



MBA em Big Data, Business Analytics e Gestão de Negócios

Disciplina:

Técnicas Avançadas de Captura e Tratamento de Dados

Captura de dados sobre registro de patentes nos Estados Unidos para o período 1976 a 2015

Cristiano de Lima Logrado

Brasília / DF

Agosto/2021

SUMÁRIO

1	Objetivo	3
2	Análise do repositório de arquivos.....	3
3	Código para captura automática	6
4	Análise preliminar da evolução do número de patentes	7
5	conclusões	8

1 OBJETIVO

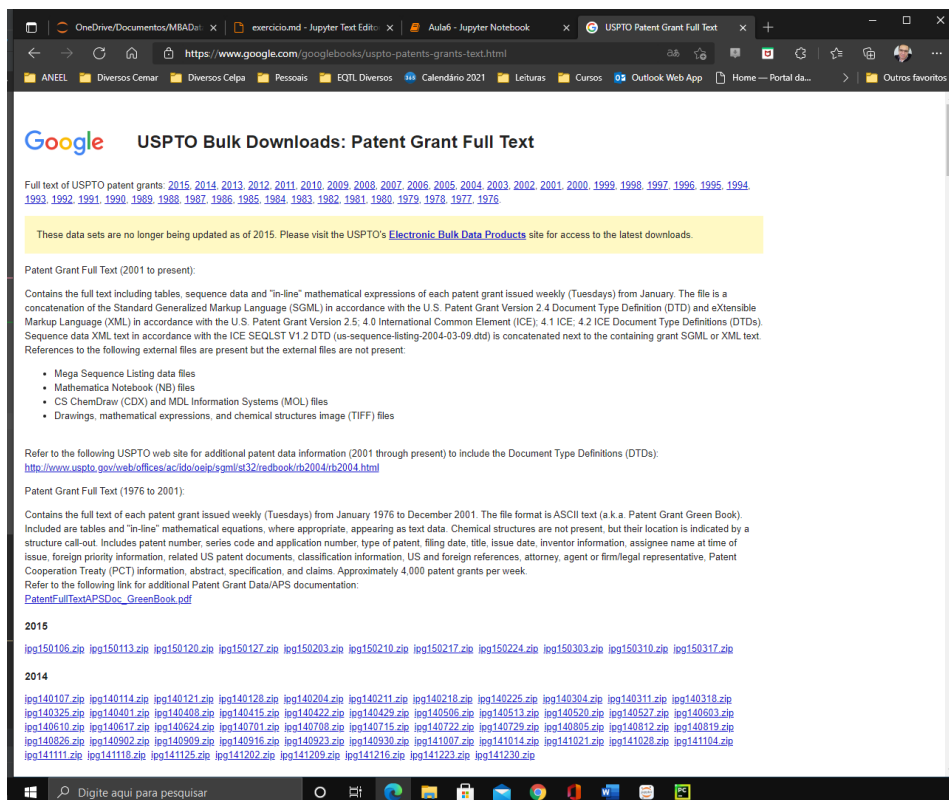
No endereço <https://www.google.com/googlebooks/uspto-patents-grants-text.html> são disponibilizados textos completos, incluindo tabelas e expressões, de patentes registradas nos Estados Unidos no período de 1976 a 2015 (parcial). Cada arquivo contempla dados semanais, ou seja, são, em média, 52 arquivos para cada ano, regra que não se aplica a 2001, como será visto à frente.

O presente trabalho tem por objetivos *i)* capturar as informações, por meio do download automatizado dos arquivos, *ii)* avaliar, preliminarmente, a evolução da quantidade de patentes registradas no tempo, mediante o tratamento dos dados capturados.

Para a automatização, será utilizado um conjunto de códigos elaborados na linguagem Python, versão 3.9.6, no IDE PyCharm. Para controle de versão, será usado o sistema GitHub (<https://github.com/cristianologrado/ExercicioPatentes>). Além do PyCharm, foi utilizado o Jupyter Notebook como ferramenta auxiliar, para testes de desenvolvimento.

2 ANÁLISE DO REPOSITÓRIO DE ARQUIVOS

Como sinalizado anteriormente, os arquivos constam do endereço <https://www.google.com/googlebooks/uspto-patents-grants-text.html>, cuja tela inicial é apresentada na Figura 1.



