvem o Chrome serão mencionados em português brasileiro.

1.3.2. Extensões do Google Chrome

Após a instalação do navegador, você pode fazer instalação a do Data Miner e do Scroll It!, acessando, no Google Chrome, os endereços abaixo:

- **Data Miner:** chrome.google.com/webstore/detail/data-scraper-easy-web-scr/nndknepjnldbdbepjfgmncbggmopgden
- **Scroll It!:** chrome.google.com/webstore/detail/scroll-it/nIndoolndemidhlomaokpfbicfnjeeed

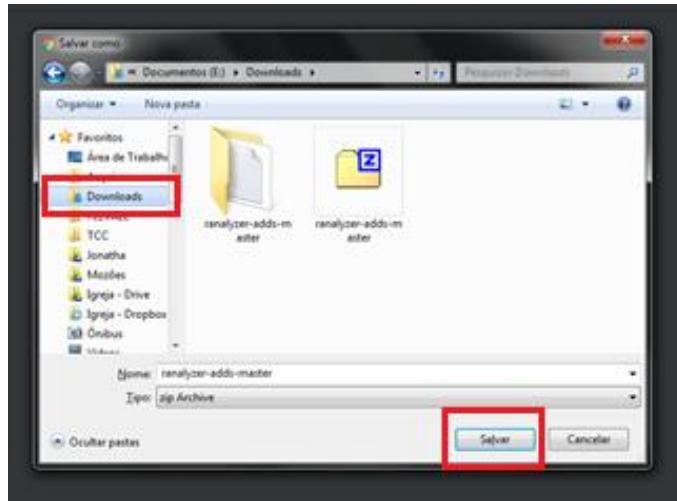
Após acessar os sites, um de cada vez, clique em “Usar no Chrome”. Na janela que aparecer, confirme clicando em “Adicionar extensão”.



1.3.3. Pacote de softwares

Devido à grande quantidade de softwares necessários, foi criado um código que faz o *download*, instalação e configuração de quase todos eles – com exceção do Google Chrome e das duas extensões, cuja instalação já foi mostrada nos subcapítulos anteriores, e de outros dois softwares, que exigirão configuração manual.

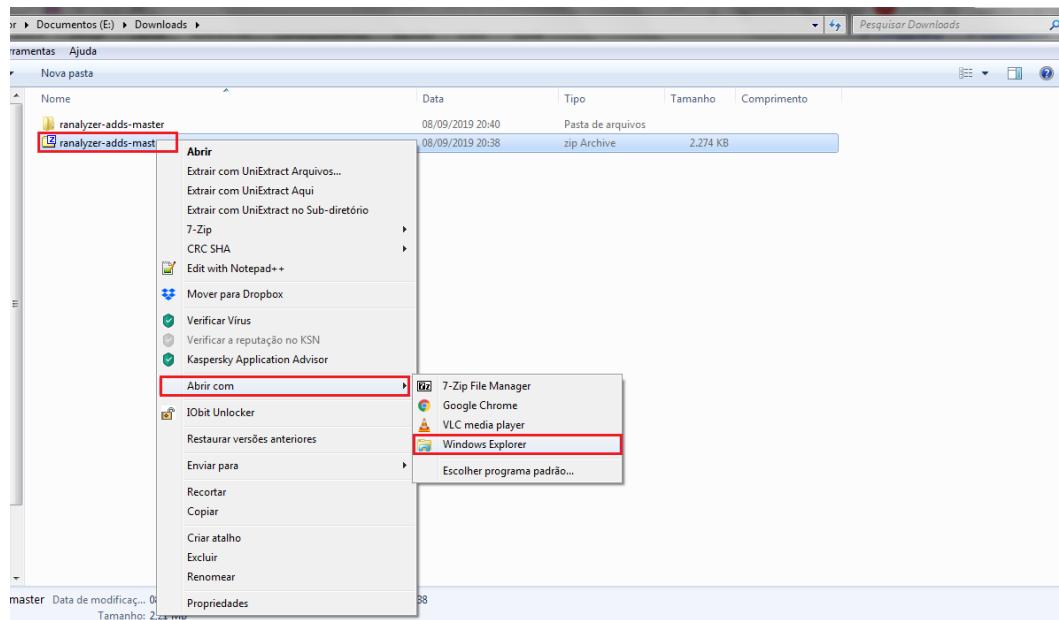
Para fazer o processo de instalação deste pacote, faça o *download* do arquivo que está no [link](https://github.com/ojonathacardoso/ralyzer-adds/archive/master.zip) github.com/ojonathacardoso/ralyzer-adds/archive/master.zip. Vai abrir uma janela do Google Chrome, fazendo o *download* automático do arquivo. Às vezes, ao invés de fazer o *download*, aparece uma janela perguntando aonde você deseja salvar o arquivo – neste caso, escolha a pasta de Downloads.



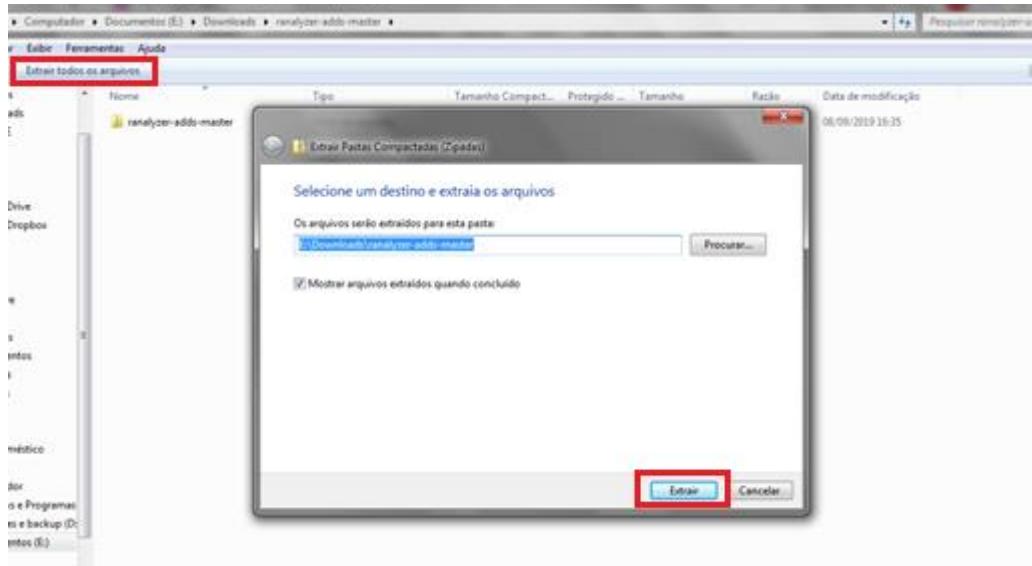
Caso após baixar o arquivo pelo Google Chrome, apareça um alerta de que o mesmo possa ser perigoso, é só clicar na seta ao lado do botão “Descartar” e clicar em “Manter”.

IMAGEM

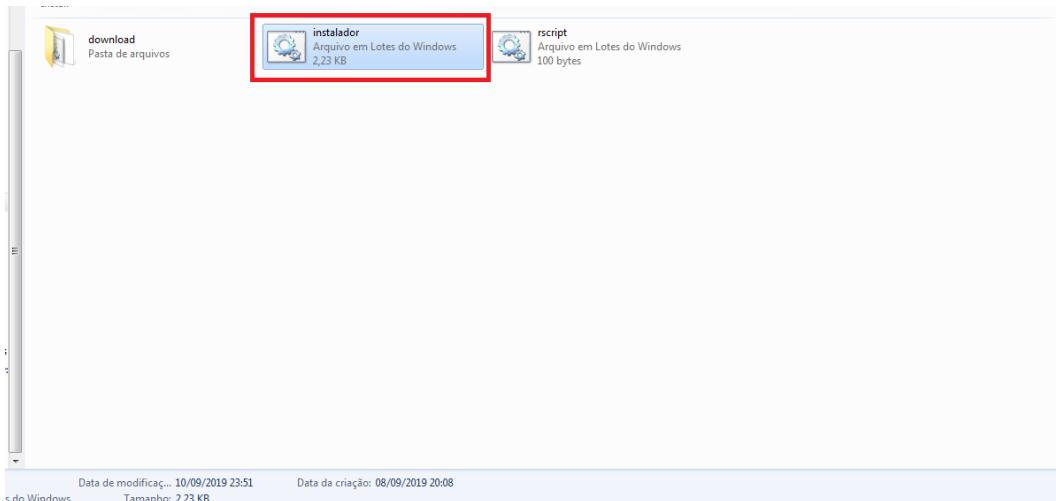
Você deve descompactar o arquivo baixado, cujo nome é “ranalyzer-adds-master.zip”. Para isto, deve-se abrir a pasta onde o mesmo foi salvo – geralmente, a de Downloads. Após isto, deve-se clicar com o botão direito em cima do mesmo, ir em “Abrir com” e clicar em “Windows Explorer”.



Abrirá uma janela, permitindo a sua extração. Então, basta clicar no botão “Extrair tudo” ou “Extrair todos os arquivos”, que aparece no alto da janela. Vai aparecer uma janela de confirmação, e é só clicar no botão “Extrair”.



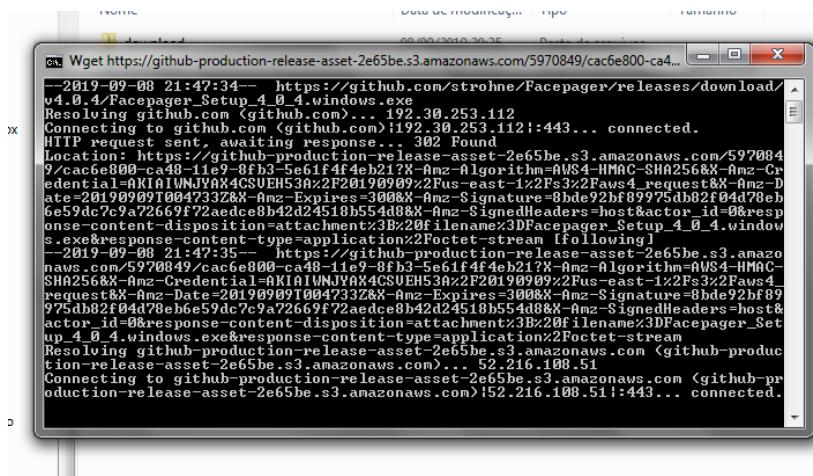
Ao final, uma nova janela aparecerá, já com os arquivos descompactados – a janela anterior, que contém o botão “Extrair tudo” ou “Extrair todos os arquivos”, deve ser fechada. Agora, após entrar na pasta “ranalyzer-adds-master”, basta entrar na subpasta “Install”, e dar dois cliques em cima do arquivo “instalador”.



Se o Windows 10 exibir uma janela informando que “protegeu o computador”, basta clicar em “Mais informações” e, então, em “Executar assim mesmo”.

IMAGEM

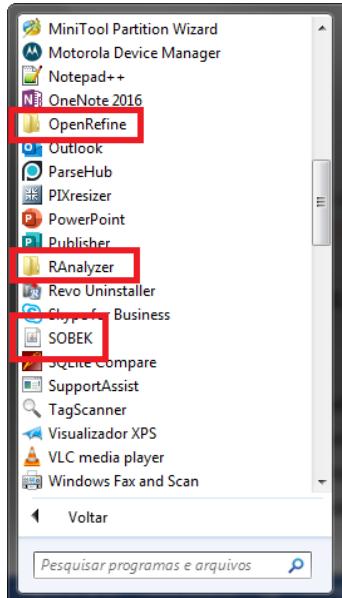
Então, abrirá uma janela do Prompt de Comando (CMD), que fará todo o procedimento: *download*, instalação e configuração. Isto poderá levar vários minutos – dependendo da Internet, de 5 a 30 minutos.



Ao longo do processo, alguns programas serão instalados, o que exigirá permissão de administração, conforme dito anteriormente. Se ele pedir alguma confirmação de autorização, basta marcar que “Sim” – se necessário, deve-se informar usuário e senha que tenha permissão de administrador.

IMAGEM

Quando a janela fechar, significa que o procedimento estará concluído. Todos os atalhos destes programas que foram instalados estarão no Menu Iniciar.



1.3.4. Facepager

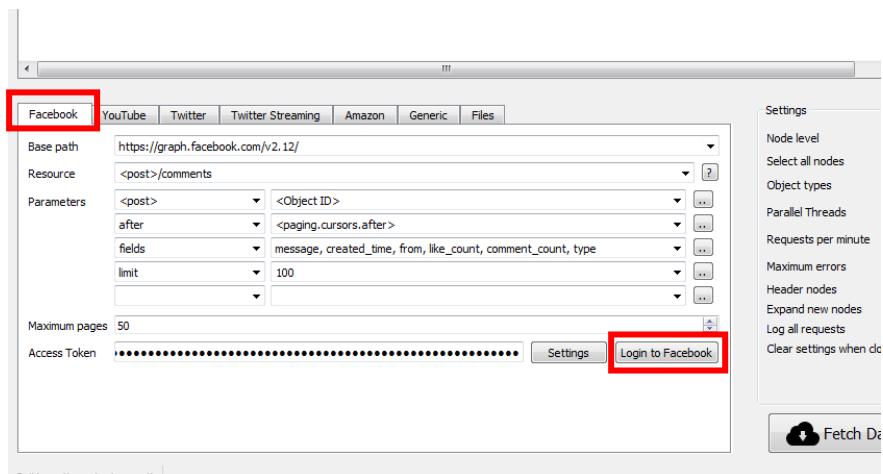
O Facepager será o responsável por obter dados, entre outros, do Facebook, Twitter e YouTube. Para isto, é necessário configurá-los. Claro, se você não possui conta nestas redes sociais, e deseja obter os dados para alguma delas, deve criar

uma conta. Por exemplo, se você quer usar só o Facebook, basta ter cadastro nesta rede.

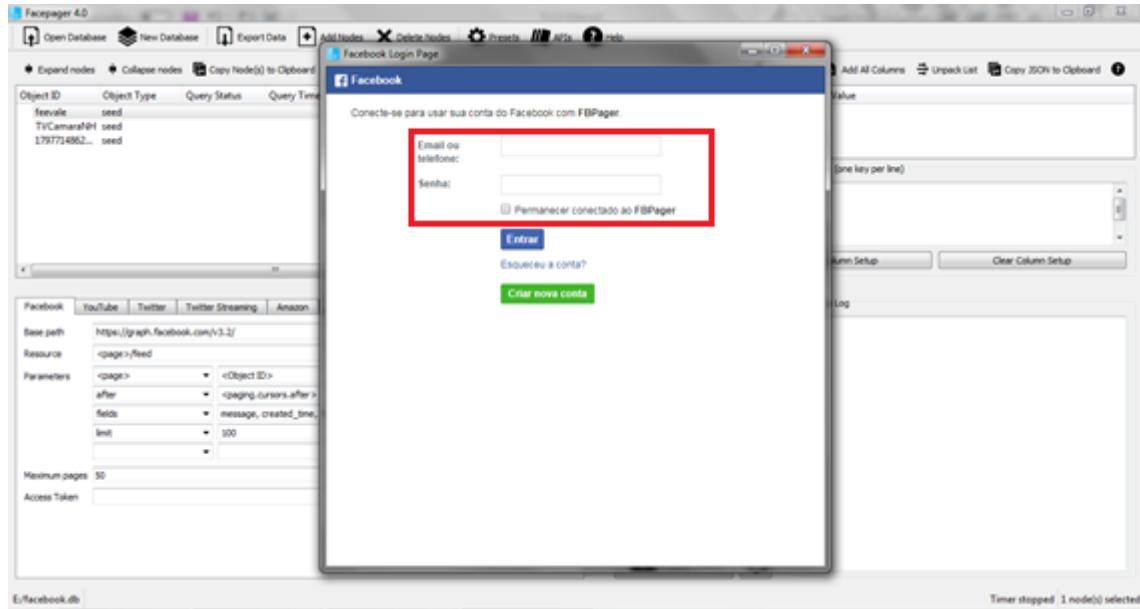
Primeiro, abra o Facepager. Vá no Menu Iniciar e procure na lista de programas o “Facepager”. Clique para abrir.



Depois de aberto, você precisa obter o código de acesso ao Facebook (*access token*). Para isto, no canto inferior esquerdo, deixe selecionada a aba “Facebook”. Então, bem na parte de baixo da janela, mais no centro, clique no botão “Login to Facebook”.



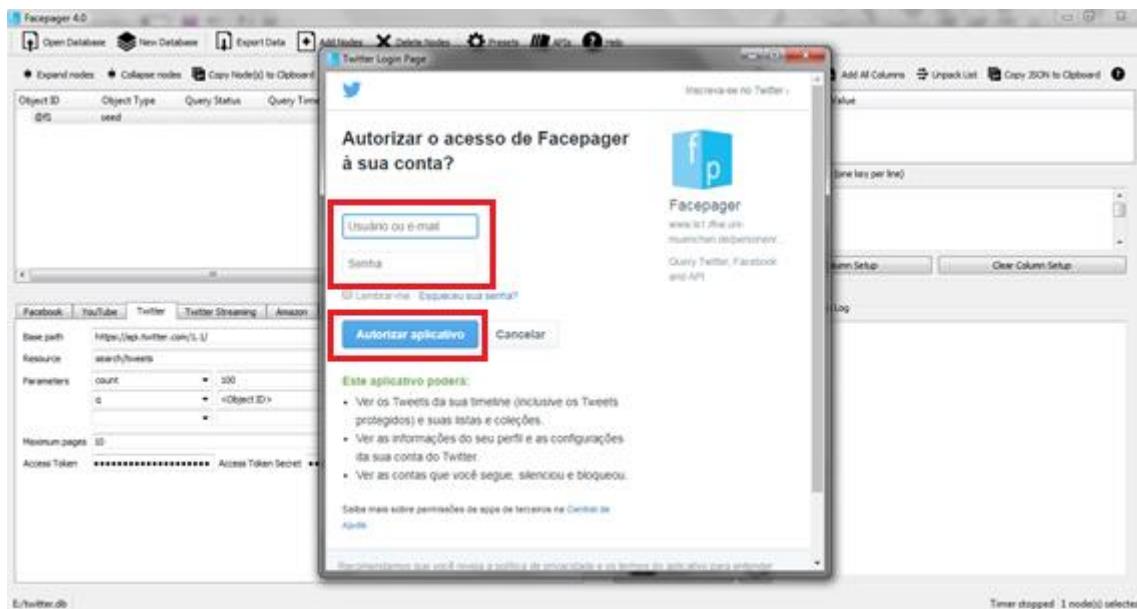
Basta acessar com o e-mail e a senha que você já usa para acessar seu Facebook, e marque a opção “Permanecer conectado ao FBPage”.



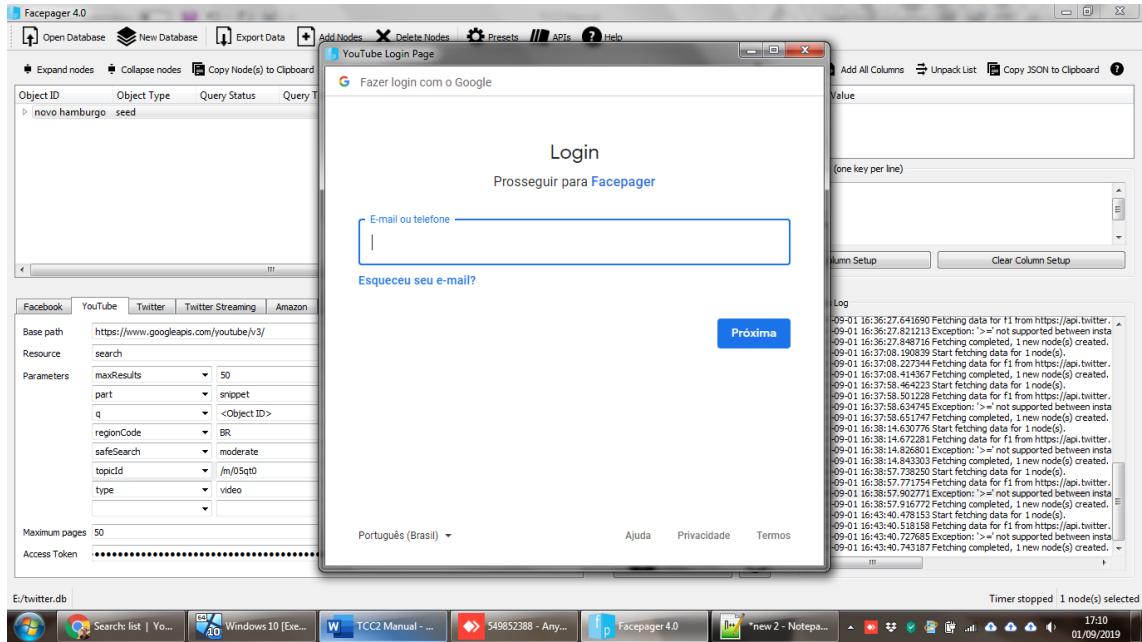
Provavelmente o Facebook vai pedir autenticação por meio de código via SMS ou via aplicativo – da mesma forma que ele iria pedir quando você abrisse o Facebook em um outro computador. Por fim, escolha a opção de “Salvar navegador”.

IMAGEM

Para o Twitter, a situação é a mesma. Clique na aba “Twitter” e, depois, no botão “Login to Twitter”. Ele também vai exigir o seu usuário ou seu e-mail, e mais a senha. Basta preencher na janela e clicar em “Autorizar aplicativo”.

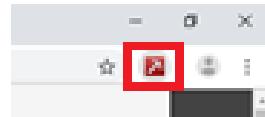


Por fim, o YouTube é igual. Clique na aba “YouTube” e no botão “Login to Google”. Preencha os dados de e-mail e, depois, senha. Se você possui uma conta do Google (para Gmail, por exemplo), pode usá-la neste *login*.

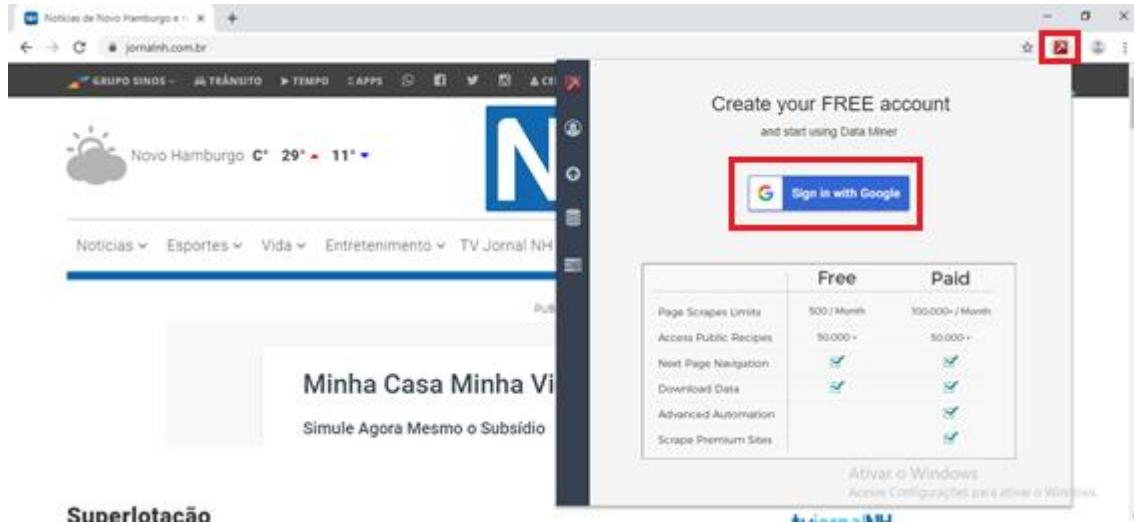


1.3.5. Data Miner

É necessário que você faça uma configuração prévia no Data Miner, para que ele possa ser usado. Inicialmente, após abrir o Google Chrome, acesse o Facebook – não é necessário ter feito o login. Depois, clique no botão vermelho, à direita da barra de endereços. Sempre que você quiser acessar o Data Miner, basta clicar ali:



Agora, você precisa fazer um cadastro com uma conta do Google, clicando em “Sign in with Google”. Como já dito antes, caso você não possua conta no Google (não usa Gmail, por exemplo), será necessário criar uma.

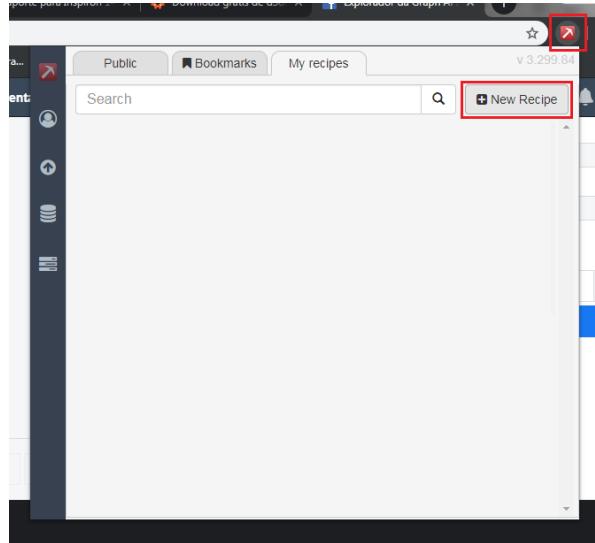


Superlotação

Depois de clicar nesse botão, vai abrir uma nova janela, com um botão igual.

Clique nele e faça o acesso – informe e-mail e senha. Então, ele estará pronto para receber as *recipes*², para que seja possível obter os dados do Facebook (perfis e grupos). Temos que criar duas.

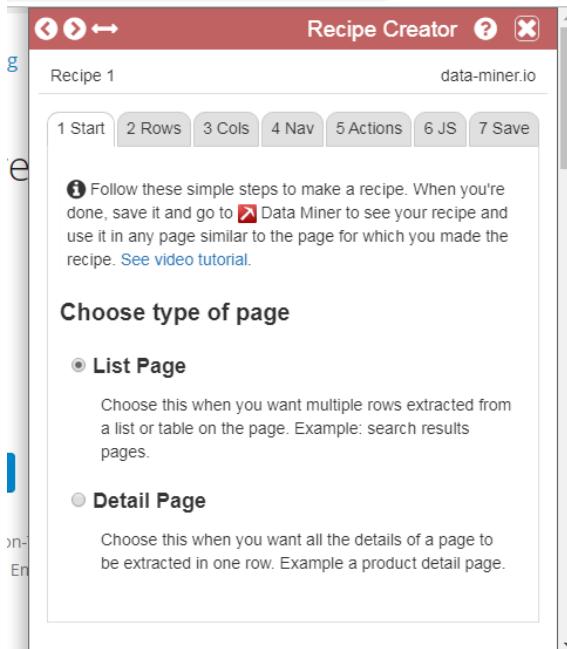
Volte para o Facebook, clique no botão vermelho novamente até aparecer a janela do Facepager. Então, clique em “New recipe”.



Agora você precisa preencher alguns campos na janela abaixo, conforme os dados que aparecem após a figura. Lembrando que são duas *recipes*, então o processo é feito duas vezes. Ah, importante lembrar que eles devem ser preenchidos conforme está destacado em amarelo, incluindo ponto, underline “ ” e

² Da mesma forma, caso o usuário possua conhecimento técnico – neste caso, em HTML e CSS – pode criar suas próprias *recipes*, de qualquer página web. Para isto, basta acessar os tutoriais disponíveis em data-miner.io

outros. Inclusive recomenda-se cuidado ao copiar os textos abaixo – após colar, confira se está igual.



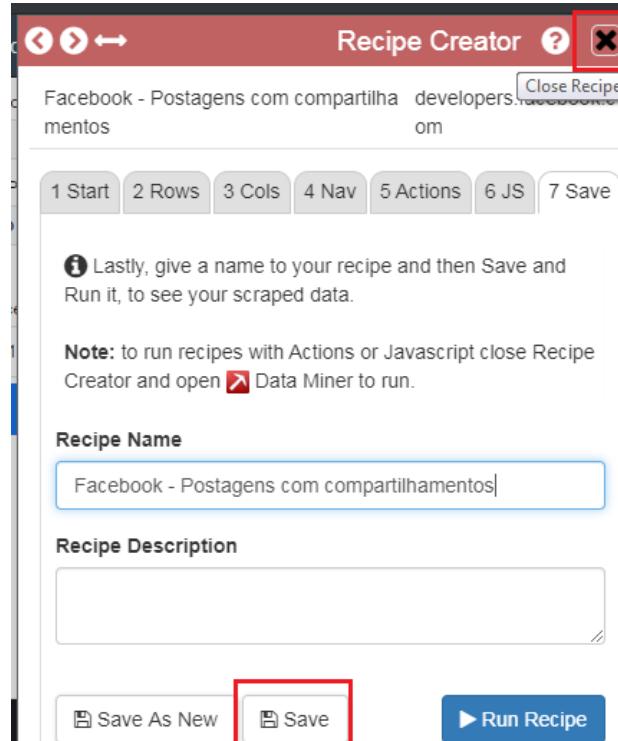
Recipe nº 1: Postagens (incluindo as compartilhadas)

- Clique na aba “**1 Start**”:
 - Campo “**Choose type of page**”: Selecionar “List page”
- Clique na aba “**2 Rows**”:
 - Campo “**Row Count**”: Preencher com **._1dwg**
 - Clique na aba “**3 Cols**” e preencha os campos “Name”, “Extract” e “Selector” com o conteúdo de cada linha da tabela abaixo. Após cada linha preenchida, clique em “Add New Column”, até todas as linhas da tabela terem sido usadas:

Campo “ Name ”	Campo “ Extract ”	Campo “ Selector ”
Autor	Text	:eq(17)
Conteúdo do autor	Text	.userContent
Autor do conteúdo compartilhado	Text	.fwb:eq(1)
Conteúdo compartilhado	Text	.mtm._5pco
Data	Text	.timestampContent
URL	URL	:eq(22)

- Clique na aba “**7 Save**”:
 - Campo “**Recipe Name**”: Preencher com “Facebook – Postagens”

- Clicar em “Save”. Após, clicar no X à direita de “Recipe Creator” e novamente no botão vermelho para retornar ao Data Miner.



- Recipe nº 2: Comentários de postagens
 - Clique na aba “1 Start”:
 - Campo “Choose type of page”: Selecionar “List page”
 - Clique na aba “2 Rows”:
 - Campo “Row”: Preencher com **_72vr**
 - Clique na aba “3 Cols” e preencha os campos “Name”, “Extract” e “Selector” com o conteúdo de cada linha da tabela abaixo. Após cada linha preenchida, clique em “Add New Column”, até todas as linhas da tabela terem sido usadas:

Name	Extract	Selector
Autor	Text	_6qw4
Conteúdo	Text	_3I3x

 - Clique na aba “7 Save”:
 - Campo “Recipe Name”: Preencher com “Facebook – Comentários”
 - Clicar em “Save”. Após, clicar no X à direita de “Recipe Creator”

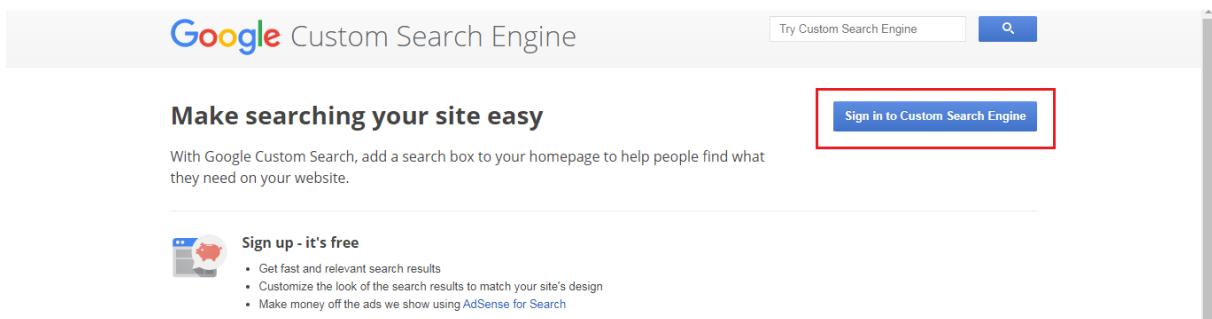
Após o procedimento, o Data Miner estará pronto para ser usados.

1.3.6. Google CSE

O Google CSE é um serviço disponibilizado pelo Google que permite a criação de motores de busca personalizados. O Google, por padrão, procura em uma lista extensa na Internet. Uma CSE criada, por sua vez, faz a busca em determinados *sites*, que foram adicionados quando da sua criação. Ele será utilizado em um dos *Presets* disponíveis no Facepager.

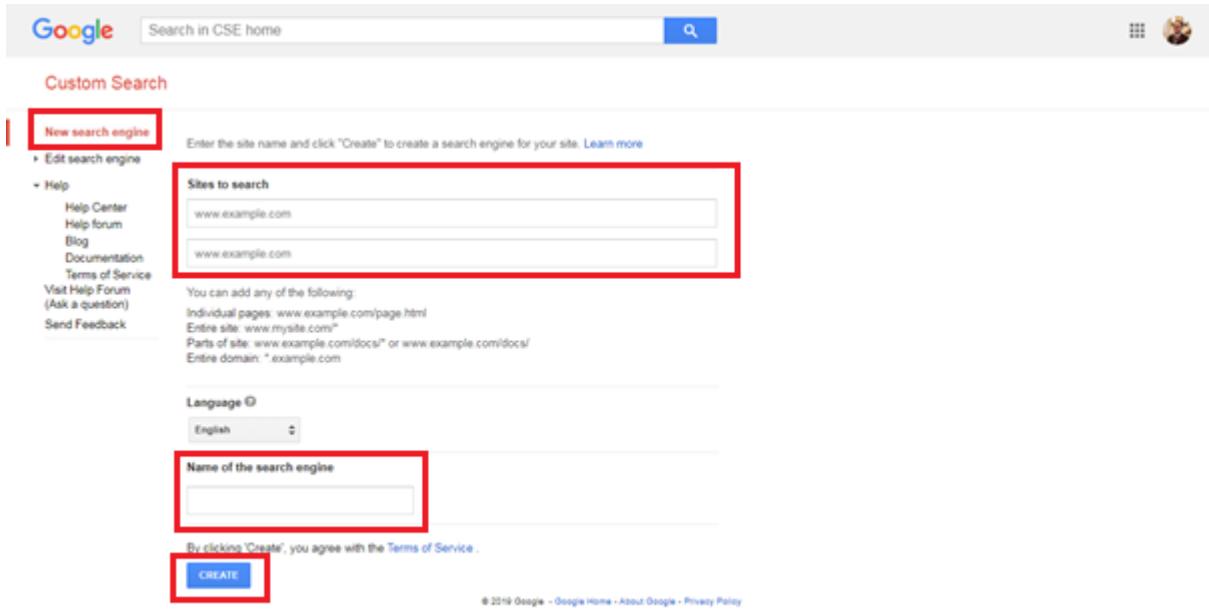
Lembrando que se você não deseja pegar dados de mídias tradicionais, como jornais e revistas, *blogs* e outros *sites*, você pode ignorar esse capítulo.

Para criar uma *CSE*, basta acessar o site cse.google.com, fazer o login com uma conta do Google, e cadastrar – pode ser necessário clicar em um botão chamado “*Sign in to Custom Search Engine*”.



Os comandos que serão mostrados podem aparecer em inglês ou português. Embora eles serão escritos em inglês, as imagens mostrarão aonde você deve ir.

Após acessar o CSE, ele exibe uma página chamada “New search engine”, para criar um motor de busca – se não exibir, é só clicar nessa opção, à esquerda da página. Em seguida, basta preencher um nome no campo “Name of the search engine” e informar um ou mais sites no campo “Sites to search”. Por fim, clique em “Create”.



Importante lembrar que, ao adicionar um site, você deve colocar apenas o endereço principal, ou seja, tirando tudo o que vem depois do “.com”, “.com.br”, “.net”, “.br” e outros. Por exemplo, se você quiser pesquisar no site www.martinbehrend.com.br/noticias/noticia/id/6732/titulo/atacama-ceu-e-chao-de-estrelas, você deve informar apenas o endereço principal, ou seja, www.martinbehrend.com.br. Isso garante que você vai procurar em todo o *site*.

Lembre-se que é desnecessário criar uma CSE para o Facebook, Twitter ou YouTube, pois os dados serão obtidos de outra forma. O CSE serve para você procurar em um determinado *site*, ou em um grupo de *sites*, que, de outra forma, não teria uma busca.

Após salvar, você pode visualizar todas as CSEs criadas, clicando em “Custom Search”. Para editar uma delas, clique em cima dela, na lista “Search engines”, ou clique na caixa logo abaixo de “Edit search engine” e escolha uma das que você criou.

The screenshot shows the 'Custom Search' interface. On the left, there's a sidebar with links like 'New search engine', 'Edit search engine' (which is selected and highlighted with a red box), 'All', 'Help', 'Visit Help Forum (Ask a question)', and 'Send Feedback'. The main area is titled 'Edit search engines' and shows a table with three rows:

Search engines	Edition	Is owner?	Public URL
<input type="checkbox"/> Jornalinh.com.br	Free	Yes	GO
<input type="checkbox"/> Novo Hamburgo	Free	Yes	GO

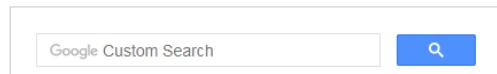
At the bottom, there's a small note: '© 2019 Google - [Google Home](#) - [About Google](#) - [Privacy Policy](#)'.

Após alguns testes, verificou-se que não é necessário fazer outras personalizações, mas alterar apenas os sites selecionados para busca, deixando os demais campos como estão.

Por fim, clique à esquerda em “Edit search engine”, depois na opção “Search features”. Depois de ir na aba “Advanced”, clique em “Results sorting”. Veja se à direita de *Result sorting* está marcado como OFF (desativado).

This screenshot shows the 'Advanced' tab selected in the 'Search features' section. The 'Results sorting' section is highlighted with a red box. To its right, a switch labeled 'OFF' is also highlighted with a red box, indicating it is disabled. Other sections like 'Promotions', 'Refinements', 'Autocomplete', and 'Synonyms' are visible at the top of the page.

Após concluir as configurações, é possível testar o motor, ao pesquisar no espaço à direita. Digite qualquer palavra e clique na lupa azul.



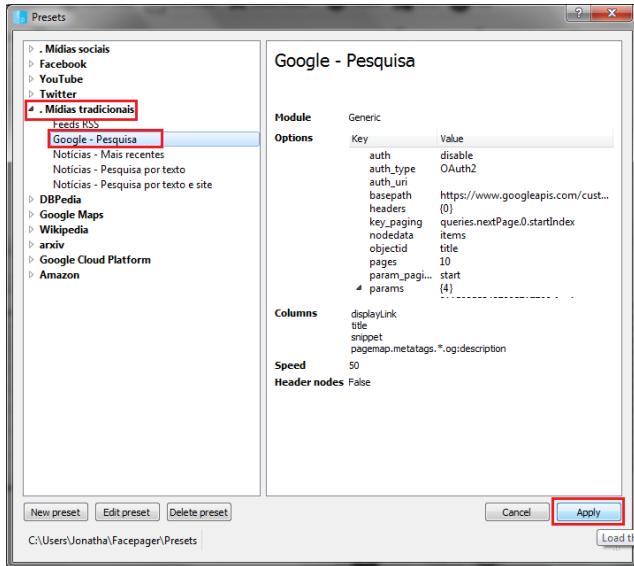
The screenshot shows the Google Custom Search interface. On the left, there's a sidebar with options like 'New search engine', 'Edit search engine' (set to 'Novo Hamburgo'), 'Setup' (which is highlighted with a red box), 'Search features', 'Statistics and Logs', 'Help', 'Visit Help Forum', 'Send Feedback', and a user profile icon. The main area has tabs for 'Basics', 'Ads', 'Admin', and 'Advanced'. Under 'Basics', there are fields for 'Search engine name' (Novo Hamburgo), 'Search engine description' (Description of search engine), 'Search engine keywords' (Search engine keywords, e.g. climate 'global warming' 'greenhouse gases'), 'Edition' (Free, with ads), 'Search engine ID' (011583553437995717708:fpv4ow-gbm), 'Public URL' (https://cse.google.com/cse?cx=011583553437995717708:fpv4ow-gbm), and dropdowns for 'Image search', 'SafeSearch', 'Region' (All Regions), and 'Language'. To the right, a search result window is open, showing results for 'Novo Hamburgo'. It includes a header with 'Aproximadamente 124.000 resultados (0.57 segundos)' and a dropdown for 'Classificar por: Date'. Below are several news snippets with titles, URLs, and small images.

Automaticamente, a pesquisa buscará por ordem de data decrescente – das mais novas para as mais antigas. Porém, caso você queira restringir

Se estiver tudo correto, clique à esquerda em Setup. Dentro da aba “Basics”, vá até “Search engine ID”, e clique no botão “Copy to clipboard”, que está logo à direita. Neste momento, o código foi copiado e está pronto para ser colado posteriormente.

This screenshot shows the same Google Custom Search interface as the previous one, but with specific elements highlighted with red boxes. The 'Setup' link in the sidebar is highlighted. In the 'Basics' tab, the 'Search engine ID' field (containing '011583553437995717708:jb81xbgjvf') and the 'Copy to clipboard' button next to it are both highlighted. The rest of the interface and the search results window on the right are identical to the first screenshot.

Agora, você deve configurar o Facepager. Vá no Menu Iniciar e abra o Facepager. Após algum tempo, a janela dele abre. Na barra superior, clique em “Presets”. Dê dois cliques no grupo “Mídias tradicionais”, clique em “Google - Pesquisa” e, após isto, clique no botão “Apply”.

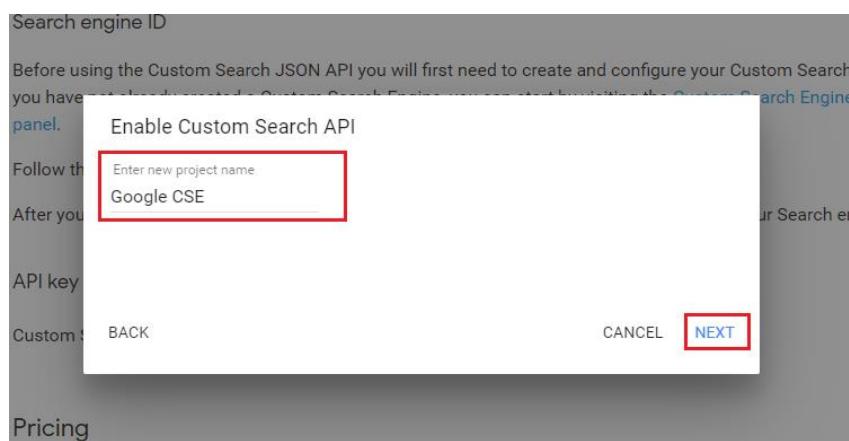


Há dois campos que devem ser preenchidos, no canto inferior esquerdo da janela, dentro de *Parameters*. O primeiro é *cx*, que deve ser substituído pelo código que você copiou antes – referente ao código do motor. Basta clicar na caixa à direita de “*cx*” e colar, pressionando Ctrl + V.

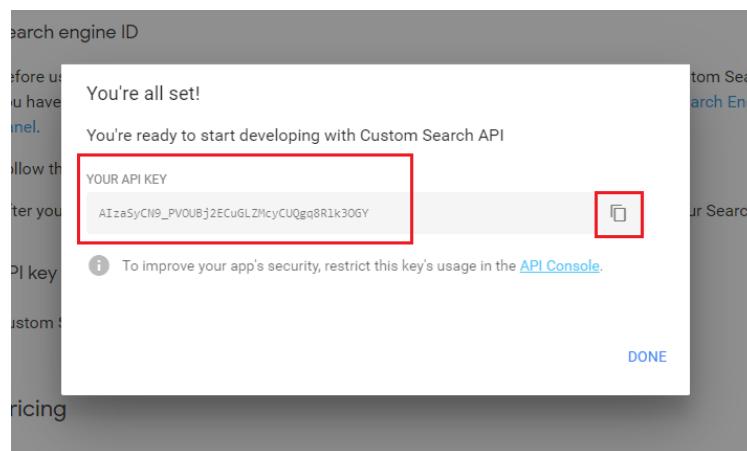
Já o segundo campo é o *key*, que se refere a uma chave de autorização fornecida pelo Google, para que você possa usar os CSEs. Para obter esta chave, acesse o site developers.google.com/custom-search/v1/overview e clique no botão azul “*GET A KEY*” – você precisa descer um pouco a página para achar o botão.

The screenshot shows the Google Custom Search API documentation. On the left, there's a sidebar with navigation links for 'STRUCTURED DATA', 'JSON API', 'Advanced Topics', and 'Custom Search Element API 1.0'. The main content area is titled 'Search engine ID' and contains text about creating a Custom Search Engine. It includes a 'GET A KEY' button, which is highlighted with a red box. To the right, there's a vertical sidebar with links to 'Contents', 'Data format', 'Related documents', 'Prerequisites', 'Search engine ID', 'API key', and 'Pricing'.

Após isto, você deve criar um projeto. Basta inserir um nome – sugere-se “Projeto” - marcar a opção “Yes” e clicar em Next.



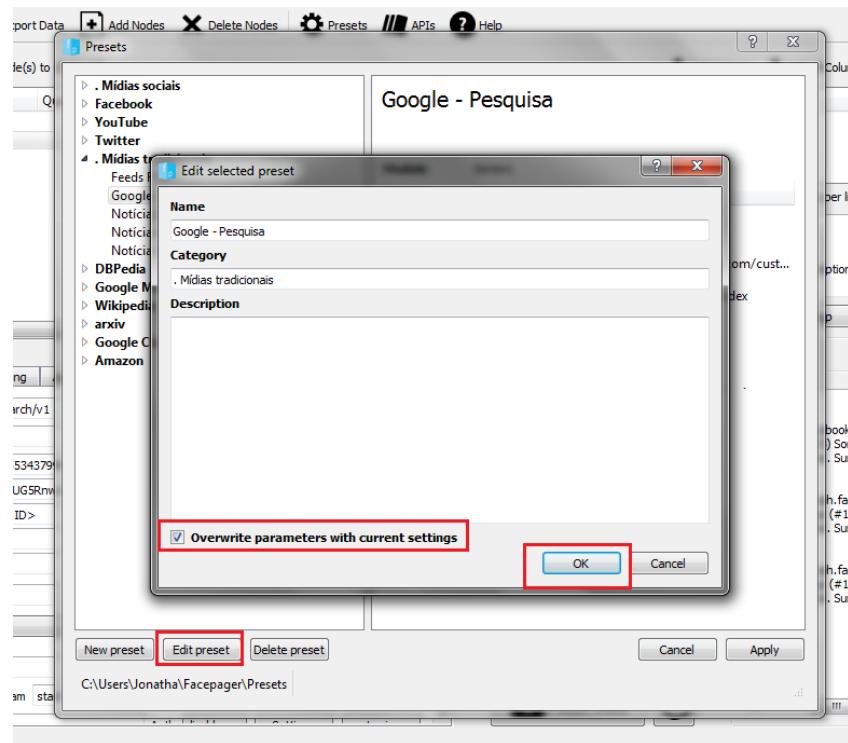
O site exibe a chave para uso, dentro de “YOUR API KEY”. Basta selecionar e copiar, clicando no botão à direita:



A chave copiada deve ser colada no Facepager, à direita de “key”.