Fonte: <https://www.google.com/googlebooks/uspto-patents-grants-text.html>

Figura 1 – Tela inicial do repositório de dados (captura de tela)

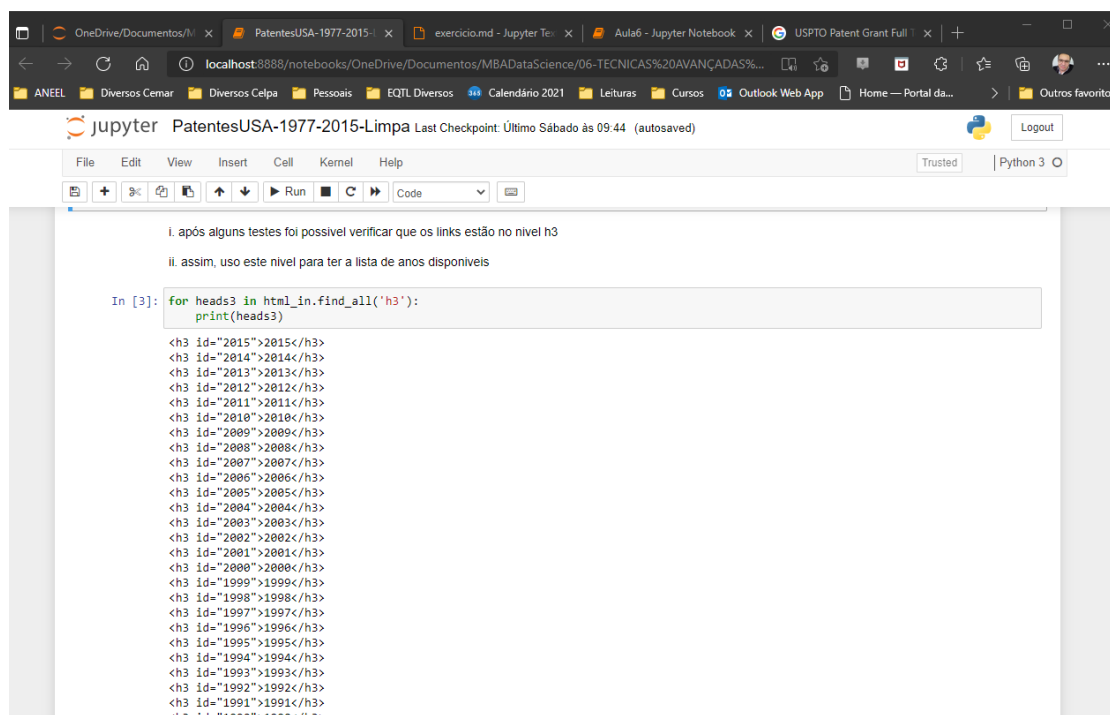
Uma avaliação identificou um repositório com estrutura HTML simples, com a simples apresentação sequencial de textos e links para os arquivos de dados, sendo que alguns pontos serão destacados a seguir.

Quanto à ferramenta de captura, foi utilizada a biblioteca BeautifulSoup, do Python, que permitiu a leitura e captura integral do texto, pelo código abaixo.

```
url_in = 'https://www.google.com/googlebooks/uspto-patents-grants-text.html'
response = urllib.request.urlopen(url_in)
html_in = BS(response)
```

Não foi detectado qualquer tipo de bloqueio ou restrição, o que simplificou o processo de captura do conteúdo HTML do repositório de dados.