Para salvar os dados alterados, clique novamente no botão “Presets”, vá até o grupo “Mídias tradicionais” com dois cliques, e clique na configuração “Google –

Pesquisa". Ao clicar em "Edit preset", na barra inferior da janela, marque a opção "Overwrite parameters with current settings" e clique em OK.



Vai aparecer uma janela perguntando se é para confirmar – “Overwrite preset” - e basta clicar em “Yes”. Assim, o Facepager estará pronto para usar a CSE configurada³.

³ Caso se queira usar outras CSEs, é possível criar presets no Facepager, apenas replicando os dados e salvando-os.

2. USO

Após instalados e configurados os softwares, é possível usar toda a estrutura para coletar os dados, e posteriormente fazer o pré-processamento e a extração das informações.

2.1. COLETA DOS DADOS

Aqui, será mostrado, para cada fonte de dados, como é feita a coleta, explicando o passo-a-passo de uso das ferramentas anteriormente explicadas, instaladas e configuradas.

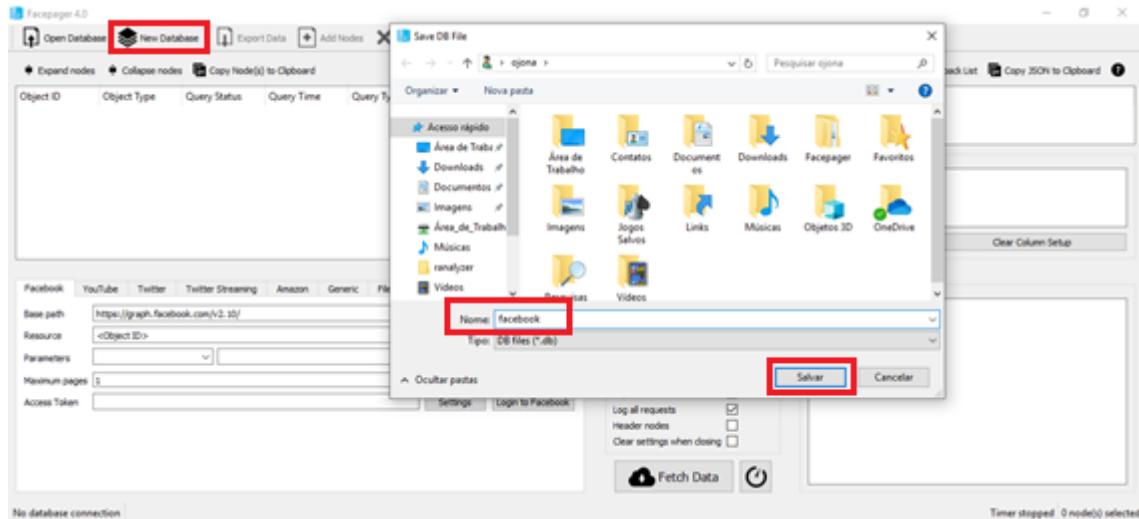
Se você vai obter postagens de perfis ou grupos do Facebook, você pode ir direto para os subcapítulos 2.1.3 (postagens) e 2.1.5 (comentários das postagens). Nos demais casos, leia o próximo subcapítulo, que explicará o uso do Facepager, e, então, siga para o subcapítulo desejado.

2.1.1. Uso do Facepager

Abra o programa, indo no Menu Iniciar e procurando-o na lista de programas.

2.1.1.1. *Databases* (bases de dados)

A primeira coisa que deve fazer, antes de iniciar cada uma das buscas, é criar uma base de dados (*database*), onde elas serão armazenadas. Basta ir à barra superior, clicar no botão “New database”, informar o nome do arquivo, aonde ele ficará salvo, e clicar em Salvar.



É possível, posteriormente, abrir uma base existente, clicando no botão “Open database”, também na barra superior. Ou seja, você pode salvar suas buscas do Facebook em uma *database*, criar outra para as do Twitter, e abri-las quando quiser.

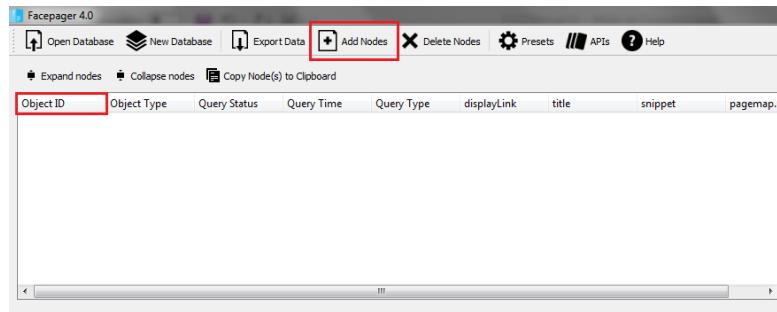
2.1.1.2. Nodes (nós)

Criada a base, você pode iniciar as buscas. Para isso, deve adicionar nós. Basicamente, um nó (*node*) é uma referência que pode ser usada para obter dados, ou pode representar os próprios dados que já foram obtidos. Por exemplo, cada página do Facebook que você adiciona para obter dados é um nó. Quando você buscar as postagens de uma página, são criados nós (nós-filhos, abaixo do nó da página) para cada postagem. Da mesma forma, destas postagens podem ser obtidos os comentários escritos, e cada comentário também vira um nó.

Veja abaixo, por exemplo, que dentro do nó “TVCamaraNH” – referente a uma página do Facebook – há nós-filhos, que representam as postagens obtidas da página. E dentro destes nós-filhos, há outros nós-filhos, que são os comentários das postagens.

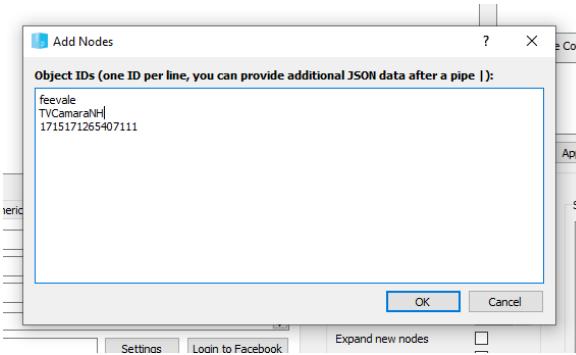
Object ID	Object Type	Query Status	Query Time	Query Type	message	from.name	created_time	shares
# 1062947...	data	fetched (200)	2019-09-01 14:3...	Facebook>pag...	NÓTA OFICIAL...	TV Câmara NH	2019-08-30T18...	
# 1062947...	data	fetched (200)	2019-09-01 13:4...	Facebook>pag...			2019-08-30T17...	2
7980...	data	fetched (200)	2019-09-02 13:4...	Facebook>pag...	Parabéns ao Gr...		2019-08-30T17...	
7980...	data	fetched (200)	2019-09-02 13:4...	Facebook>pag...	Acessibilidade é...		2019-08-30T17...	
7980...	data	fetched (200)	2019-09-02 13:4...	Facebook>pag...	Bom tarde		2019-08-30T17...	
7980...	data	fetched (200)	2019-09-02 13:4...	Facebook>pag...	Acompanhando		2019-08-30T17...	
7980...	data	fetched (200)	2019-09-02 13:4...	Facebook>pag...	Citada adequadamente		2019-08-30T17...	
11902...	official	fetched (200)	2019-09-02 13:4...	Facebook>pag...	O tema da pág...	TV Câmara NH	2019-08-30T18...	3
# 1062947...	data	fetched (200)	2019-09-01 14:3...	Facebook>pag...	O tema da pág...	TV Câmara NH	2019-08-30T18...	

Para adicionar nós, basta ir à barra superior e clicar no botão “Add nodes”.



A chave de identificação do nó está na coluna “Object ID”, que representa o identificador que foi ou será buscado. Cada página do Facebook possui um identificador. Ao buscar as postagens, cada postagem possui um identificador. Comentários, idem.

Por exemplo, ao clicar no “Add nodes”, para adicionar três nós de páginas do Facebook, colocamos os respectivos identificadores:



O Facepager precisa de nós. Só é possível obter os dados para um nó adicionado. Entretanto, nem sempre o nó será um código identificador. Por exemplo, quando você ver sobre Twitter e YouTube, verá que os nós adicionados são frases, palavras, nomes de usuário, que serão usados na pesquisa. Se eu quiser obter os tweets com a hashtag #FEEVALE, o nó que eu adicionaria no Facepager seria “#FEEVALE”.

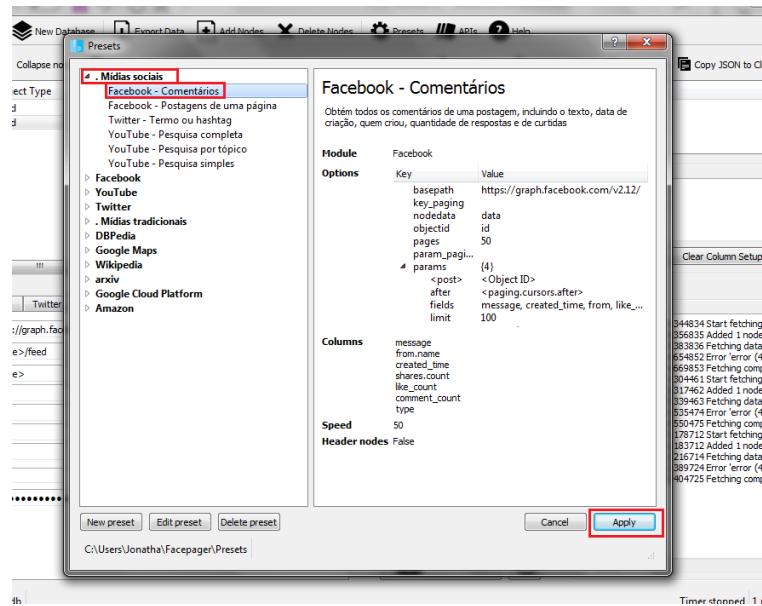
Em outros casos, como nos subcapítulos que vão falar de mídias tradicionais, o nó pode ser qualquer coisa. Por exemplo, se quiser pegar os dados de um feed RSS, eu posso criar o nó com qualquer valor, como “meuRSS”, “dadosRSS”, entre outros.

Em suma: Para cada subcapítulo, será mencionado qual o valor que deve ser adicionado no nó.

2.1.1.3. Presets (predefinições)

Como já explicado no início, ao falar sobre o Facepager, ele trabalha com presets, que são pré-configurações prontas para obter os dados de diversas fontes, como Facebook, Twitter, YouTube, RSS e Google CSE – sem contar outros que o software possui e que também podem ser criados pelo usuário.

Sempre que for solicitado para carregar um *preset*, basta ir à barra superior e clique no botão “Presets”. Na janela que abrir, selecione o *preset* desejado e clique em “Apply”, no canto inferior direito. Por exemplo, para carregar o *preset* “Facebook – Comentários”, após dar dois cliques no grupo “Mídias sociais”:

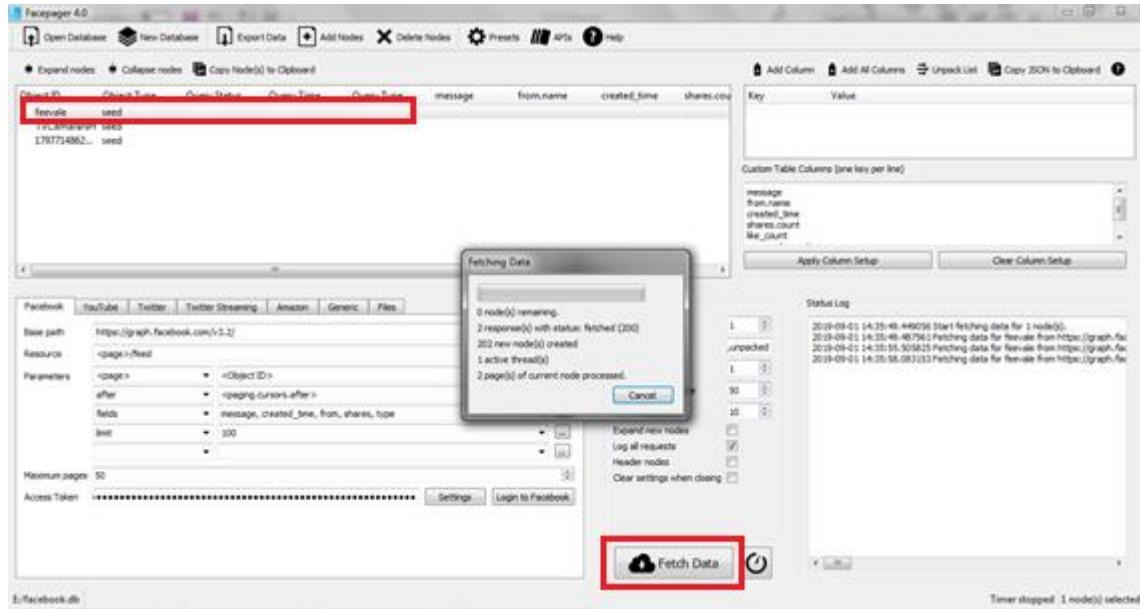


2.1.1.4. *Fetch data* (buscar dados)

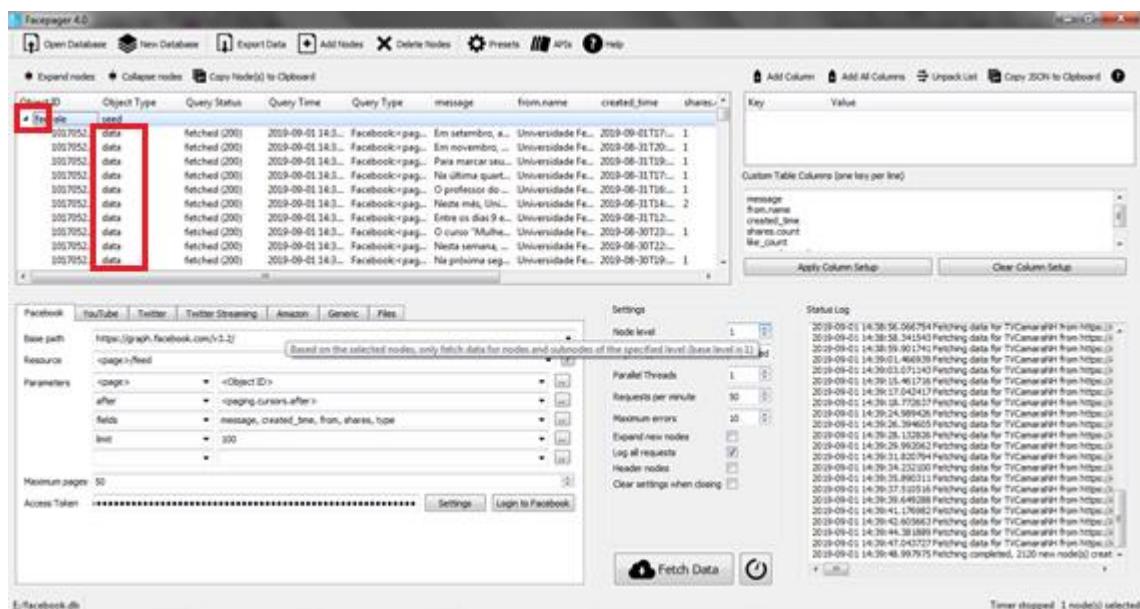
Para obter os dados, basta selecionar o nó desejado e clicar em “Fetch data”. Os nós que podem ser usados para obter dados são aqueles cujo “Object Type” é “seed” ou “data”. O primeiro são os nós que você adicionou. O segundo, os que o Facepager buscou.

Cada fonte de dados possui um limite de postagens, para evitar eventuais bloqueios. Por exemplo, o Facebook está limitado a 5 mil postagens. Não se preocupe com a demora da busca - o processo é propositalmente demorado para evitar esse bloqueio⁴.

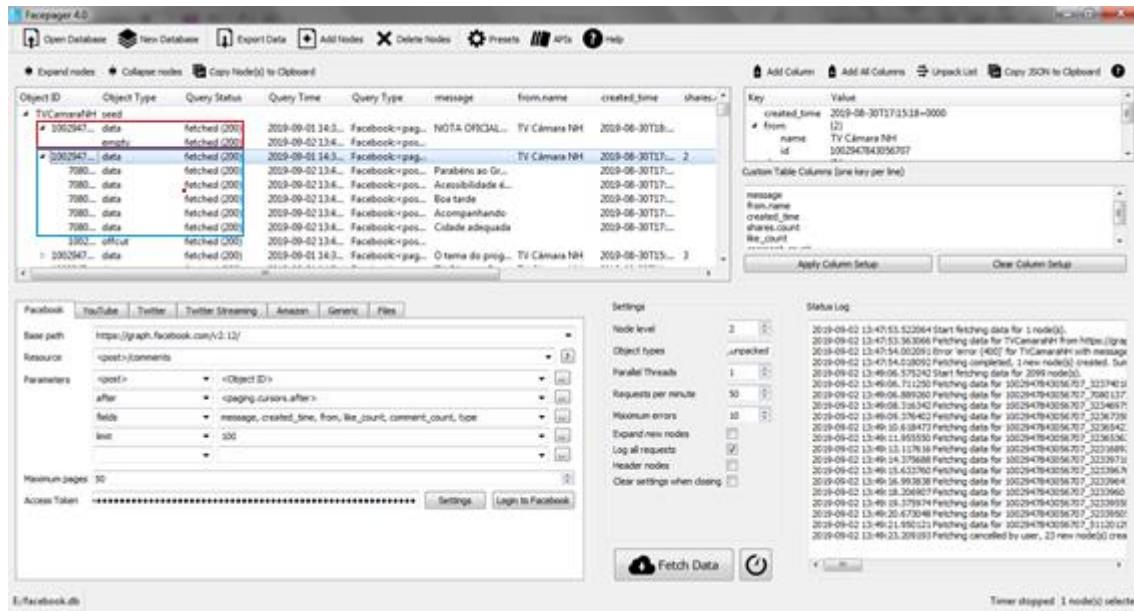
⁴ Esta e outras configurações de pesquisa podem ser ajustadas. Claro que, com conhecimento da documentação, é possível ajustar isto e outras questões, como data inicial/final, entre outros. Porém, para deixar a busca o mais simples e abrangente possível, foi estabelecido desta forma.



Para os dados obtidos clicando na seta à esquerda do “Object ID”.

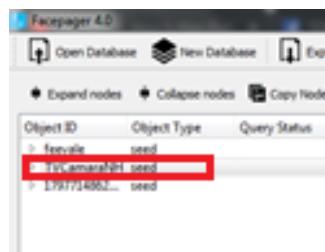


Para saber se e quantos dados foram coletados, basta clicar na seta à esquerda do “Object ID”, para expandir a busca. Caso haja apenas uma linha, cujo dado na coluna “Object Type” seja “empty” (na imagem abaixo, destacado em vermelho), significa que não há dados. Se aparecerem uma ou mais linhas, cujo valor seja “data” (destacado em azul), significa que há dados, sendo que cada linha representa um.



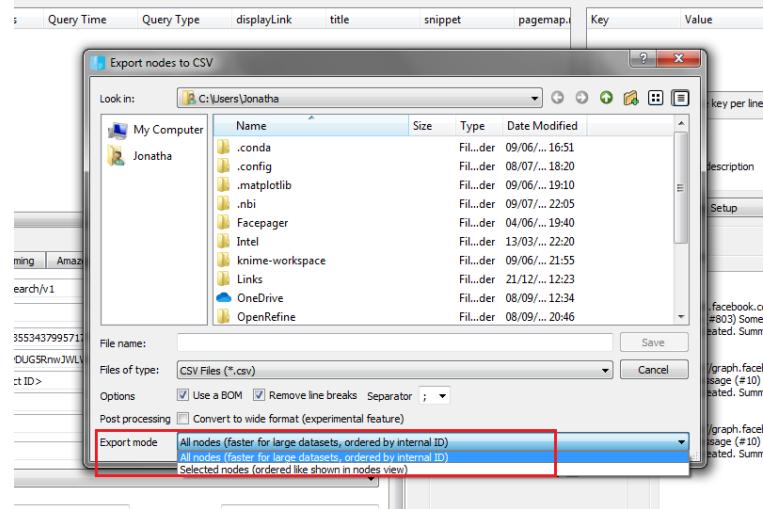
2.1.1.5. Export data (exportar dados)

Para exportar os dados, primeiro, você deve selecionar as páginas desejadas. Você pode selecionar um nó pai ou nós filhos. Por exemplo, pode selecionar um nó que você adicionou, para uma página do Facebook, e automaticamente os nós filhos – postagens obtidas – serão incluídas na exportação. Por sua vez, se você selecionar apenas um nó de uma postagem obtida, apenas ela será exportada. Por exemplo, das três páginas abaixo, serão exportados apenas os dados obtidos de uma delas.



Selecionados os nós, você pode clicar, lá na barra superior, no botão “Export Data”.

Se você deseja exportar todos os nós, selecionados ou não, marque, no campo “Export mode”, a opção “All nodes”. Porém, se deseja exportar apenas os nós que selecionou antes, marque a opção “Selected nodes”.



Escolha o nome e o local, e salve em um local desejado. Não altere as demais opções.

2.1.2. Facebook (páginas)

Etapa 1: Criar uma *database*, conforme explicado no subcapítulo 2.1.1.1.

Etapa 2: Adicionar nós, conforme explicado no subcapítulo 2.1.1.2.

Para nó, será utilizado o código identificador da página. Para obter o código identificador de uma página, abra a página no Facebook, copie o endereço e cole no site findmyfbid.com e clique em “Find numeric ID”.

Find your Facebook ID

To find your Facebook personal numeric ID for fb:admins, social plugins, and more, enter your **Facebook personal profile URL** below:

Find numeric ID →

Depois de alguns segundos, vai aparecer o identificador, que deve ser copiado.

Success!

Your Facebook personal numeric ID is:

388002091282590

[Find another →](#)

Em algumas páginas, é possível obter esse identificador sem precisar no site. A dica é que boa parte das páginas possuem um nome personalizado, como na página da FEEVALE (www.facebook.com/feevale), onde deve ser considerado o conteúdo que está após o “.com/” – neste caso, o identificador da página é “**feevale**”.



É possível usar tanto o número quanto o texto. Mas, se você tiver dúvidas em qual é o identificador da página, siga o primeiro passo – até porque algumas páginas nem possuem identificador em texto.

Etapa 3: Carregar o preset “Facebook - Postagens” (dentro do grupo “Mídias sociais”), conforme explicado no subcapítulo 2.1.1.3.

Etapa 4: Obter os dados, conforme explicado no subcapítulo 2.1.1.4. O limite de dados obtidos foi estabelecido em 5 mil.

Etapa 5: Exportar os dados, conforme explicado no subcapítulo 2.1.1.5.

2.1.3. Facebook (perfis e grupos)

Obter dados de perfis e grupos passa por um trabalho diferente – o único que não usa API e o Facepager -, pois o processo usado para páginas do Facebook,

explicado no subcapítulo anterior, foi desativado. Assim, é necessário acessar os perfis e grupos por dentro do Facebook.

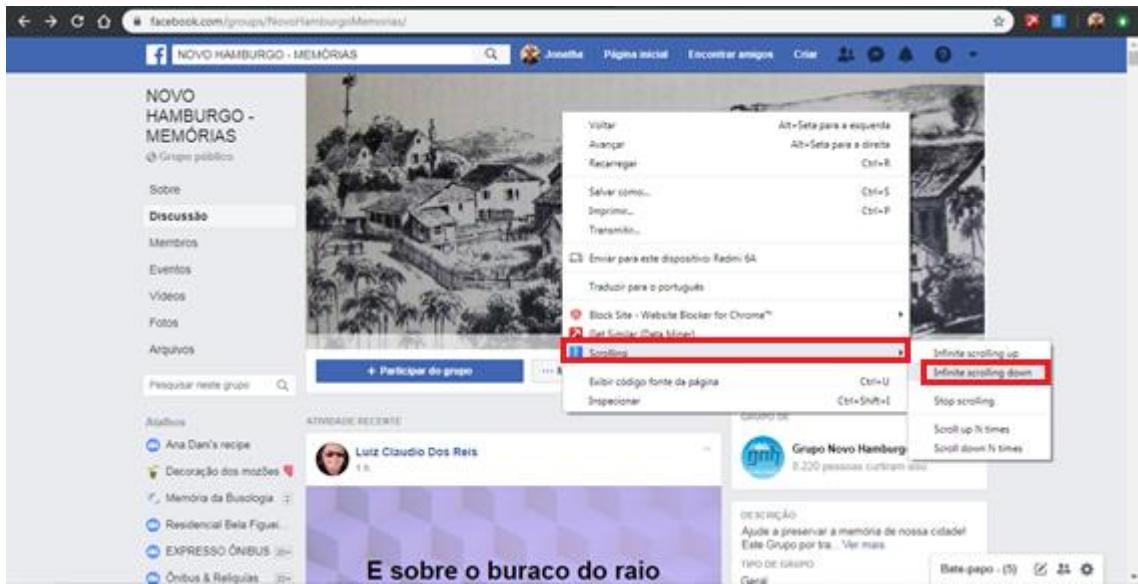
Desde já, saiba que as técnicas que mostraremos neste subcapítulo não são restritas a perfis e grupos, mas também podem ser usadas em páginas, visto que a estrutura no Facebook é a mesma.



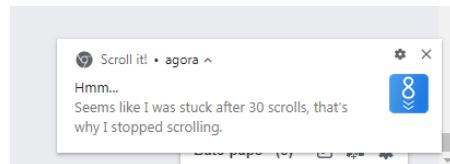
2.1.3.1. Rolagem de página

Após abrir uma página ou um grupo desejado, é necessário rolar a página. Existem duas formas: Manualmente, como geralmente é feito quando se acessa o Facebook, ou com uma ferramenta semiautomática: o Scroll It! – adicionado lá no primeiro capítulo.

Para realizar a rolagem, basta clicar com o botão direito em cima da página, ir em “Scrolling” e escolher “Infinite scrolling down”.



Automaticamente, o Facebook começa a rolar a página. Por padrão, ele rola até um certo limite – às vezes, o Chrome exibe uma mensagem de que o Scroll It! travou.

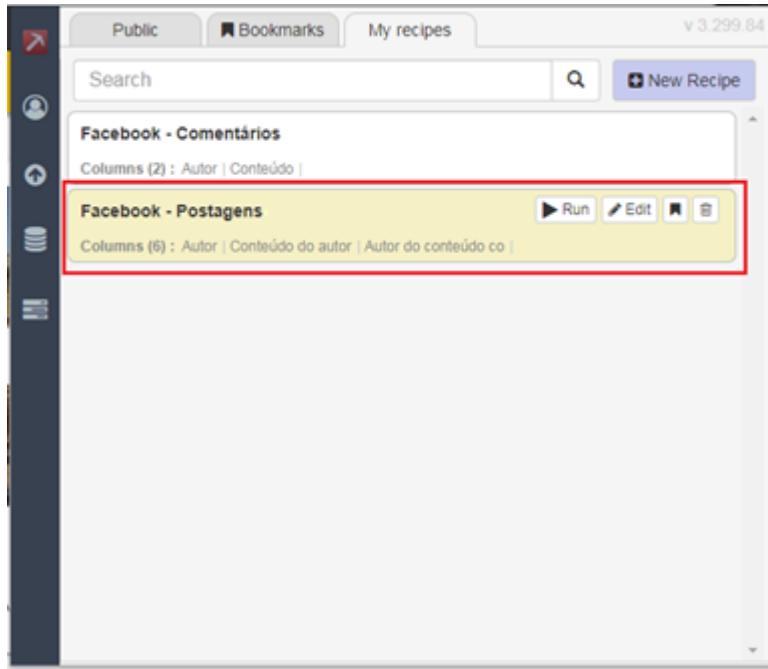


Repita o procedimento até que seja atingido um ponto ou data que for de interesse do usuário. Ao chegar ao ponto ideal, é possível obter os dados usando o Data Miner.

2.1.3.2. Busca e exportação

Basta clicar no botão vermelho da ferramenta, à direita da barra de endereços. Posicione o mouse em cima da opção “Facebook – Postagens” e clique no botão “Run”.

Importante mencionar que ela vai obter todas as postagens exibidas. Porém, as criadas pelo autor da página aparecerão na coluna “Conteúdo do autor”, e as que foram compartilhadas a partir de postagens de outras pessoas vão aparecer na coluna “Conteúdo compartilhado”.



Então, a ferramenta realiza a coleta das postagens, exibindo os dados em uma nova janela. Você pode, se quiser, fazer pesquisas nesses dados – o que já pode lhe dar algumas informações prévias, como as pessoas que postaram, alguma palavra ou tema específico, entre outros.

This screenshot shows the Data Miner interface with a search filter applied to the 'Facebook - Postagens com compartilhamentos' section. The filter is set to 'Contains' 'compartilhou'. The results table shows various posts with their authors and content. The 'Download' button is highlighted with a red box.

Autor	Conteúdo do autor	Autor do conteúdo compartilhado	Conteúdo compartilhado
Douglas Oliveira Friedl comp...	Contains compartilhou	Ónibus gaúcho	524 da vicsa-Neobus
Edival Pugsley compartilhou ...		Marcos Block para ANTIGA...	Imagens dos passageiros
Edival Pugsley compartilhou ...		Paulo José da Costa para A...	Quando a carroça puxava
Douglas Oliveira Friedl comp...		Ônibus Antigos de Brasília	TCB DF Decada de 80
Manfred Luis Huppes compa...		Fotos Antigas Rio Grande do...	Santa Maria - Av Rio Br...

Para salvá-los, basta clicar em “Download” e escolher a opção “as .XLSX”. Por fim, basta escolher o nome desejado para o arquivo e aonde ele será salvo. Lembrando que todas as linhas serão salvas, independente dos filtros feitos, como acima.

The screenshot shows the OpenRefine interface with a recipe titled "Facebook - Posts with shares". The "Extracted Data" panel displays a table with two columns: "Autor" and "Conteúdo do autor". The "Autor" column lists names like Harff Haus, Rodrigo Leo Santos, LUIZ Cláudio Dos Reis, Mônica Pilger, etc. The "Conteúdo do autor" column contains their respective posts. At the top right of the "Extracted Data" panel, there is a "Download" button with options for "as CSV" and "as XLSX", and a "To Clipboard" button. Both the "Download" and "To Clipboard" buttons are highlighted with red boxes.

2.1.4. Facebook (comentários em páginas)

Etapa 1: Criar uma *database*, conforme explicado no subcapítulo 2.1.1.1.

Etapa 2: Não é necessário adicionar nós, pois serão utilizados os nós que foram obtidos no subcapítulo 2.1.2

Você deve selecionar todas as linhas de postagens – aquelas que, na coluna “Object Type”, aparecem como “data”..

The screenshot shows the OpenRefine interface displaying a results table. The columns are: Object ID, Object Type, Query Status, Query Time, Query Type, displayLink, title, snippet, and page. The "Object Type" column contains numerous entries labeled "data", which are all highlighted with red boxes. This indicates that these rows represent the posts being analyzed.

Saiba que, neste momento, não é possível separar postagens e comentários – caso se deseje exportar apenas um ou outro. Ela poderá ser feita no OpenRefine, conforme subcapítulo 2.3.3.

Etapa 3: Carregar o preset “Facebook - Comentários” (dentro do grupo “Mídias sociais”), conforme explicado no subcapítulo 2.1.1.3.

Etapa 4: Obter os dados, conforme explicado no subcapítulo 2.1.1.4. O limite de dados obtidos foi estabelecido em 5 mil.

Etapa 5: Exportar os dados, conforme explicado no subcapítulo 2.1.1.5.

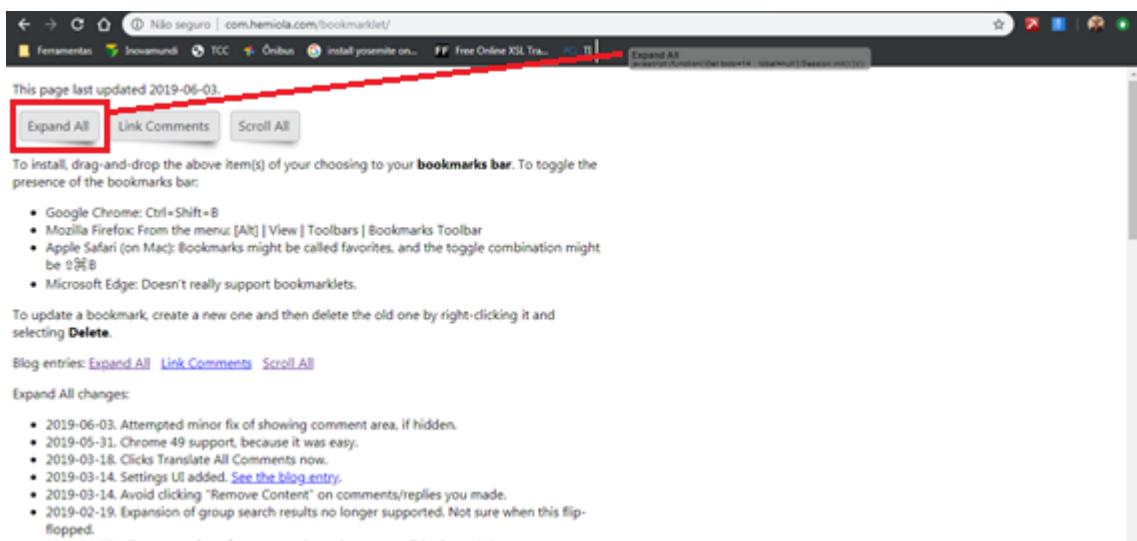
2.1.5. Facebook (comentários em perfis e grupos)

Etapa 1: Rolar a página, conforme explicado no subcapítulo 2.1.3.1.

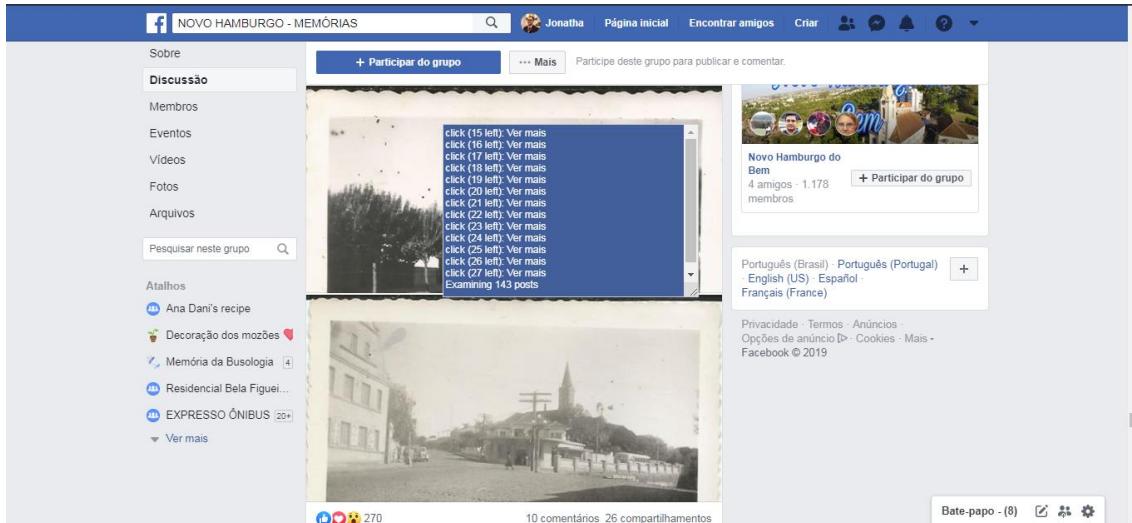
Etapa 2: Expandir os comentários.

Perceba que, por padrão, o Facebook não exibe todos os comentários. Para que todos eles sejam exibidos, logo após a rolagem, sem a necessidade de um procedimento manual, existe uma ferramenta que auxilia isto.

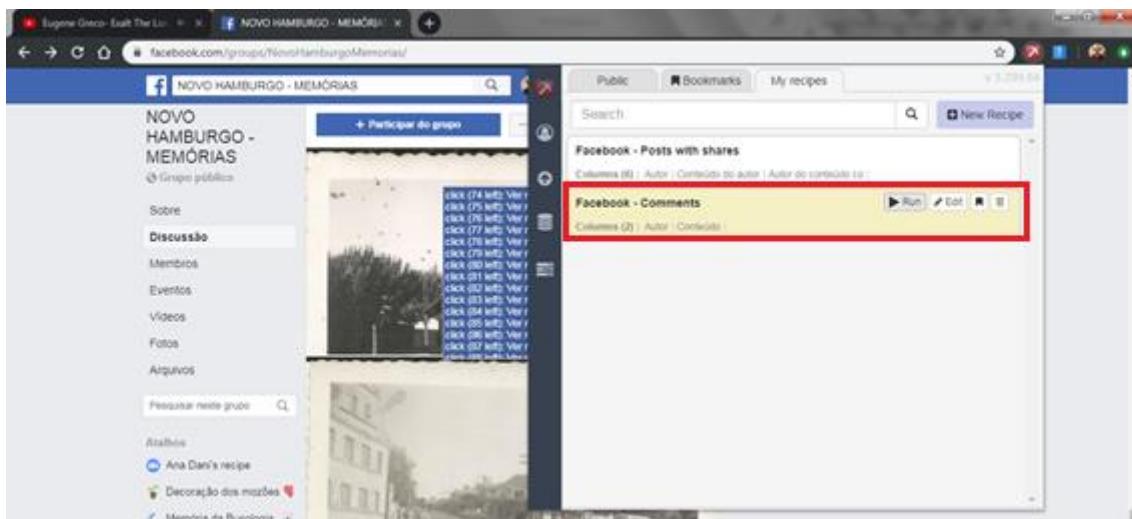
Ela chama-se “Expand all”, e permite expandir todos os comentários exibidos no Facebook. Para isto, acesse o site com.hemiola.com/bookmarklet. Então, clique com o botão direito do mouse em cima de “Expand All”, mantenha-o pressionado e arraste-o para a barra de favoritos. Caso a barra não esteja aparecendo, logo abaixo da barra de endereços (como na imagem abaixo), basta apertar as teclas Ctrl + Shift + B.



Com o favorito adicionado, é só voltar à página do Facebook, já com as postagens roladas, e clicar no botão “Expand All”, que você adicionou há pouco na barra de favoritos. Então, a ferramenta iniciará o procedimento de expandir todos os comentários. Às vezes ele é demorado, dependendo da quantidade de postagens



Etapa 3: Buscar e exportar os dados, utilizando a opção “Facebook – Comentários”, conforme explicado no subcapítulo 2.1.3.2.



2.1.6. Twitter

Etapa 1: Criar uma *database*, conforme explicado no subcapítulo 2.1.1.1.

Etapa 2: Adicionar nós, conforme explicado no subcapítulo 2.1.1.2.

Para nós, serão utilizadas palavras, frases ou expressões, *hashtags* (com o # antes) ou nomes de usuários (com o @ antes).

Importante salientar que, infelizmente, vários testes foram feitos e não foi possível pegar os *tweets* de um usuário (sua *timeline*).

Etapa 3: Carregar o *preset* “Twitter – Pesquisa ou hashtag” (dentro do grupo “Mídias sociais”), conforme explicado no subcapítulo 2.1.1.3.

Etapa 4: Obter os dados, conforme explicado no subcapítulo 2.1.1.4. O limite de dados obtidos foi estabelecido em 5 mil.

Etapa 5: Exportar os dados, conforme explicado no subcapítulo 2.1.1.5.

2.1.7. YouTube

Etapa 1: Criar uma *database*, conforme explicado no subcapítulo 2.1.1.1.

Etapa 2: Adicionar nós, conforme explicado no subcapítulo 2.1.1.2.

Para nó, informe uma palavra ou expressão a ser buscada. Se você quiser pesquisar no YouTube sobre a FEEVALE, pode adicionar o nó “FEEVALE”.

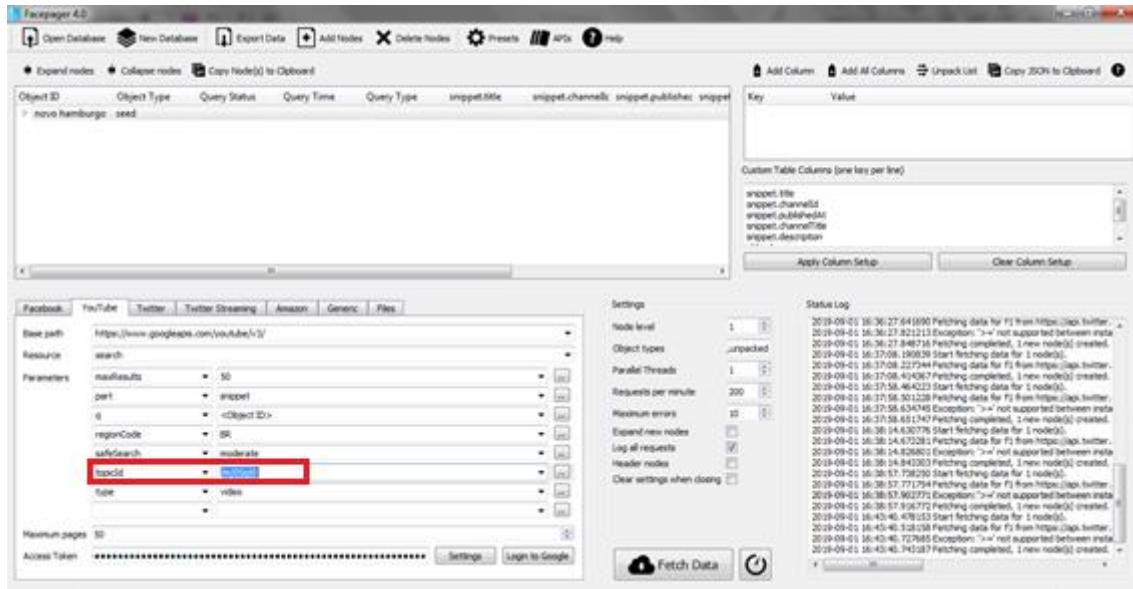
Etapa 3: Carregar um dos *presets* “YouTube” (dentro do grupo “Mídias sociais”), conforme explicado no subcapítulo 2.1.1.3.

Existem três *presets*: “YouTube - Pesquisa simples”, “YouTube - Pesquisa completa” ou “YouTube - Pesquisa por tópico”. “Pesquisa simples” procura apenas em vídeos; “Pesquisa completa” procura em vídeos, canais e *playlists*; “Pesquisa por tópico” consiste na Pesquisa simples segmentada por tópicos.

Caso tenha sido escolhida a opção por tópicos, é necessário definir o código no campo *topicId*. Alguns tópicos estão abaixo, e outros podem ser consultados em developers.google.com/youtube/v3/docs/search/list, na seção *topicId*.

Negócios	/m/09s1f
Conhecimento	/m/01k8wb
Sociedade	/m/098wr

Saúde	/m/0kt51
-------	----------



Etapa 4: Obter os dados, conforme explicado no subcapítulo 2.1.1.4. O limite de dados obtidos foi estabelecido em 2,5 mil.

Etapa 5: Exportar os dados, conforme explicado no subcapítulo 2.1.1.5.

2.1.8. Jornais, revistas, blogs e sites que usam RSS

Etapa 1: Criar uma *database*, conforme explicado no subcapítulo 2.1.1.1.

Etapa 2: Adicionar nós, conforme explicado no subcapítulo 2.1.1.2.

Para nós, você pode usar qualquer valor. Por exemplo, se for pegar o RSS da Folha de São Paulo, ao adicionar o nó, você pode colocar qualquer nome, como “folha”, “folhasp”, “folharss”, entre outros.

Etapa 3: Carregar o preset “Feeds RSS” (dentro do grupo “Mídias tradicionais”), conforme explicado no subcapítulo 2.1.1.3.

Etapa 4: Buscar e configurar a fonte de dados.

Vários *sites* de notícias, como jornais e revistas, bem como blogs, e mesmo outros *sites*, utilizam a tecnologia RSS, que é um formato de distribuição de conteúdo muito usado para compartilhar notícias e informações. Caso uma determinada mídia tradicional utilize RSS, o processo é simples.

Primeiro, é necessário buscar o endereço do seu *feed*. Muitas vezes, há um pouco de trabalho para saber se o RSS está disponível, mas geralmente os *sites* exibem um ícone no padrão RSS, conforme destacado abaixo.



Após acessar o *feed*, ele possui a seguinte aparência:

```

<rss xmlns:atom="http://www.w3.org/2005/Atom" xmlns:media="http://search.yahoo.com/mrss/" version="2.0">
  <channel>
    <item>
      <category><![CDATA[ futebol feminino ]]>
      <title>
        <![CDATA[
          Brasil perde do Chile e fica com vice no 1º torneio de Pia Sundhage no comando
        ]]>
      </title>
      <author>Agência Estado</author>
      <link>
        https://www.jornaldocomercio.com/_conteudo/esportes/2019/09/701011-brasil-perde-do-chile-e-fica-com-vice-no-1-torneio-de-pia-sundhage-no-comando.html
      </link>
      <description>
        Neste domingo, na decisão de um torneio amistoso no estádio do Pacaembu, em São Paulo, o Brasil ficou no empate sem gols contra o Chile e foi derrotado na decisão por pênaltis
        po 5 a 4
      </description>
      <pubDate>Sun, 1 Sep 2019 16:54:27 -0300</pubDate>
      <image>
        <![CDATA[ ]]>
      </image>
      <linkfoto>
        <![CDATA[ ]]>
      </linkfoto>
      <creditfoto>
        <![CDATA[ ]]>
      </creditfoto>
      <image>
        <![CDATA[ ]]>
      </image>
      <media:content type="image/jpg" expression="full" width="" height="">
        <![CDATA[ ]]>
      </media:content>
      <media:description type="plain">
        <![CDATA[ ]]>
      </media:description>
    </item>
  </channel>
</rss>

```

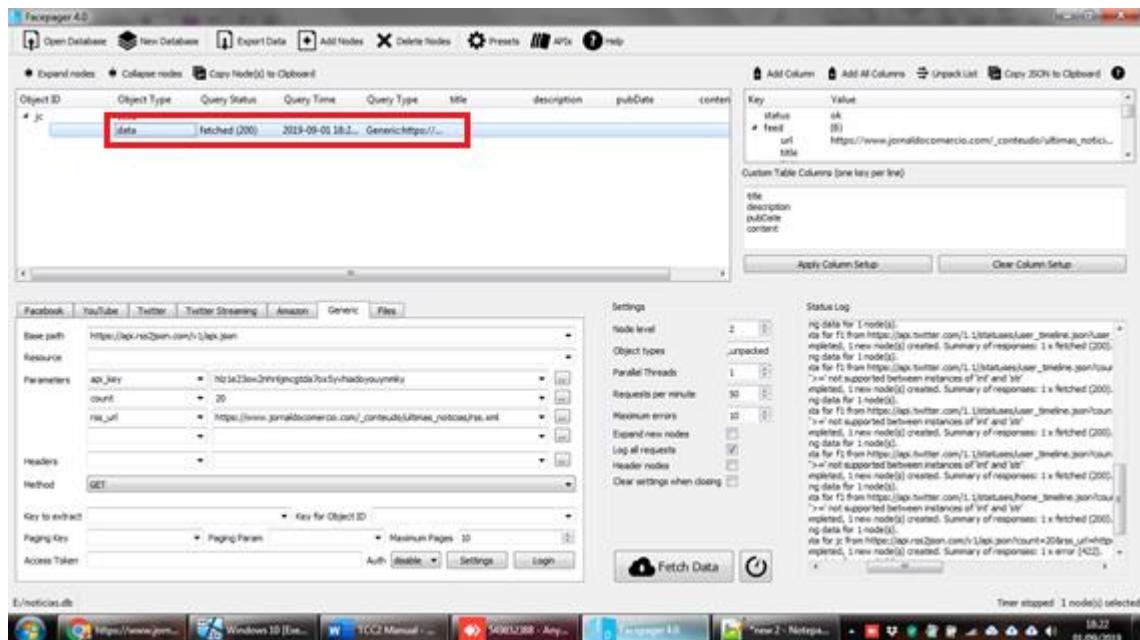
Basta copiar o endereço para colocar no Facepager. O endereço copiado anteriormente deve ser colado no campo “rss_url”.