Após análise do arquivo, com a leitura direta do código HTML, identificou-se que os arquivos foram agrupados segundo os anos, usando-se estruturas de título H3. Deste modo, a partir da identificação de tais estruturas, foi possível a captura dos anos para os quais havia arquivos de patentes disponibilizados. A Figura 2, abaixo, apresenta um extrato do Notebook usado no exercício exploratório.



```
In [3]: for heads3 in html_in.find_all('h3'):
        print(heads3)

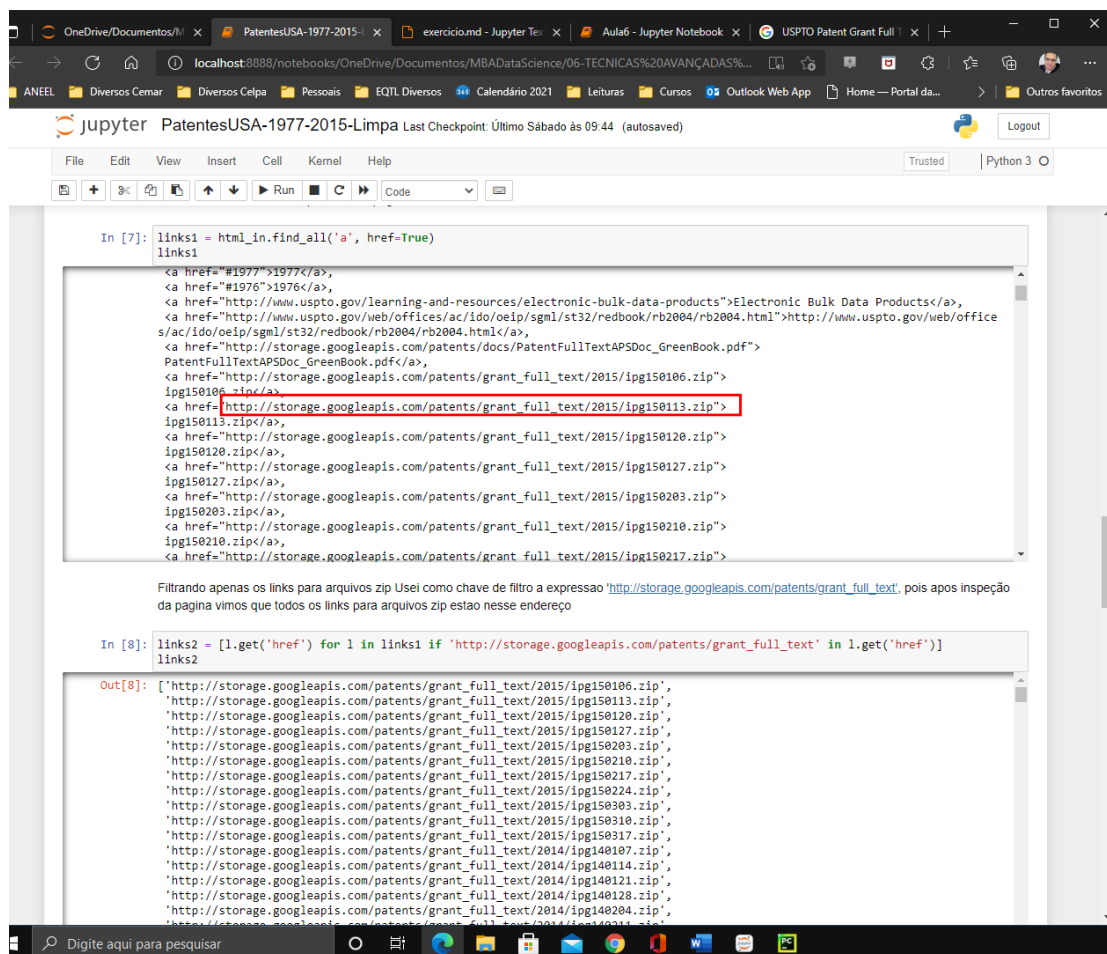
<h3 id="2015">2015</h3>
<h3 id="2014">2014</h3>
<h3 id="2013">2013</h3>
<h3 id="2012">2012</h3>
<h3 id="2011">2011</h3>
<h3 id="2010">2010</h3>
<h3 id="2009">2009</h3>
<h3 id="2008">2008</h3>
<h3 id="2007">2007</h3>
<h3 id="2006">2006</h3>
<h3 id="2005">2005</h3>
<h3 id="2004">2004</h3>
<h3 id="2003">2003</h3>
<h3 id="2002">2002</h3>
<h3 id="2001">2001</h3>
<h3 id="2000">2000</h3>
<h3 id="1999">1999</h3>
<h3 id="1998">1998</h3>
<h3 id="1997">1997</h3>
<h3 id="1996">1996</h3>
<h3 id="1995">1995</h3>
<h3 id="1994">1994</h3>
<h3 id="1993">1993</h3>
<h3 id="1992">1992</h3>
<h3 id="1991">1991</h3>
<h3 id="1990">1990</h3>
```

Fonte: própria

Figura 2 – Identificação dos anos, a partir dos cabeçalhos H3

A sequência do processo exploratório levou a localização dos links para os arquivos de dados de patentes. Todos os arquivos constam do repositório