The screenshot shows the Facepager configuration interface. The 'rss_url' field is highlighted with a red box. The 'Settings' pane includes fields for 'Node level' (set to 3), 'Object types' (set to 'unchecked'), 'Parallel Threads' (set to 1), 'Requests per minute' (set to 50), 'Maximum errors' (set to 10), and various checkboxes for 'Expand new nodes', 'Log all requests', 'Header nodes', and 'Clear settings when closing'. The 'Status Log' pane displays a summary of responses for the specified URL, showing 1 node(s) created and 1 x fetched (200). The bottom status bar indicates 'Total elapsed: 3 ms'.

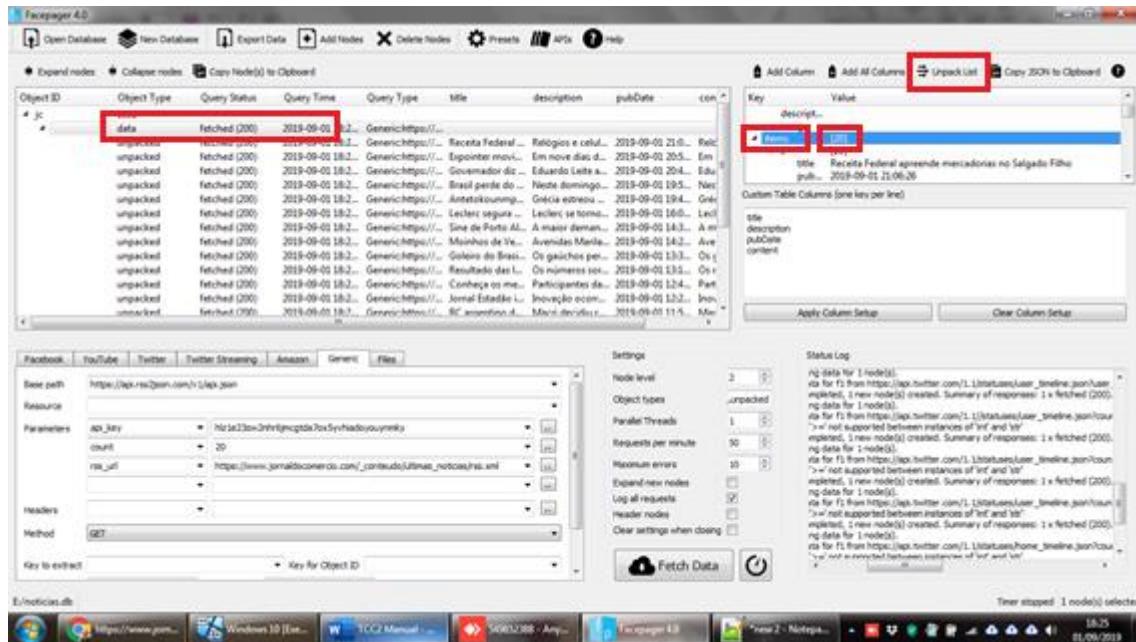
Etapa 5: Obter os dados, conforme explicado no subcapítulo 2.1.1.4. O limite de dados obtidos varia conforme o RSS.

Etapa 6: Expandir os dados obtidos.

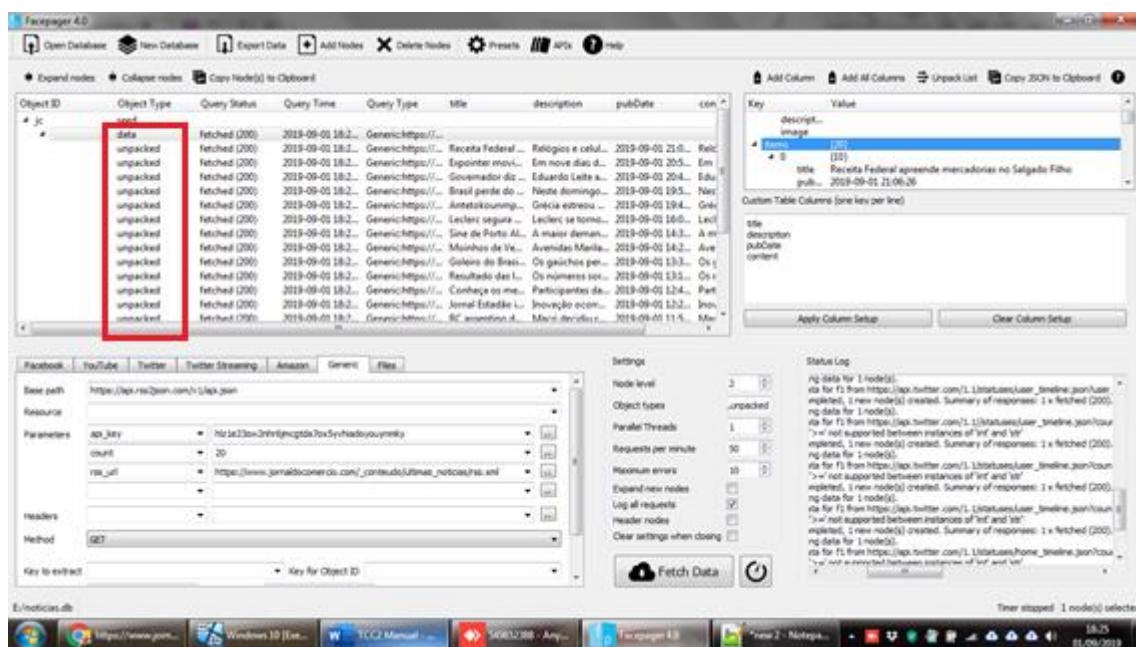
Após a busca, diferentemente das demais buscas, aparece apenas uma linha, cujo valor na coluna “Object Type” é “data”. Isso não quer dizer que apenas um registro retornou, mas que deve ser feito um procedimento adicional.



Para exibir todos os registros obtidos, é necessário selecionar esta única linha. Em seguida, na caixa “Key/Value”, que fica ali no canto superior direito, role pra baixo e procure a chave (*key*) cujo valor é “items”. Provavelmente, na coluna “value”, virá entre colchetes a quantidade de itens que a busca retornou – geralmente “[10]”, “[20]” ou “[30]”. Então, para exibir todos estes registros, clique no botão “Unpack List”.



Após este procedimento, aparecem linhas abaixo daquela linha “data”, cujo tipo vai ser “unpacked”. Cada uma delas representa um item buscado no RSS.



Etapa 7: Exportar os dados, conforme explicado no subcapítulo 2.1.1.5.

2.1.9. Demais sites e mídias tradicionais

Etapa 1: Criar uma *database*, conforme explicado no subcapítulo 2.1.1.1.

Etapa 2: Adicionar nós, conforme explicado no subcapítulo 2.1.1.2.

Para nós, será utilizada a palavra ou expressão a ser buscada. Se você quiser pesquisar nos *sites* configurados sobre a FEEVALE, pode adicionar o nó “FEEVALE”.

Etapa 3: Carregar o *preset* “Google – Pesquisa” (dentro do grupo “Mídias tradicionais”), conforme explicado no subcapítulo 2.1.1.3.

Etapa 4: Obter os dados, conforme explicado no subcapítulo 2.1.1.4. O limite de dados obtidos é de 100.

Antes de obter os dados, é necessário que você tenha lido e feito as configurações orientadas no subcapítulo 1.3.6, para configurar sua CSE.

Devido à baixa quantidade de registros obtidos, você pode, se quiser, filtrar a busca, especificando um período específico. Para isto, altere o campo “sort” e troque “date” por

```
date:r:DATAINICIAL:DATAFINAL
```

sendo que ambas as datas devem estar no formato AAAAMMDD (ano com 4, mês com 2 e dia com 2, tudo junto). Por exemplo, se você quiser buscar entre 1º de fevereiro de 2018 e 1º de março de 2019, ficaria assim:

```
date:r:20180201:20190301
```

Etapa 5: Exportar os dados, conforme explicado no subcapítulo 2.1.1.5.

2.2. CONVERSÃO DOS ARQUIVOS

Antes de passar para a próxima fase, é necessário converter os arquivos gerados para um formato específico de planilha: o XLS.

Basicamente, o Facepager gera arquivos em formato CSV, e o Data Miner gera em formato XLSX. Até foi testado usar estes arquivos direto na próxima fase, mas não deu certo. Então, temos que converter. Para isto, você precisa ter um software de planilha eletrônica, que pode ser o Excel (Microsoft Office) ou o Calc (LibreOffice).

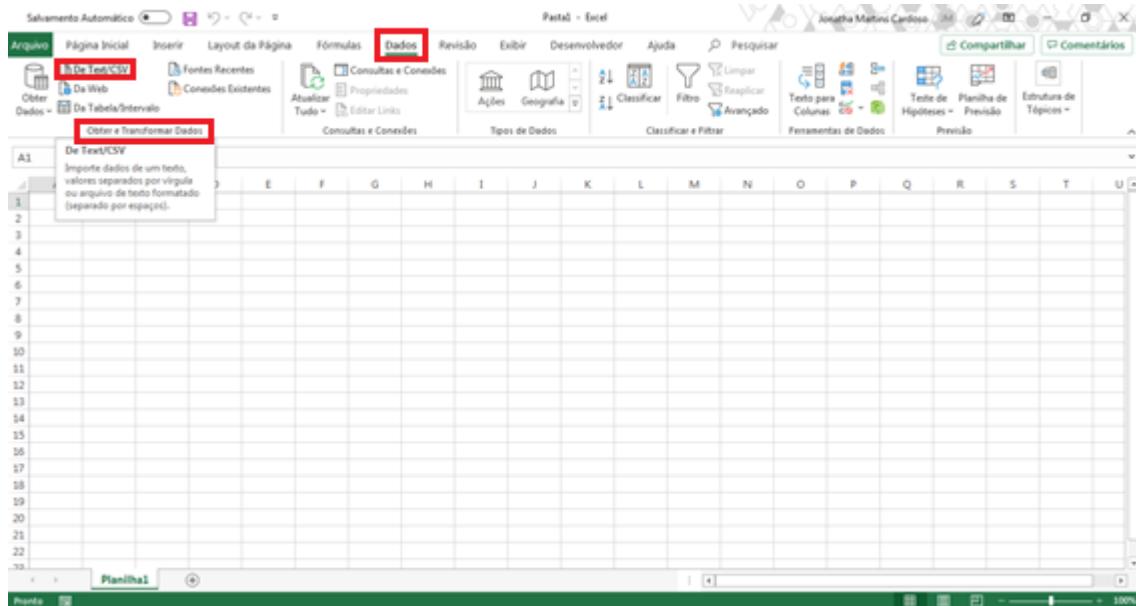
Caso você não tenha nenhuma das opções instaladas, sugerimos instalar o LibreOffice, por ser gratuito e trabalhar bem com estes formatos. Ele está disponível no endereço pt-br.libreoffice.org/baixe-ja/libreoffice-novo.



2.2.1. Arquivos CSV (Facepager)

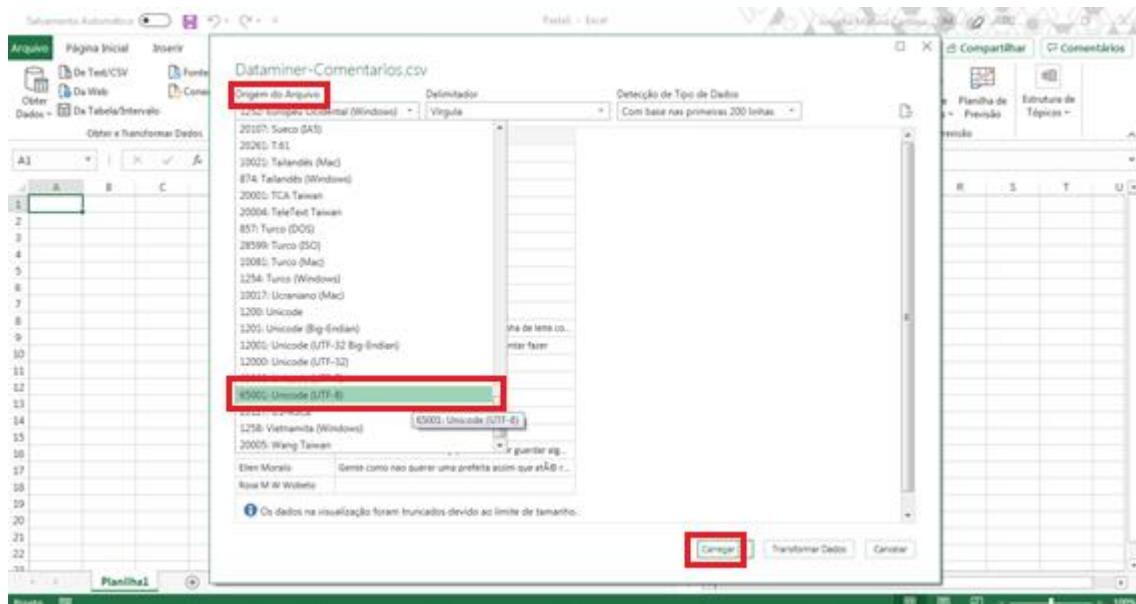
No **Excel**:

Abra um arquivo novo. Após isto, selecione a guia “Dados”, e clique na “De Texto” (em versões antigas) ou “De Text/CSV” (em versões mais novas), dentro do grupo Obter Dados.

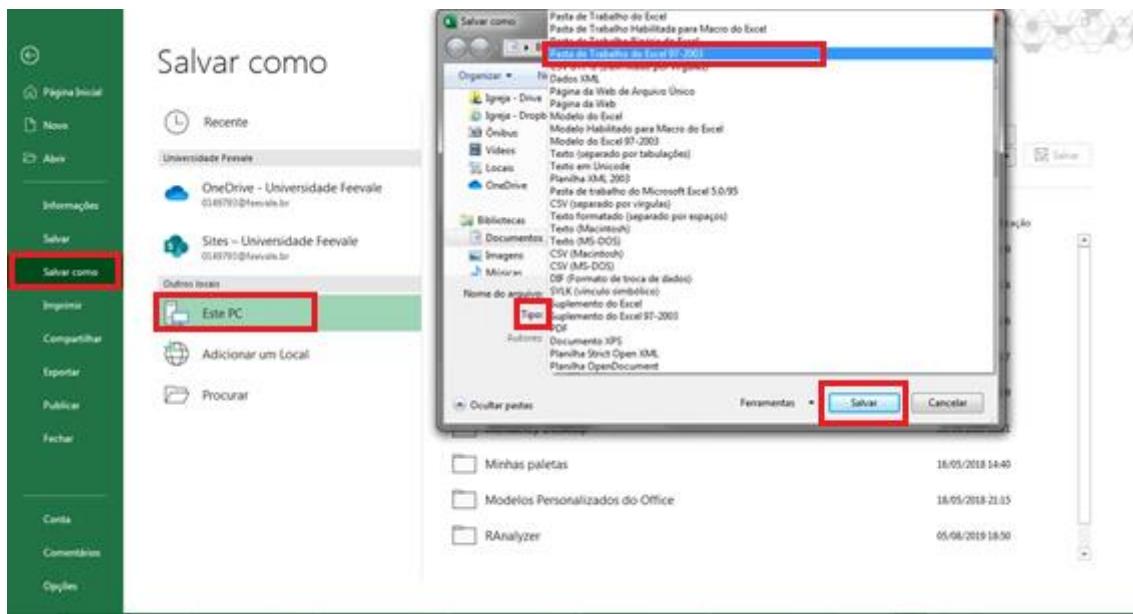


Abrirá uma janela para abrir o arquivo CSV. Escolha o arquivo desejado – um daqueles que você exportou no Facepager.

Em seguida, vai aparecer uma janela de importação. Selecione, no item “Origem do Arquivo”, no topo da janela, a opção “Unicode (UTF-8)”. Por fim, clique em “Carregar”.



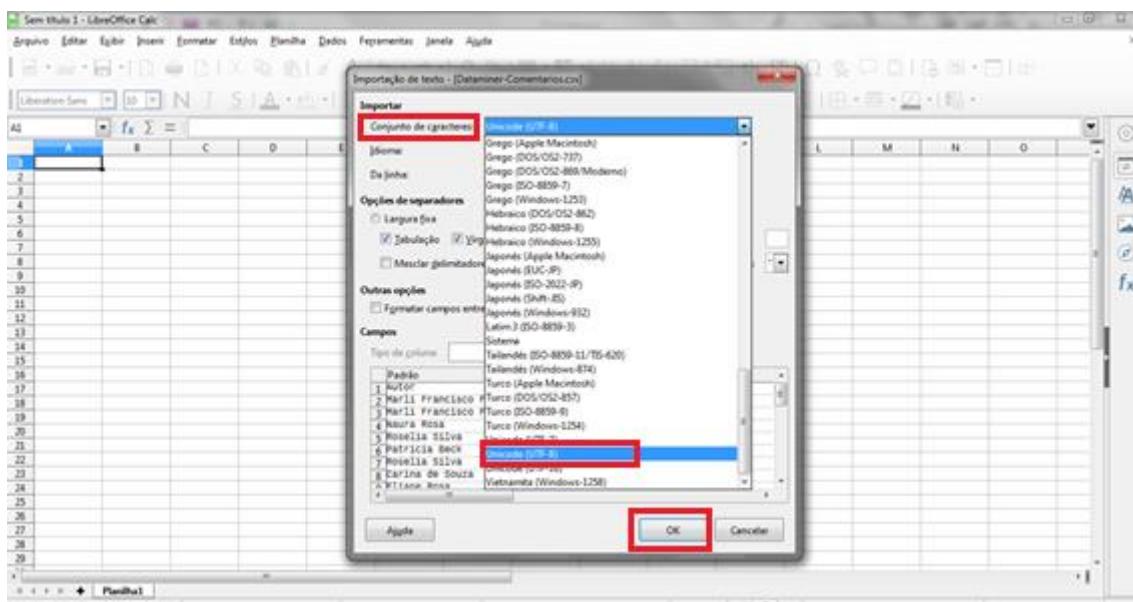
Após carregar o arquivo, você já pode salvá-lo. Para isto, clique no menu Arquivo – lá no canto superior esquerdo. Escolha a opção “Salvar como”. Nas versões mais novas, é necessário clicar em “Este PC”, para aparecer a janela de salvar arquivo. Então, é só alterar o tipo, selecionando a opção “Pasta de Trabalho do Excel 97-2003”. Salve aonde quiser e com o nome que desejar.



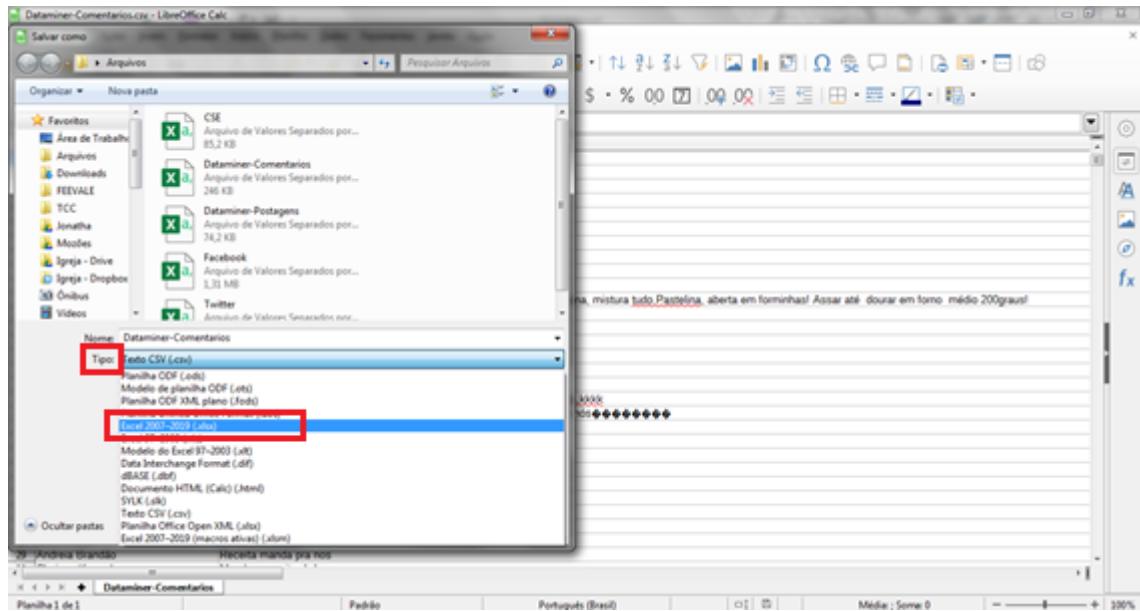
No Calc:

Abra um arquivo novo. Vá no menu “Arquivo” (o primeiro da barra superior) e clique em “Abrir”. Escolha o arquivo CSV desejado – um dos que você exportou no Facepager.

Após abrir o arquivo, irá aparecer uma janela de importação de texto. A única coisa que você deve ver é a opção selecionada em “Conjunto de caracteres”, que deve ser “Unicode (UTF-8)”. Por fim, só clicar em OK.



Após a importação, você pode salvar o arquivo. Basta ir novamente no menu “Arquivo”, clicando na opção “Salvar como”. A opção selecionada, em “Tipo”, deve ser “Excel 97-2003 (.xls)”.



Após informar o nome que quiser e selecionar a pasta que desejar, vai aparecer uma janela de confirmação. Basta clicar na opção “Utilizar o formato Excel...”

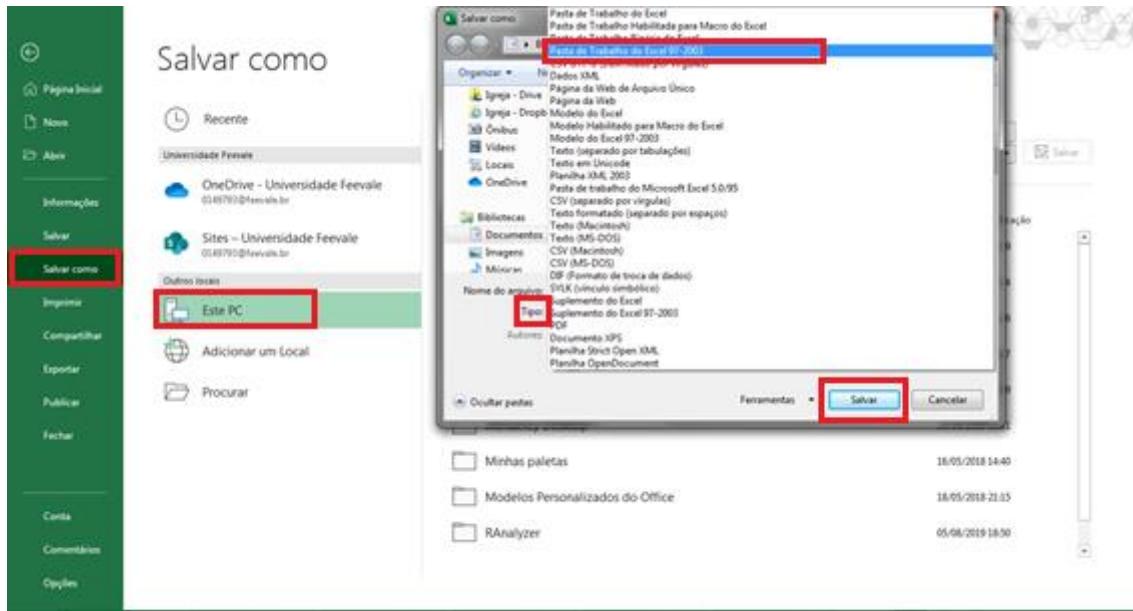
2.2.2. Arquivos XLSX

Para os arquivos em formato XLSX, gerados pelo *Data Miner*, a situação é muito mais simples. Basta abrir o arquivo, clicando duas vezes em cima dele. O mesmo abrirá no Excel ou no Calc – tanto faz em qual abrir.

Em seguida, para salvar em XLS, basta repetir os procedimentos mencionados antes.

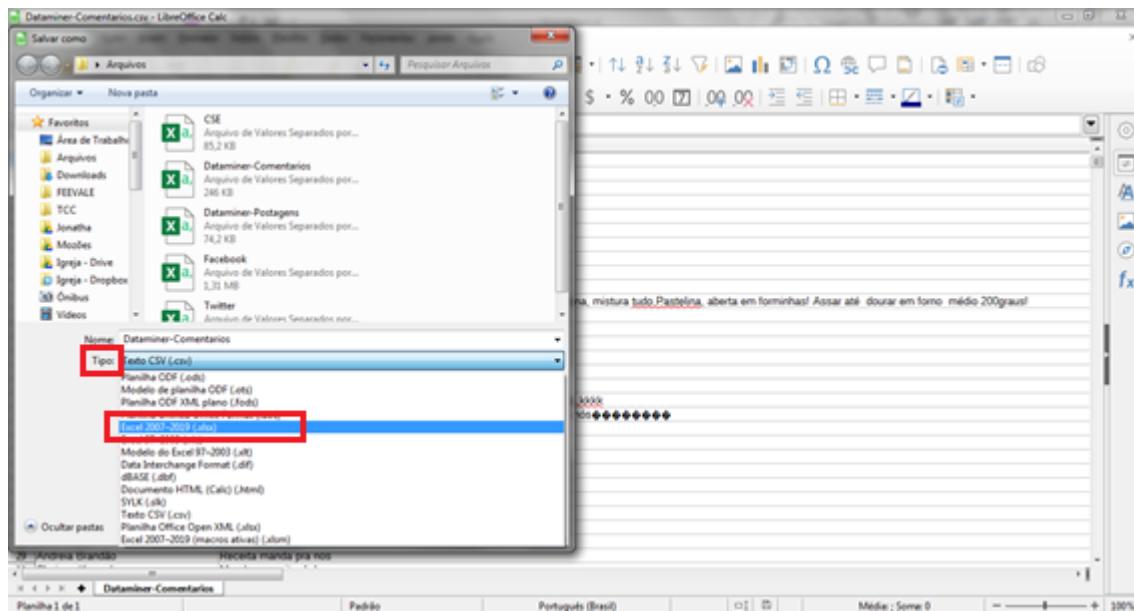
No Excel:

Clique no menu Arquivo – lá no canto superior esquerdo. Escolha a opção “Salvar como”. Nas versões mais novas, é necessário clicar em “Este PC”, para aparecer a janela de salvar arquivo. Então, é só alterar o tipo, selecionando a opção “Pasta de Trabalho do Excel 97-2003”. Salve aonde quiser e com o nome que desejar.



No Calc:

Basta ir no menu “Arquivo”, clicando na opção “Salvar como”. A opção selecionada, em “Tipo”, deve ser “Excel 97-2003 (.xls)”.



Após informar o nome que quiser e selecionar a pasta que desejar, vai aparecer uma janela de confirmação. Basta clicar na opção “Utilizar o formato Excel...”

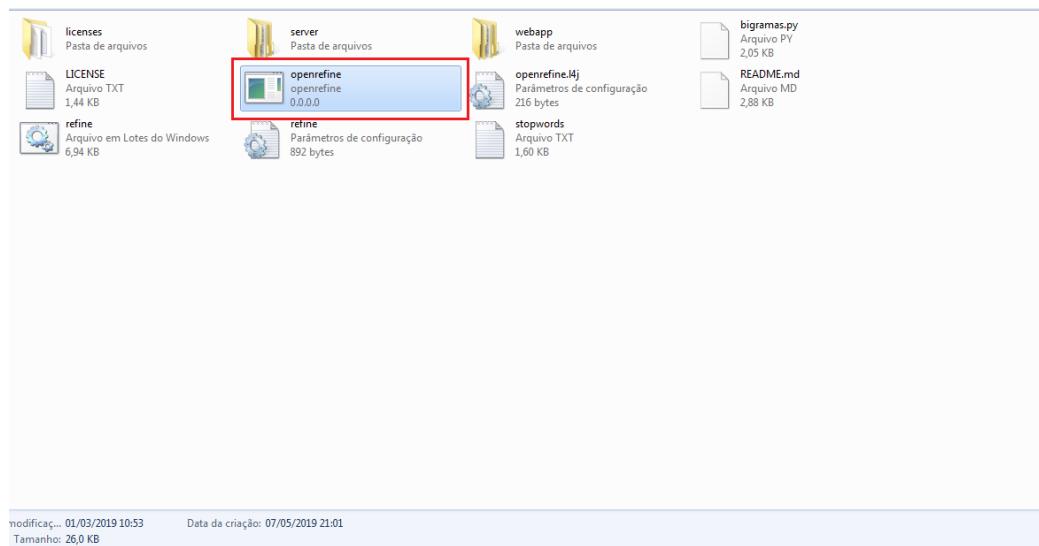
2.3. PRÉ-PROCESSAMENTO

Nesta fase, os dados que você coletou serão preparados, antes de que suas informações sejam extraídas. Essa preparação consiste em uma limpeza dos dados coletados. Principalmente no Facepager, muitos dados são desnecessários e podem ser limpos/excluídos.

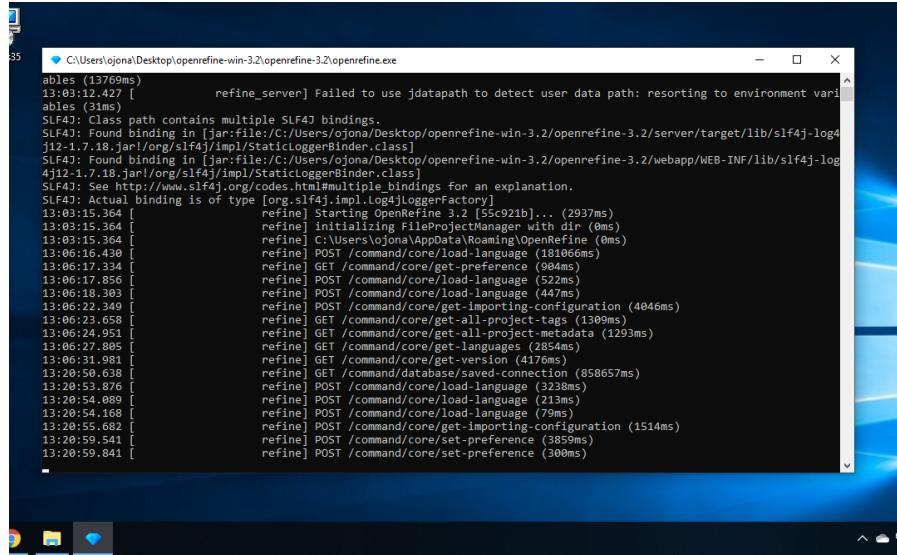
Outro ponto interessante é que ele permite fazer alguns filtros. Por exemplo, após converter as datas, você pode filtrar e selecionar o período que lhe interessa para, depois, exportar os dados e extrair informações deles. Todo o processo será feito no software OpenRefine.

IMAGEM

Para abri-lo, basta ir ao Menu Iniciar, acessar a lista de programas e clicar em “OpenRefine”. Vai abrir uma janela, e basta dar dois cliques no arquivo “openrefine”.



Depois de dar dois cliques, vai abrir uma janela preta, exibindo vários comandos. É uma janela do *prompt* de comando (CMD). Não a feche, pois é necessária para que o OpenRefine funcione.



```

35  C:\Users\ojona\Desktop\openrefine-win-3.2\openrefine-3.2\openrefine.exe

ables (13769ms)
13:03:12.427 [           refine_server] Failed to use jdatapath to detect user data path: resorting to environment vari
ables (31ms)
SLF40: Class path contains multiple SLF4J bindings.
SLF40: Found binding in [jar:file:/C:/Users/ojona/Desktop/openrefine-win-3.2/openrefine-3.2/server/target/lib/slf4j-log4
j12-1.7.18.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF40: Found binding in [jar:file:/C:/Users/ojona/Desktop/openrefine-win-3.2/openrefine-3.2/webapp/WEB-INF/lib/slf4j-log
4j12-1.7.18.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF40: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF40: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
13:03:15.364 [           refine] Starting OpenRefine 3.2 [5c921b]... (2927ms)
13:03:15.364 [           refine] Refining command line arguments with dir (0ms)
13:06:16.430 [           refine] POST /command/core/load-language (48100ms)
13:06:17.334 [           refine] GET /command/core/get-preference (904ms)
13:06:17.856 [           refine] POST /command/core/load-language (522ms)
13:06:18.303 [           refine] POST /command/core/load-language (447ms)
13:06:22.349 [           refine] POST /command/core/get-importing-configuration (4046ms)
13:06:23.658 [           refine] GET /command/core/get-all-project-tags (1309ms)
13:06:24.951 [           refine] GET /command/core/get-all-project-metadata (1293ms)
13:06:27.385 [           refine] GET /command/core/get-languages (2854ms)
13:06:31.981 [           refine] GET /command/core/get-version (4176ms)
13:20:50.638 [           refine] GET /command/database/saved-connection (858657ms)
13:20:53.876 [           refine] POST /command/core/load-language (3238ms)
13:20:54.089 [           refine] POST /command/core/load-language (213ms)
13:20:54.168 [           refine] POST /command/core/load-language (79ms)
13:20:55.682 [           refine] POST /command/core/get-importing-configuration (1514ms)
13:20:59.541 [           refine] POST /command/core/set-preference (3859ms)
13:20:59.841 [           refine] POST /command/core/set-preference (380ms)

```

Após alguns segundos, vai abrir uma janela do Google Chrome, exibindo o software, pronto para uso.

2.3.1. Uso do OpenRefine

Com o *software*, você poderá filtrar e organizar os dados que quiser. Vamos dar alguns exemplos:

- No Facebook com Facepager, filtrando por data e tipo, você pode apenas exibir as postagens de vídeos entre março e maio de um ano.
- No Facebook para Data Miner, filtrando por autor, você pode filtrar apenas os comentários feitos pelo próprio autor da página, e não feitas por outras pessoas.

Veja bem: Você conseguirá obter informações bem interessantes! Por exemplo, no Facebook, saber quantas postagens você fez entre um período, a distribuição de tipos (quantos vídeos, quantas fotos...), ver quem é que mais comenta, entre outras.

2.3.1.1. *Projects* (Projetos)

Cada arquivo que é importado torna-se um projeto. Assim, para iniciar o pré-processamento, clique primeiro em “Create Project”; na opção “Get data from”, clique em “This computer”. Por fim, clique no botão “Escolher arquivos”.

OpenRefine A power tool for working with messy data.

Create Project

Open Project Import Project Language Settings

Get data from

This Computer Escolher arquivos Nenhum arquivo selecionado

Web Addresses (URLs) Clipboard Database Google Data

Next »

Version 3.2-beta [8d89a2a]

Preferences Help About

Após selecionar o arquivo – escolha um dos que você salvou em XLS no subcapítulo anterior – clique em “Next”.

DICA: Você pode salvar projetos e reabri-los posteriormente. Basta, na página inicial, clicar em “Open project” e selecionar o arquivo de projeto desejado.

Após isto, mantenha todas as opções conforme o padrão. Apenas defina um nome para o projeto, no alto da tela, e clique em “Create project”.

OpenRefine A power tool for working with messy data.

Create Project

Open Project Import Project Language Settings

Start Over Configure Parsing Options Project name: CSE.xls Create Project

level	id	parent_id	object_id	object_type	query_status	query_time	query_type	displayLink	title	snippet
1	0.0	1.0	versadores	seed	None	None	None	www.martinbehrend.com.br	Biblioteca	há 5 hora
2	1.0	2.0	Biblioteca Municipal Machado de Assis - Martin Behrend	data	fetched (200)	2019-06-05 19:51:49,319072	Generic https://www.googleapis.com/customsearch/v1	Municipal Machado de Assis - Martin Behrend	de Vassouras	esta com reformas
3	1.0	3.0	1.0	data	fetched (200)	2019-06-05 19:51:49,319072	Generic https://www.googleapis.com/customsearch/v1	novachambugo.rs.gov.br	Festa marcial	atrasadas final de se

Parse data as:

Excel files JSON files Line-based text files CSV / TSV / separator-based files Fixed-width field text files PC-Axis text files MARC files JSON-LD files RDF/NDL files

Worksheets to import: CSE.xls#CSE 112 rows

Ignore first: 0 line(s) at beginning of file
Parse next: 1 line(s) as column headers
Discard initial: 0 row(s) of data
Load at most: 0 row(s) of data

Store blank rows
Store blank cells as nulls
Store file source (file names, URLs) in each row

Update Preview

Version 3.2-beta [8d89a2a]

Preferences Help About

Após isto, os dados estão prontos para serem filtrados.

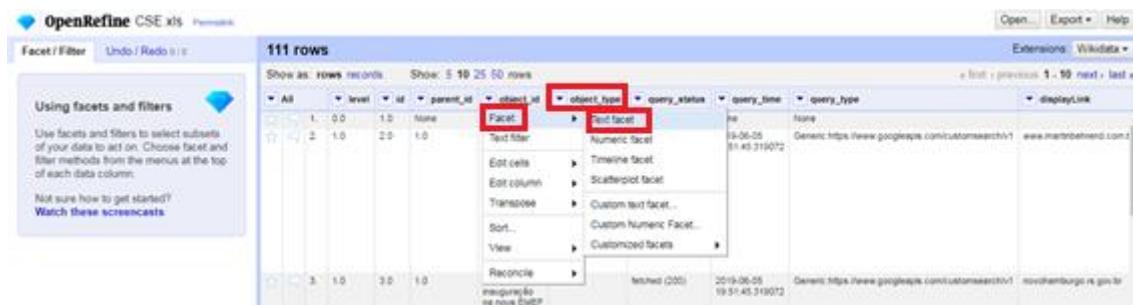
2.3.1.2. *Text facet* (filtro por texto)

O filtro de texto não serve para buscar uma “expressão” nos dados da coluna. Na verdade, ele agrupa os valores/categorias diferentes que aparecem em uma

coluna. Por exemplo, ao agrupar os dados da coluna “type”, o OpenRefine encontrou 4 valores diferentes: “link”, “photo”, “status” e “video”, além de vários registros com valor em branco (“blank”).



Para filtrar/agrupar por texto, clique na seta que fica à esquerda da coluna desejada, vá na opção “Facet” e clique em “Text facet”.



Após isto, na coluna à esquerda da janela irão aparecer as diferentes categorias/valores da coluna e a quantidade de ocorrências de cada uma delas. Acima, aparece o nome da coluna em cima.

Para filtragem em texto, as categorias/valores vão aparecer em uma listagem. Veja abaixo o nome da coluna (“type”), a quantidade de categorias/valores diferentes (um número acompanhado da palavra “choices”) e quais são essas categorias/valores (destacado em azul), com a respectiva quantidade de ocorrências à direita (destacado em cinza).



2.3.1.3. Choices (opções)

Após fazer um filtro de texto, caso você deseje ver apenas os dados de uma das categorias/valores, clique em cima dele. Automaticamente, na listagem, vão aparecer apenas os dados daquele valor, e o mesmo estará destacado em laranja:

The screenshot shows the OpenRefine interface with a Facebook dataset. A facet for 'from.name' is selected, showing 'Patricia Beck' (901) as the chosen value. The main list displays 901 matching rows out of 3695 total. The results show four posts from Patricia Beck:

- POR QUE O GOVERNO INSISTE EM NÃO APRESENTAR A PRESTAÇÃO DE CONTAS?
- TEMOS CONSULTA COM PROCTOLOGISTA E EXAMES DE COLONOSCOPIA PARA QUEM PRECISA? Esta foi uma das perguntas que fiz à Prefeitura Municipal, pelo grande número de pessoas que fazem contato pedindo orientações pela demora ou falta de consultas com proctologista ou exames de colonoscopia. Queremos saber quanto tempo é a fila de quem aguarda. Dentre as respostas recebidas estão: 1) que a média de espera por consulta com proctologista é de 89,80 dias! 2) que existem na fila de espera para consulta com proctologista 1.342 pacientes! 3) e que na fila de espera para exame de colonoscopia estão 1.357 pacientes!
- Bom dia!
- Linda minha mãe né!

To the right, a detailed view of the posts is shown, with Patricia Beck's name and timestamp highlighted in orange.

Lembre-se que, no alto da janela, aparecem as informações de quantas linhas você possui no total, e quantas linhas foram filtradas. Por exemplo, na imagem acima, existem 3.695 postagens, mas como foi filtrado com uma das categorias/valores da coluna “from.name”, aparecem apenas 901 postagens.

Perceba que o OpenRefine não exibe todas as linhas, mas as divide em páginas. Você pode navegar entre elas, conforme clica nos botões à direita (*first* para primeira página, *previous* para página anterior, *next* para próxima página e *last* para última página). Você ainda pode definir quantas linhas são exibidas por página, de 5 a 50 registros.

Você pode também exibir mais de uma categoria/valor de coluna. Na imagem acima, apenas uma das categorias/valores foi filtrada. Mas para filtrar mais categorias/valores, você deve posicionar o mouse em cima destas bem para o lado direito. Vão aparecer duas opções: “edit” e “include”. Clique em “include”. Repita isto para todas as categorias/valores que deseja exibir.

The screenshot shows the OpenRefine interface with a Facebook dataset. A facet for 'object_type' is selected, showing several choices: 'data' (3703), 'empty' (654), 'offcut' (359), and 'seed' (1). The 'edit' and 'include' buttons are visible next to the choice list.

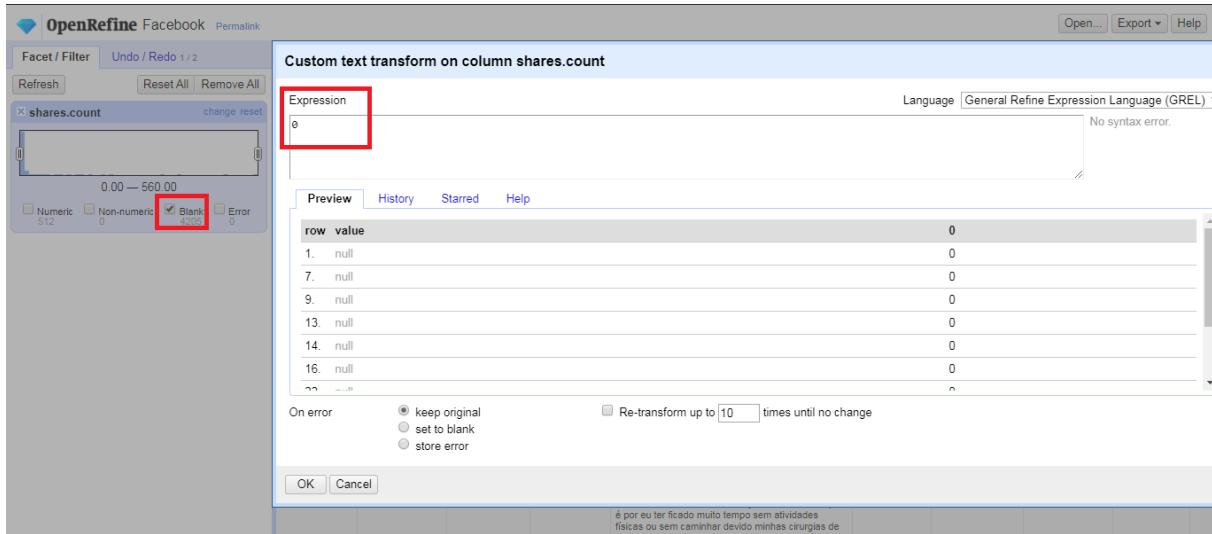
Da mesma forma, as categorias/valores selecionadas estarão em laranja. Se quiser ocultar algum deles da listagem, só clicar em “exclude”, à direita.

2.3.1.4. Numeric facet (filtro por texto)

O filtro numérico exibe um gráfico, mostrando a evolução dos números de uma coluna. Primeiro, é necessário preparar a coluna. Clique na seta da coluna desejada na tabela. No menu “Edit cells”, vá em “Common transforms” e clique em “To number”. Automaticamente a conversão é efetuada.

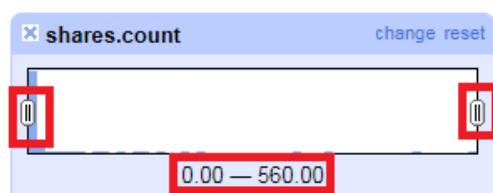
Porém, pode haver linhas cujo valor da coluna está em branco. Desta forma, esses campos não são entendidos como zero. Se você não corrigir isto, eles vão aparecer, no filtro, como “Blank”, e não como “Numeric”.

Para que eles sejam classificados como “zero”, vá no menu “Edit cells”, clique em “Transform”. No campo “Expression”, basta colocar um zero (“0”) e clicar em OK.



Agora já é possível fazer o filtro. Para filtrar/agrupar por número, clique na seta que fica à esquerda da coluna desejada, vá na opção “Facet” e clique em “Numeric facet”.

Vai aparecer, na barra à esquerda, um gráfico, mostrando a evolução dos números. É possível filtrar, definindo o limite inicial e final. Basta clicar e arrastar nos botões que estão em extremidade da barra.



Conforme você ajusta, ali embaixo do gráfico ele exibe os valores que você delimitou.

2.3.1.5. *Timeline facet* (filtro por data/tempo)

O filtro de data exibe um gráfico, mostrando a evolução das datas de uma coluna desse tipo. Primeiro, é necessário preparar a coluna. Clique na seta da coluna desejada na tabela. No menu “Edit cells”, clique em “Transform”.

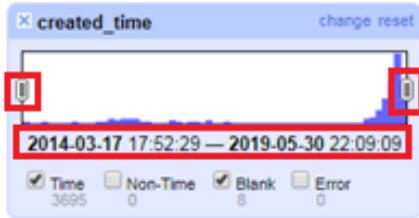
No campo “Expression”, você deve colocar um ou mais códigos. Esses códigos serão exibidos ao longo dos próximos subcapítulos. Basta copiar aqui do texto e colar no campo indicado abaixo – um código de cada vez.

Se à esquerda desse campo aparecer um erro em vermelho, verifique se o código foi colocado da mesma forma que mostrado e explicado neste manual. Do contrário, se estiver tudo certo, basta clicar em OK.

IMAGEM

Agora já é possível fazer o filtro. Para filtrar/agrupar por número, clique na seta que fica à esquerda da coluna desejada, vá na opção “Facet” e clique em “Timeline facet”.

Vai aparecer, na barra à esquerda, um gráfico, mostrando a evolução do tempo. É possível filtrar, definindo o limite inicial e final. Basta clicar e arrastar nos botões que estão em extremidade da barra. Conforme você ajusta, ali embaixo do gráfico ele exibe o período que você delimitou.

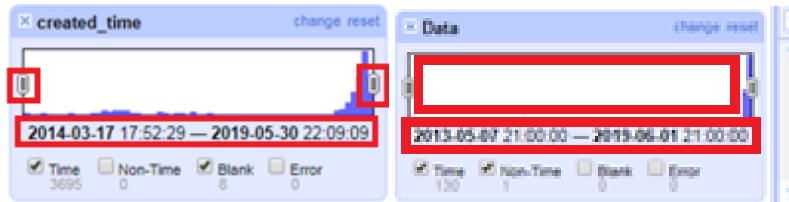


Algumas vezes, aparecerão valores classificados como “Non-Time/Non-Numeric”, “Blank” e “Error”. Significa que esses valores não puderam ser classificados como data ou número.

Você pode exclui-los, conforme mostramos acima – basta deixar marcadas essas opções, para exibir na lista os valores errados – ou corrigi-los. Por exemplo, às vezes a data está com algum erro de escrita, e para isto basta que você faça a edição do campo, conforme descrito no subcapítulo 2.3.1.7.

2.3.1.6. Outliers

Podemos chamar de *outliers*, neste manual, os valores que estão muito fora do comum, seja em data ou em número. Para ficar mais claro, veja os exemplos:

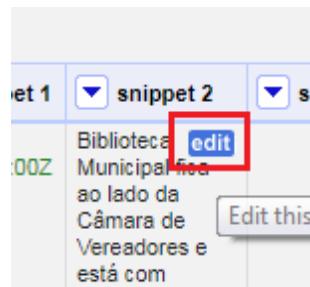


Cada barra roxa indica uma quantidade de registros (quanto maior, mais registros), conforme a imagem da esquerda. Às vezes, como mostra a imagem da direita, há um grande período com poucos ou nenhum registro – representada por uma grande área em branco. Neste caso, você pode, se quiser, excluir esses registros. Na imagem, os dados começam em 2013 e vão até 2019. Porém, há um grande espaço em branco, com poucas ou nenhuma postagem.

Para remover as linhas destes valores ou datas, basta selecionar esse período, usando as barras nas laterais. Em seguida, faça a exclusão das linhas, conforme mostramos no subcapítulo 2.3.1.12.

2.3.1.7. Edit value (editar valor)

Caso você precise alterar um valor da tabela, basta posicionar o mouse em cima dele e clicar em “Edit”.



2.3.1.8. Join choices (juntar/agrupar opções)

Em algumas situações, podem aparecer várias linhas com valores diferentes em uma coluna, mas que na verdade o valor deveria ser o mesmo. Por exemplo, na imagem abaixo, há várias postagens feitas pelo autor da página do Facebook, mas que, por terem sido marcadas com outras pessoas, aparecem como “Fulano está com...”, mas o autor da postagem é sempre o mesmo: “Fulano”.

The screenshot shows the OpenRefine interface with a list of 124 records. On the left, there is a facet for the 'Autor' column, which lists several entries under 'Cluster' 1, all of which are 'Patricia Beck'. A red box highlights this cluster. On the right, a specific record for 'Patricia Beck' is selected. A modal dialog box is open over the record, also labeled 'Patricia Beck'. The modal contains buttons for 'Apply' and 'Cancel'.

Neste caso, valeria a pena agrupar os autores para “Fulano”. Basta editar essas categorias/valores e renomeá-los para um só. Para isto, posicione o mouse em cima deles bem para o lado direito. Escolha a opção “edit”.

The screenshot shows the OpenRefine interface with a facet titled 'object_type'. It lists four choices: 'data' (3703), 'empty' (654), 'offcut' (359), and 'seed' (1). The 'edit' button next to 'seed' is highlighted with a red box.

Defina o novo nome dessa categoria/valor. Repita isto para todas as categorias/valores que deseja agrupar.

2.3.1.9. Rename choices (renomear opções)

Às vezes alguma categoria/valor está com um nome que você deseja alterar. Veja, por exemplo, que os comentários obtidos pelo Facepager aparecem como o tipo “blank”. Para deixar esse tipo da mesma forma que estão os demais, basta renomeá-lo.

The screenshot shows the OpenRefine interface with a facet titled 'type'. It lists five choices: 'link' (78), 'photo' (656), 'status' (145), 'video' (118), and '(blank)' (265). A modal dialog box is open, showing the word 'comentários' in the input field. The 'Apply' button is highlighted with a red box.

Basta posicionar o mouse em cima deles bem para o lado direito. Escolha a opção “edit”. Defina o novo nome dessa categoria/valor.

2.3.1.10. Close filter (fechar filtro).

Caso não queira mais ver as opções de filtro para alguma coluna, basta clicar no X à esquerda do nome da coluna, conforme a imagem:

The screenshot shows the OpenRefine interface with a facet titled 'type'. It only shows one choice: 'link' (78). The 'X' icon next to 'type' is highlighted with a red box.

2.3.1.11. Remove rows (Excluir linhas)

Em algumas situações, pode ser necessário excluir as linhas filtradas. Por exemplo, se você deseja excluir um determinado autor de comentários, após fazer o filtro, fica mais fácil de excluir todos os comentários de uma só vez, do que ter que ir em um a um dos comentários e fazer a exclusão.

Para excluir todas as linhas que foram filtradas, a opção é simples. Basta clicar na seta da primeira coluna, denominada “All”. Em seguida, na opção “Edit rows”, clique em “Remove all matching rows”.

The screenshot shows the OpenRefine interface with a table titled "11 matching rows (111 total)". The first column has a dropdown arrow pointing down, which is highlighted with a red box. A context menu is open at this position, with the option "Remove all matching rows" also highlighted with a red box. The menu path "Edit rows > Remove all matching rows" is visible. The table contains several columns: object_type, id, parent_id, object_id, query_status, query_time, query_type, and displayLink. The data rows show various entries, mostly with "object_type" values like "official" and "unofficial".

Automaticamente, os registros são excluídos e aquelas categorias/valores agrupadas na coluna – as que apareciam destacados em laranja - deixam de existir. Na imagem acima, foram filtrados 11 valores de duas categorias selecionadas da coluna “object_type”. Esses valores não existem mais, tampouco as duas categorias.

2.3.1.12. Remove column (remover coluna)

Para remover uma coluna, você deve clicar na seta à esquerda da coluna, ir à opção “Edit column” e escolher “Remove this column”.

The screenshot shows the OpenRefine interface with a table titled "100 rows". The first column has a dropdown arrow pointing down, which is highlighted with a red box. A context menu is open at this position, with the option "Remove this column" highlighted with a red box. The menu path "Edit column > Remove this column" is visible. The table contains several columns: Facet, id, parent_id, object_id, object_type, query_status, query_time, query_type, and displayLink. The data rows show various entries, including some with "object_type" values like "official" and "unofficial".

2.3.1.13. Export (exportação)

Depois que você terminar todos os processos de limpeza e filtragem, você pode exportar os textos para a próxima fase. Basta clicar em “Export”, lá no canto superior direito, e escolher a opção “Custom tabular exporter...”.

The screenshot shows the OpenRefine interface with a list of 3695 rows. The 'Export' menu is open, and the 'Custom tabular exporter...' option is highlighted with a red box. Other options like 'Tab-separated value', 'Comma-separated value', 'HTML table', and 'Excel (.xls)' are also visible in the menu.

Antes de salvar, você precisa fazer duas seleções. A primeira é selecionar as colunas que serão exportadas. Clique na aba “Content” e marque as opções de colunas, conforme descrito em cada subcapítulo.

The screenshot shows the 'Content' tab of the 'Custom Tabular Exporter' dialog. The 'message' column is selected and highlighted with a red box. Other columns listed include 'from.name', 'created_time', 'shares.count', 'like_count', 'comment_count', and 'type'. There are also options for selecting column headers and outputting empty rows.

Após isto, clique na aba “Download” e marque a opção “Custom separator \t”. Por fim, clique no botão “Download”. Escolha o nome do arquivo e aonde deseja salvar.

The screenshot shows the 'Download' tab of the 'Custom Tabular Exporter' dialog. The 'Custom separator \t' option is selected and highlighted with a red box. Other options include 'Tab-separated values (TSV)', 'Comma-separated values (CSV)', 'Excel (.xls)', 'Excel in XML (.xlsx)', and 'HTML table'. There are also 'Preview' and 'Download' buttons at the bottom.

2.3.2. Limpeza inicial

Existem filtros e limpezas particulares a cada dado coletado – se foi do Facebook, ou Twitter, ou outro site. Basicamente, serão explicados todos os passos, além de mostrar as colunas exibidas na tela, o que elas representam e – o mais importante – se elas devem ou não serem exportadas, ao final do processo. Também será mencionado se é necessário fazer alguma conversão de coluna e quais filtros, agrupamentos e alterações em categorias/valores devem ser feitos.

Se você coletou os dados no OpenRefine, pode seguir direto para o capítulo 2.3.4. Do contrário, deve executar os passos a seguir, antes de seguir para o capítulo desejado. Isto porque são necessárias algumas limpezas inerentes às consultas obtidas no Facepager.

Etapa 1: Filtrar por texto na coluna “object_type”, conforme o subcapítulo 2.3.1.2.

Etapa 2: Marcar para exibir todas as *choices* da coluna “object_type”, exceto “data” e “unpacked” – se existirem. Para isto, basta seguir o subcapítulo 2.3.1.3.

Etapa 3: Excluir todas as linhas filtradas, conforme subcapítulo 2.3.1.11.

O Facepager, quando busca os dados, gera várias linhas, como já foi dito ao longo do subcapítulo 2.1. As linhas que nos interessam são as que, no “object_type”, estão com os valores de “data” ou “unpacked” (para RSS). Ou seja, as linhas que aparecem com valor “seed”, “error” ou “offcut” podem ser eliminadas.

Etapa 4: Excluir as colunas “level”, “id”, “parent_id”, “object_id”, “object_type”, “query_status”, “query_time” e “query_type”, conforme subcapítulo 2.3.1.12.

2.3.3. Facebook com Facepager

Após a limpeza inicial, conforme subcapítulo 2.3.2, serão exibidas as seguintes colunas:

Coluna	Descrição
from.name	Quem publicou a postagem
message	Conteúdo da postagem
created_time	Data em que foi publicada a postagem
shares.count	Total de compartilhamentos
like_count	Total de curtidas/reações
comment_count	Total de comentários
type	Tipo do conteúdo

Etapa 1: Converter a coluna “created_time”. O código a ser usado está a seguir (destacado em amarelo). Basta copiá-lo e colá-lo, conforme o subcapítulo 2.3.1.5:

```
value.toDate("yyyy-MM-dd'T'hh:mm:ssZ")
```

Etapa 2: Filtrar por texto na coluna “from.name”, conforme subcapítulo 2.3.1.2.

Fazendo isto, você pode, se quiser, definir quais autores de postagens deseja manter ou excluir. A categoria (*choice*) “blank” se refere as pessoas que postaram na página, mas que não são donas dela.

A categoria “blank” é para as demais pessoas que postaram na página. Você pode renomeá-la para “Outros”, “Demais”, entre outras opções, conforme descrito no subcapítulo 2.3.1.9.

Etapa 3: Filtrar por texto na coluna “type”, conforme subcapítulo 2.3.1.2.

Basicamente, vão aparecer as opções de “video” para vídeos, “photo” para fotos, “status” ou “link”. Caso você tenha pego comentários junto, vai aparecer a opção “blank” para comentários.

Você pode renomear a categoria “blank” para “comentários”. Se quiser, pode, inclusive, renomear as demais para o português. Para estas coisas, basta seguir conforme descrito no subcapítulo 2.3.1.9.



Você pode, também, marcar para exibir apenas as postagens que sejam do tipo “vídeo”. Basta seguir o subcapítulo 2.3.1.2.

Com relação à criação de agrupamentos, digamos que você quer deixar apenas dois tipos: postagens e comentários. Ou seja, tudo o que é “link”, “photo”, “video” e “status” seja do tipo “postagens”. Simples: Basta renomear todos para o mesmo nome, como explicado no subcapítulo 2.3.1.8. Após o processo, vão aparecer apenas dois valores em “type”, sendo “comentários” e “postagens”.

Etapa 4: Exportar os dados, conforme subcapítulo 2.3.1.13. A única coluna a ser marcada é a “message”.

2.3.4. Facebook com Data Miner

Visto que não há limpeza inicial , serão exibidas as seguintes colunas:

Coluna	Descrição
Autor	Quem publicou a postagem
Conteúdo (do autor)	Conteúdo da postagem
Autor do conteúdo compartilhado	A postagem é o compartilhamento de outra postagem feita por outra pessoa
Conteúdo compartilhado	Conteúdo dessa postagem que foi compartilhada
Data	Data em que foi publicada a postagem
URL	Endereço para acessar a postagem

Etapa 1: Converter a coluna “Data”, conforme subcapítulo 2.3.1.5. Os passos necessários para converter são mais complexos, pois é necessário transformá-la em um padrão único. Algumas postagens contêm “Ontem”, “2h”, “12h”, e é necessário corrigir isto.

Sendo assim, serão usados quatro códigos, explicados a seguir (destacados em amarelo). Basta copiar e colar, um de cada vez. Leia primeiro todo o texto desta etapa e, então, copie, cole e altere os códigos.

Primeiro, vamos ajustar as linhas que contém a expressão “Ontem”. Para isto, converta a coluna “Data” usando o código abaixo.

```
replace(value, 'Ontem', 'DATAONTEM')
```

A expressão **DATAONTEM** deve ser substituída pela data na qual você obteve os dados. Por exemplo, se os dados foram obtidos no Data Miner em 20 de agosto, a expressão “Ontem” se relaciona a “19 de agosto”. Lembre-se que não se pode usar o “º” em 1º.

```
replace(value, 'Ontem', '19 de agosto')
```

Segundo, vamos ajustar as demais linhas que contém quantidade de horas, como “2 h” ou “12 h” – ou seja, que se referem a “Hoje”. Elas devem ser substituídas pelo dia da coleta, assim como acima mostrado.

```
if(indexOf(value, ' de ') == -1, 'DATAHOJE', value)
```

A expressão **DATAHOJE** ser substituída pela data na qual você obteve os dados. Por exemplo, se os dados foram obtidos no Data Miner em 20 de agosto, a expressão deve ser substituída por “20 de agosto”. Lembrando que não se pode usar o “º” em 1º.

```
if(indexOf(value, ' de ') == -1, '20 de agosto', value)
```

Terceiro, vamos remover a parte de horário, para deixar apenas a data. Em vários testes, mostrou-se melhor trabalhar sem a hora do que com a mesma. Para isto, basta usar o código a seguir:

```
if(indexOf(value, ' às ') == -1,
    if(value.split(" ").length() > 3,
        if(indexOf(value.split(" ")[4], '201') == -1,
            (value.split(" ")[0] + '-' + value.split(" ")[2] + '-' + '2019'),
            (value.split(" ")[0] + '-' + value.split(" ")[2]
                + '-' + value.split(" ")[4]))
        ),
        (value.split(" ")[0] + '-' + value.split(" ")[2] + '-' + '2019')
    ),
    substring(value,0,indexOf(value,' de '))
    + '-' + substring(value,indexOf(value,' de '))
    + 4,indexOf(value,' às ')) + '-' + '2019'
)
```

Quarto, realizar a última e definitiva conversão, usando a seguinte expressão:

```
value.toDate("dd-MMMM-yyyy")
```

Etapa 3: Filtrar por texto na coluna “Autor”, conforme subcapítulo 2.3.1.2.

É importante agrupar comentários ou postagens feitas pelo mesmo autor, mas que, por estarem marcadas com outras pessoas, aparecem como “Fulano está com...”, mas o autor da postagem é sempre o mesmo: “Fulano”.

Etapa 4: Remover datas com poucas ou nenhuma postagem, conforme subcapítulo 2.3.1.6.

É possível excluir os períodos que possuem poucos ou nenhum registro.

Etapa 5: Exportar os dados, conforme subcapítulo 2.3.1.13. As colunas a serem marcadas são, para comentários, a coluna “Conteúdo”, e para postagens, as colunas “Conteúdo do autor” e “Conteúdo compartilhado”.

2.3.5. Twitter com Facepager

Após a limpeza inicial, conforme subcapítulo 2.3.2, serão exibidas as seguintes colunas:

Coluna	Descrição
user.name	Quem publicou o <i>tweet</i>
user.screen_name	Quem publicou o <i>tweet</i> (nome do usuário)
text	Conteúdo do <i>tweet</i>
created_at	Data em que foi publicada o <i>tweet</i>
entities.hashtags	Conjunto de <i>hashtags</i> usado no <i>tweet</i> .
is_quote_status	Se o <i>tweet</i> é uma menção à outro.
favorite_count	Total de favoritos
retweeted	Se o <i>tweet</i> foi <i>retweetado</i> .
in_reply_to_screen_name	Se o <i>tweet</i> é uma resposta a outro, quem é o usuário que escreveu este outro <i>tweet</i> .
retweet_count	Total de <i>retweets</i> .
user.location	Localização do <i>tweet</i> .

Etapa 1: Converter a coluna “created_at”. O código a ser usado está a seguir (destacado em amarelo). Basta copiá-lo e colá-lo, conforme o subcapítulo 2.3.1.5:

```
value.toDate("EEE MMM dd hh:mm:ss Z yyyy")
```

Etapa 2: Exportar os dados, conforme subcapítulo 2.3.1.13. A única coluna a ser marcada é a “text”.

2.3.6. YouTube com Facepager

Após a limpeza inicial, conforme subcapítulo 2.3.2, serão exibidas as seguintes colunas:

Coluna	Descrição
--------	-----------

snippet.title	Título do vídeo
snippet.channelId	Código (ID) do canal onde foi publicado o vídeo.
snippet.publishedAt	Data em que foi publicado o vídeo.
snippet.channelTitle	Nome do canal onde foi publicado o vídeo.
snippet.description	Descrição do vídeo
id.kind	Tipo do conteúdo: <i>video</i> , <i>channel</i> ou <i>playlist</i> .
id.videoId	Código (ID) do vídeo.

Etapa 1: Converter a coluna “snippet.publishedAt”. O código a ser usado está a seguir (destacado em amarelo). Basta copiá-lo e colá-lo, conforme o subcapítulo 2.3.1.5:

```
value.toDate("yyyy-MM-dd'T'hh:mm:ssZ")
```

Etapa 2: Exportar os dados, conforme subcapítulo 2.3.1.13. As colunas a serem marcadas são “snippet.title” e “snippet.description”.

2.3.7. RSS com Facepager

Após a limpeza inicial, conforme subcapítulo 2.3.2, serão exibidas as seguintes colunas:

Coluna	Descrição
title	Título da notícia.
description	Descrição da notícia.
pubDate	Data em que foi publicada a notícia.
content	Texto resumido da notícia.

Etapa 1: Converter a coluna “pubDate”. O código a ser usado está a seguir (destacado em amarelo). Basta copiá-lo e colá-lo, conforme o subcapítulo 2.3.1.5:

```
value.toDate("yyyy-mm-dd hh:mm:ss")
```

Etapa 2: Exportar os dados, conforme subcapítulo 2.3.1.13. As colunas a serem marcadas são “title” e “content”.

2.3.8. CSE com Facepager

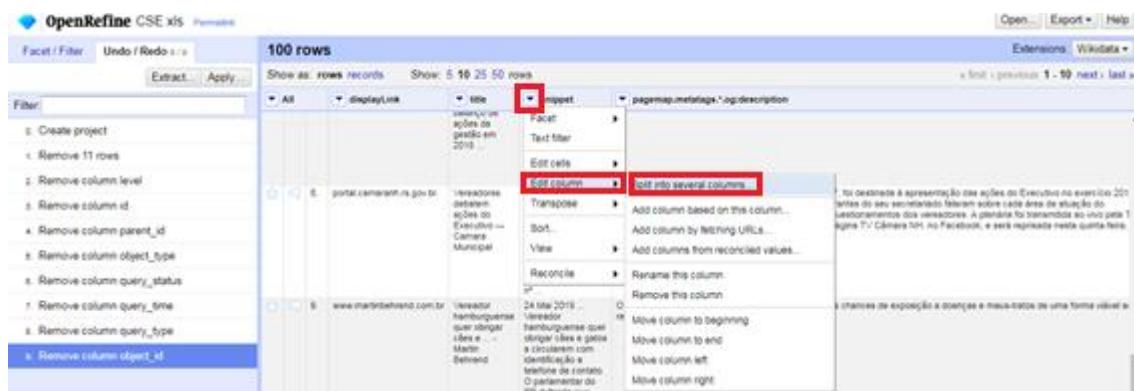
Após a limpeza inicial, conforme subcapítulo 2.3.2, serão exibidas as seguintes colunas:

Coluna	Descrição
displayLink	Site aonde foi publicado o conteúdo.
title	Título do conteúdo
snippet*	Linha que contém a data de publicação e trecho do conteúdo.
pagemap.metatags.*.og:description	Texto ou parte do texto do conteúdo.

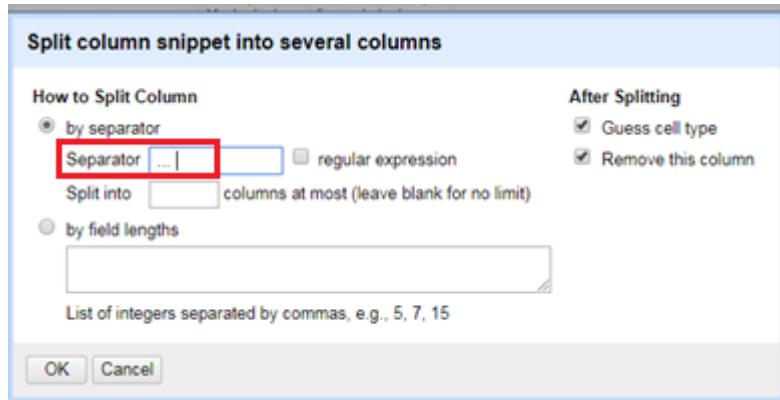
Etapa 1: Divisão da coluna

Diferente dos demais, é necessário realizar alguns procedimentos para conversão. Isto porque, a coluna original “snippet” contém junto a data e o texto. Então, antes mesmo de converter, precisamos quebrar essa coluna.

O processo funciona da seguinte forma: Primeiro, é necessário dividir a coluna. Basta clicar na seta à esquerda da coluna “snippet”, ir em “Edit column” e clicar em “Split into several columns...”.



Preencha o campo “Separator” com reticências (“...”), incluindo um espaço antes e outro depois. Após isto, basta clicar em OK.



Após isto, vão aparecer duas colunas, sendo “snippet1”, com a data, e “snippet2”, com o texto. Às vezes, pode surgir uma terceira coluna (“snippet3”), com a continuação do texto da coluna “snippet2”.

Etapa 2: Converter a coluna “snippet1”. O código a ser usado está a seguir (destacado em amarelo). Basta copiá-lo e colá-lo, conforme o subcapítulo 2.3.1.5:

Assim como acontece com o Data Miner, descrito no subcapítulo 2.3.4, é necessário arrumar as datas inválidas. Algumas postagens contêm “há 2 horas”, “há 1 dia”.

Primeiro: Vamos converter a coluna “snippet1”, usando o código a seguir.

```
if(indexOf(value,'horas') == -1,
    if(indexOf(value,'dia') != -1,
        ('DATAHOJE'),
        (value)
    ),
    'DATAONTEM'
)
```

As expressões **DATAHOJE** e **DATAONTEM** devem ser substituídas pelas datas corretas. Por exemplo, se os dados foram obtidos em 20 de agosto, **DATAHOJE** deve ser “20 ago 2019” e **DATAONTEM** deve ser “19 ago 2019”:

```
if(indexOf(value,'horas') == -1,  
    if(indexOf(value,'dia') != -1,  
        ('20 ago 2019'),  
        (value)  
    ),  
    '19 ago 2019'  
)
```

Segundo: Devemos preparar a coluna “snippet1”, convertendo com o código:

```
replace(value, ' ', '-')
```

Terceiro: Converta a coluna com a expressão:

```
value.toDate("dd-MMM-yyyy")
```

Etapa 3: Exportar os dados, conforme subcapítulo 2.3.1.13. As colunas a serem marcadas são “title”, “pagemap.metatags.*.og:description”, bem como todas as colunas “snippet”, com exceção da “snippet1”, que contém a data.

2.4. EXTRAÇÃO DAS INFORMAÇÕES

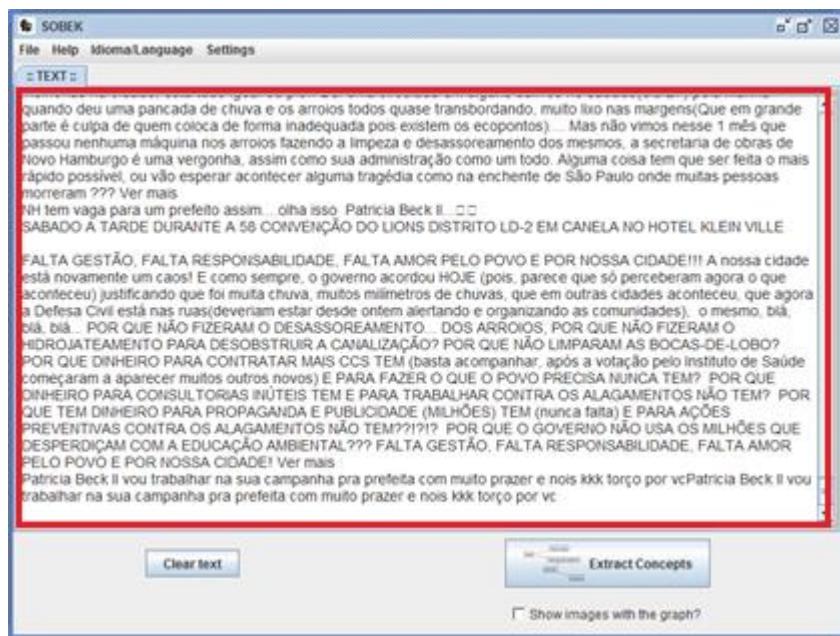
Após exportados os arquivos de texto, eles podem ser importados no SOBEK e no RAnalyzer.

2.4.1. SOBEK

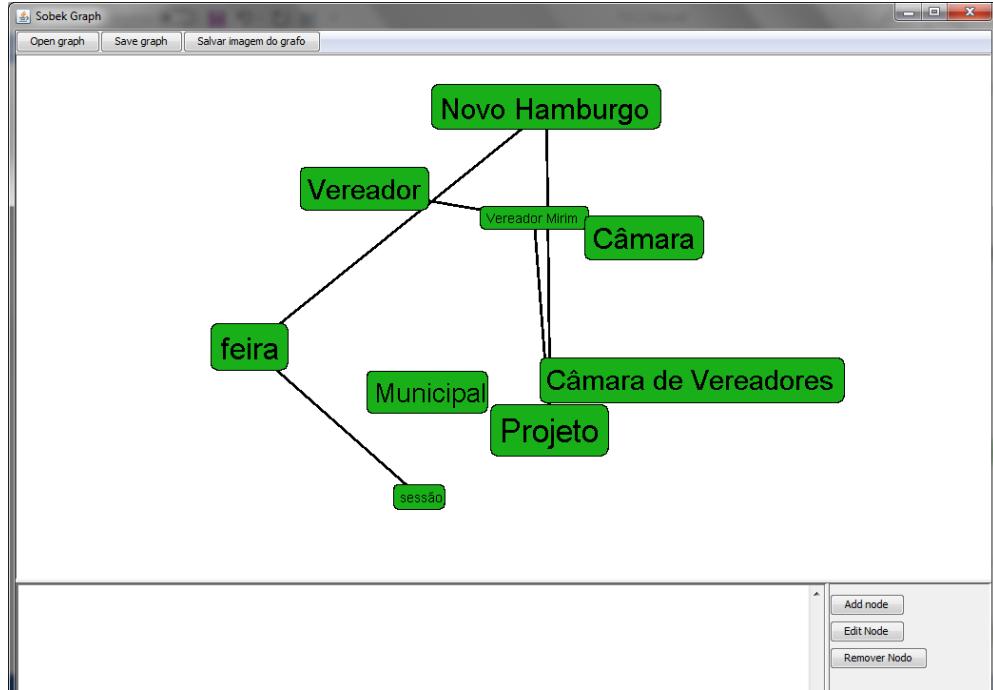
Para abrir o programa, basta ir ao Menu Iniciar e clicar em SOBEK. Após alguns instantes, vai abrir a janela, pronta para o uso.

Embora a ferramenta possua importação de texto, o melhor que recomendamos é você copiar o conteúdo do arquivo desejado e colar dentro do SOBEK. Abra um dos arquivos que você exportou no OpenRefine – como é arquivo texto, provavelmente abrirá no Bloco de Notas do Windows. Geralmente, os arquivos salvos do OpenRefine ficam na pasta de Downloads.

Selecione o texto, copie-o e cole dentro da caixa de texto do SOBEK, usando Ctrl + V. Dependendo do tamanho do texto colado, ele demora um certo tempo – de segundos a, até mesmo, minutos. Então, o texto aparecerá:



Então, basta clicar no botão “Extract concepts”. Depois de um tempo, vai ser carregada uma janela com grafos de palavras. Cuide porque, às vezes, ela aparece atrás da janela de texto. Aí, é só clicar na outra janela.



Alguns detalhes para você saber:

1. Quanto maior a palavra, significa que mais vezes ela aparece no texto.
2. Ao clicar em cima da palavra, aparece, abaixo da janela, uma listagem de ocorrências.

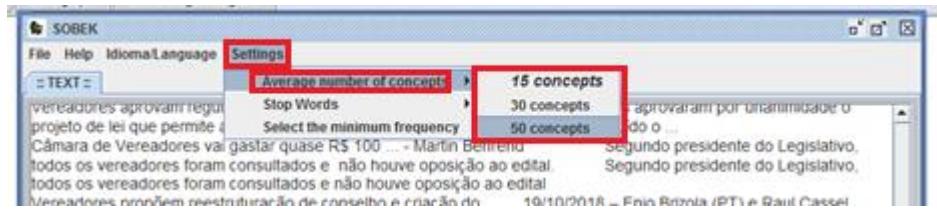


É possível movimentar os grafos, de forma a separar ou aproximar os retângulos com as palavras/expressões.

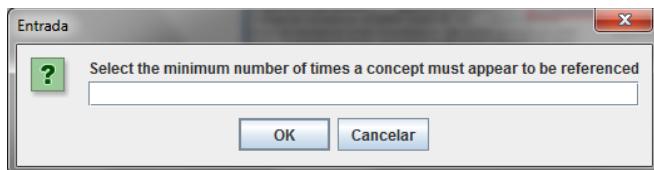
Também é possível salvar o grafo em formato de imagem ou em formato de grafo – para visualização em softwares como *Gephi*. É só clicar nos botões que estão acima do grafo.



Além disto, se você quiser aumentar a quantidade de palavras/expressões exibidas no grafo, basta ir à janela principal do SOBEK, clicar no menu “Settings”, ir em “Average number of concepts” e altere a opção para 15, 30 ou 50 palavras/expressões.



Outra opção que pode ser alterada é a de frequência mínima de palavras. Ou seja, definir que as palavras, para serem exibidas no grafo, devem aparecer, no mínimo, um determinado número de vezes.



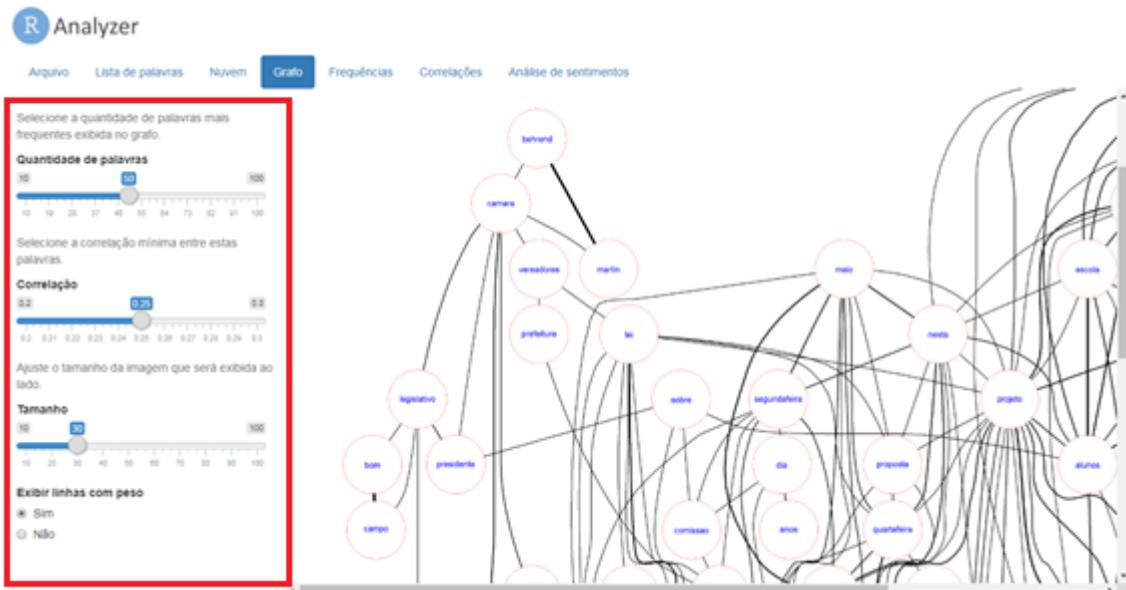
2.4.2. RAnalyzer

Para abrir o programa, basta ir ao Menu Iniciar e clicar em RAnalyzer. Na janela que abrir, basta dar dois cliques no arquivo “ranalyzer”. Vai abrir uma janela e, em seguida, o Google Chrome exibindo a ferramenta, pronta para o uso.

Basicamente, a ferramenta possui uma barra de opções (abas), onde cada opção direciona para um determinado recurso, apresentando a análise e extração de informação do texto.

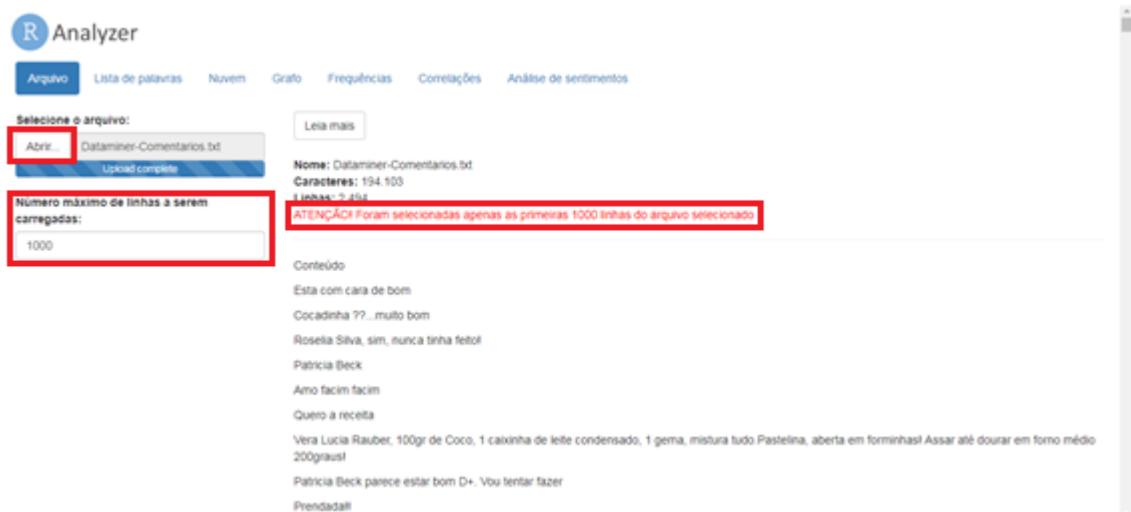


Além disto, à esquerda de cada aba, há algumas opções de filtro, que alteram as informações exibidas no gráfico, tabela ou outro.



Todo o funcionamento do programa ocorre por meio da primeira aba - a “Arquivo”. Ela serve para carregar o arquivo a ser processado e analisado. Quando você carrega o arquivo, o RAnalyzer executa alguns procedimentos, como remoção de acentos e *stopwords*. Embora o segundo passo seja comum, o primeiro ocorre devido à dificuldade em processar alguns caracteres contidos no texto.

Basta clicar no botão “Abrir” e esperar o arquivo ser carregado - diferente do SOBEK, a importação de texto funciona perfeitamente. Lembre-se que o arquivo deve ser do formato TXT. Você pode escolher qualquer arquivo de texto, mas recomendamos, claro, que você escolha um dos arquivos exportados no OpenRefine.



É possível definir o número máximo de linhas que será carregado, com o objetivo de não prejudicar o processamento. Ou seja, se você tiver um arquivo com

10 mil linhas, pode solicitar que sejam carregadas apenas as mil primeiras, sendo que isso é exibido em vermelho.

Após o carregamento, são exibidas as informações básicas do arquivo, como nome, quantidade de linhas (parágrafos), quantidade de palavras e o texto na íntegra. Quando isso ocorrer, já é possível ver as outras abas.

A primeira aba é a “**Lista de Palavras**”. Ela exibe a lista completa de palavras, com a sua respectiva frequência, ou seja, quantas vezes apareceu no texto.

Palavra	Frequência
vereadores	113
camara	77
vereador	64
projeto	55
novo	50
hamburgo	44
municipal	32
sessao	27
minim	27

Nessa tela você pode:

- filtrar as palavras, conforme você digita em “Pesquisar”;
- ordenar as colunas por palavra ou frequência em ordem crescente ou decrescente, conforme clica no título da coluna;

- estabelecer a frequência mínima, movendo a barra à esquerda – lembrando que ao selecionar o valor mínimo de 1, todas as palavras são exibidas.

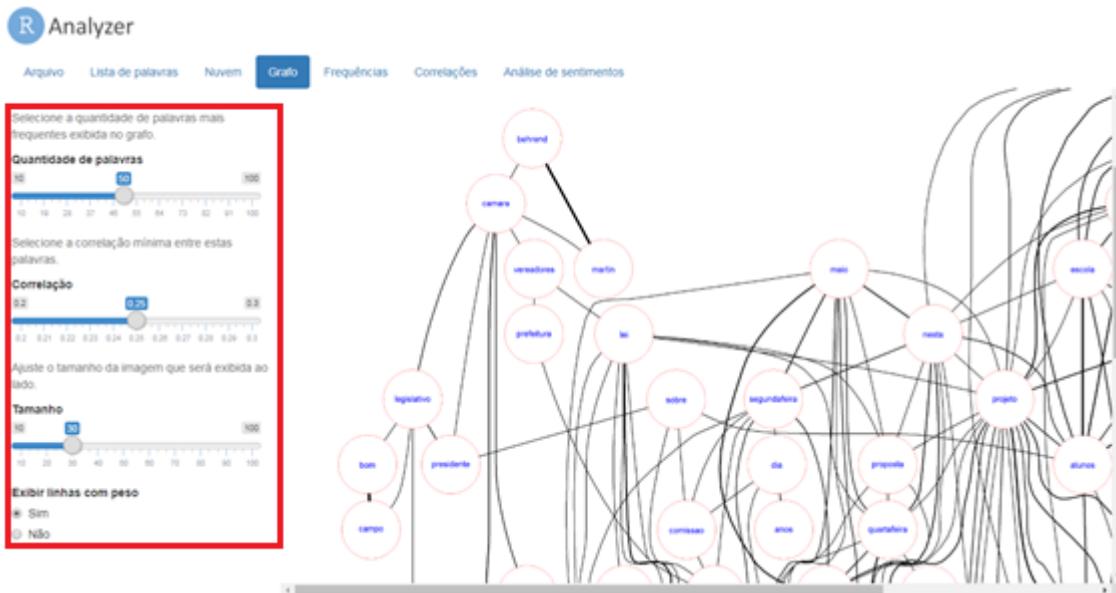
A aba seguinte é “**Nuvem**”, que exibe uma nuvem de palavras. Basicamente, ela permite ver o grau de frequência das palavras no texto de forma proporcional – ou seja, quanto maior o tamanho da palavra, maior a sua frequência.



Nessa tela você pode:

- ver quantas vezes uma palavra apareceu, ao passar o mouse em cima dela;
 - ajustar a frequência de aparição das palavras no texto, em uma variação entre 0.5 e 10, sendo que quanto maior o valor, menor é a frequência máxima considerada – ou seja, quando maior, mais vezes as palavras devem aparecer no texto e, consequentemente, menos palavras aparecem na nuvem.

Na aba “**Grafo**”, é possível visualizar a relação entre as palavras, de forma semelhante ao SOBEK – embora sem o mesmo nível de interatividade.



Nessa tela você pode:

- escolher a quantidade de palavras exibida no grafo, selecionando a exibição de 10 a 100 palavras, dentre as mais frequentes;
- definir o grau mínimo de correlação entre estas palavras, em uma escala entre 0.2 e 0.3, sendo que quanto menor o valor, mais palavras podem aparecer;
- definir o tamanho da imagem que será exibida na tela;
- exibir ou não do peso da correlação, sendo que, quanto maior a correlação entre duas palavras, mais espessa é a linha;
- fazer o *download* da imagem, em alta resolução, para melhor análise, tendo em vista que o Ranalyzer não possui um mecanismo prático para visualizar o grafo, tal qual no SOBEK.

A aba “**Frequências**” exibe, de forma simples, a frequência das palavras por meio de um gráfico de barras (histograma)



Nessa tela você pode:

- visualizar a variação de frequência das palavras, sendo que quanto mais claro o azul, maior é a frequência;
- selecionar o número máximo de palavras a serem exibidas – entre 5 e 30;
- definir a frequência de ocorrência das palavras, da mesma forma que ocorre na aba “Grafo”.

Outro histograma é disponibilizado na aba “Correlações”, mas com uma funcionalidade muito interessante: detalhar correlações entre palavras. A partir de uma palavra selecionada, dentre as que estão no texto, são exibidas as palavras que mais se relacionam com esta – relação expressa em porcentagem.



Nessa tela você pode:

- definir a exibição entre 5 e 30 palavras;
- ajustar a correlação mínima para exibição.

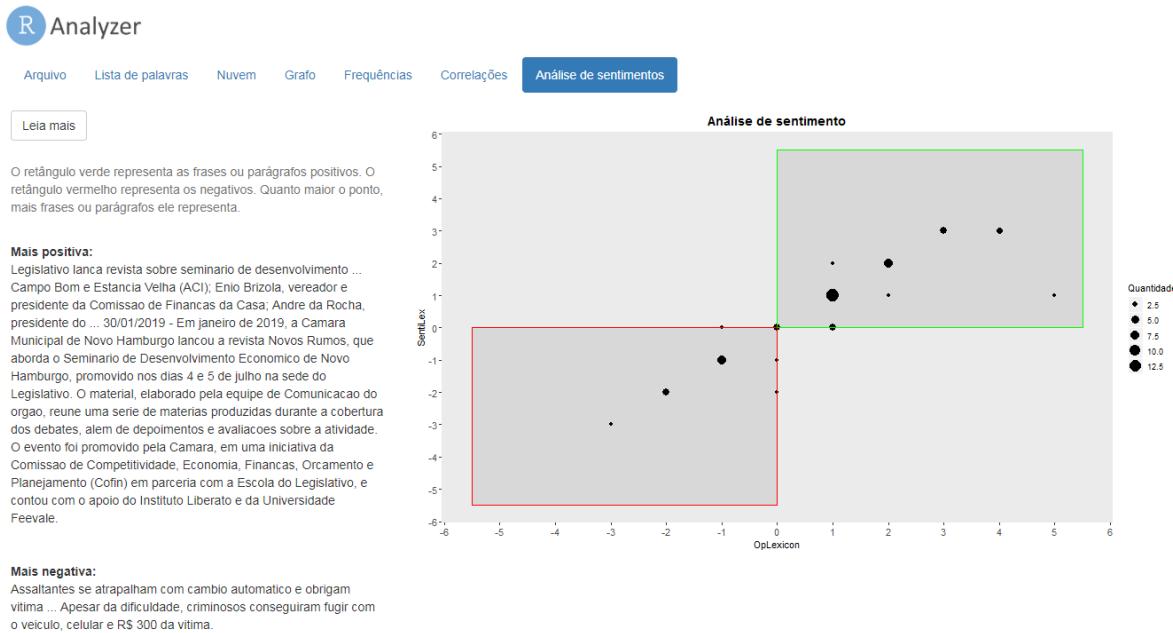
Cabe salientar que a quantidade de palavras exibidas nestas duas abas é influenciada pela correlação mínima. Ou seja, mesmo que na barra “Quantidade de palavras” esteja selecionada a quantidade de 30 palavras, se a “Correlação mínima” escolhida for muito alta, há a possibilidade de não aparecerem as 30 no gráfico.

Para enriquecer a ferramenta, há uma aba de **“Análise de Sentimentos”**. O funcionamento desta é relativamente simples. Inicialmente o texto puro, sem modificações, é passado por uma análise em dois dicionários léxicos: o OpLexicon e o SentiLex. Eles têm como objetivo analisar cada palavra de uma frase ou parágrafo e indicar se a mesma é positiva, neutra ou negativa.

Após isto, o sistema faz um cálculo a partir dos pontos atribuídos para, então, chegar-se à conclusão se o conjunto de frases ou parágrafos do texto é negativo ou positivo.

Essa conclusão é apresentada ao usuário por meio de um gráfico de distribuição, onde o eixo X (na horizontal) representa o dicionário SentiLex e o eixo Y (na vertical) representa o dicionário OpLexicon. Isso significa que, com base no SentiLex, quanto mais à direita, mais positivo o texto é, e quanto mais à esquerda, mais negativo é. Já com base no OpLexicon, quanto mais para cima, mais positivo o texto é, e quanto mais pra baixo, mais negativo é.

Para ficar mais claro, há um retângulo destacado em verde, indicando frases ou parágrafos que são positivos para ambos os dicionários. Por sua vez, há um retângulo destacado em vermelho, indicado os que são negativos para ambos. Além disto, quanto maior o ponto, mais frases ou parágrafos ele representa.



Por exemplo, na imagem acima, há mais pontos no retângulo verde que no vermelho. Além disto, os pontos do retângulo verde são predominantemente maiores do que os do retângulo vermelho. Isto permite concluir que há predominantemente mais textos positivos do que negativos.

Podem aparecer pontos nas áreas não marcadas em verde e vermelho. Isso indica que são frases ou parágrafos que são categorizados como positivo para um dicionário, mas negativo para o outro.

Para completar a tela, foi adicionada uma ferramenta que auxilia nesta análise de sentimentos: é exibida a frase ou parágrafo com pontuação mais positiva e a com mais negativa. Tais escolhas levam em consideração os dois dicionários analisados.

Para auxiliar no uso do RAnalyzer, cada componente possui um texto resumido de ajuda, acima do mesmo. Basta clicar em “Leia mais”