“http://storage.googleapis.com/patents/grant_full_text”, com diferenciação apenas do ano e do arquivo de dados, como indicado na Figura 3.



```

In [7]: links1 = html_in.find_all('a', href=True)
links1
<a href="#1977">1977</a>,
<a href="#1976">1976</a>,
<a href="http://www.uspto.gov/learning-and-resources/electronic-bulk-data-products">Electronic Bulk Data Products</a>,
<a href="http://www.uspto.gov/web/offices/ac/ido/oeip/sgml/st32/redbook/rb2004/rb2004.html">http://www.uspto.gov/web/offices/ac/ido/oeip/sgml/st32/redbook/rb2004/rb2004.html</a>,
<a href="http://storage.googleapis.com/patents/docs/PatentFullTextAPSDoc_GreenBook.pdf">PatentFullTextAPSDoc_GreenBook.pdf</a>,
<a href="http://storage.googleapis.com/patents/grant_full_text/2015/ipg150106.zip">ipg150106.zip</a>,
<a href="http://storage.googleapis.com/patents/grant_full_text/2015/ipg150113.zip">ipg150113.zip</a>,
<a href="http://storage.googleapis.com/patents/grant_full_text/2015/ipg150120.zip">ipg150120.zip</a>,
<a href="http://storage.googleapis.com/patents/grant_full_text/2015/ipg150127.zip">ipg150127.zip</a>,
<a href="http://storage.googleapis.com/patents/grant_full_text/2015/ipg150203.zip">ipg150203.zip</a>,
<a href="http://storage.googleapis.com/patents/grant_full_text/2015/ipg150210.zip">ipg150210.zip</a>,
<a href="http://storage.googleapis.com/patents/grant_full_text/2015/ipg150217.zip">ipg150217.zip</a>

Filtrando apenas os links para arquivos zip Usei como chave de filtro a expressao 'http://storage.googleapis.com/patents/grant_full_text', pois apos inspeção da pagina vimos que todos os links para arquivos zip estao nesse endereço

In [8]: links2 = [l.get('href') for l in links1 if 'http://storage.googleapis.com/patents/grant_full_text' in l.get('href')]
links2
Out[8]: ['http://storage.googleapis.com/patents/grant_full_text/2015/ipg150106.zip',
'http://storage.googleapis.com/patents/grant_full_text/2015/ipg150113.zip',
'http://storage.googleapis.com/patents/grant_full_text/2015/ipg150120.zip',
'http://storage.googleapis.com/patents/grant_full_text/2015/ipg150127.zip',
'http://storage.googleapis.com/patents/grant_full_text/2015/ipg150203.zip',
'http://storage.googleapis.com/patents/grant_full_text/2015/ipg150210.zip',
'http://storage.googleapis.com/patents/grant_full_text/2015/ipg150217.zip',
'http://storage.googleapis.com/patents/grant_full_text/2015/ipg150224.zip',
'http://storage.googleapis.com/patents/grant_full_text/2015/ipg150303.zip',
'http://storage.googleapis.com/patents/grant_full_text/2015/ipg150310.zip',
'http://storage.googleapis.com/patents/grant_full_text/2015/ipg150317.zip',
'http://storage.googleapis.com/patents/grant_full_text/2014/ipg140107.zip',
'http://storage.googleapis.com/patents/grant_full_text/2014/ipg140114.zip',
'http://storage.googleapis.com/patents/grant_full_text/2014/ipg140121.zip',
'http://storage.googleapis.com/patents/grant_full_text/2014/ipg140128.zip',
'http://storage.googleapis.com/patents/grant_full_text/2014/ipg140204.zip',
'http://storage.googleapis.com/patents/grant_full_text/2014/ipg140211.zip']

```

Fonte: própria

Figura 3 – Identificação dos links para arquivos de dados

Tendo-se em vista a estrutura simplificada da página e o fato de que todos os arquivos estão no mesmo repositório, foi possível, a partir do endereço do repositório, filtrar a totalidade dos links desejados.

Cada link aponta para um arquivo específico, compactado (.zip), o qual contém um arquivo em formato texto, mas com uma diferenciação. Para os anos mais recentes, os anos mais recentes, os registros são armazenados em um formato XML – esta abordagem foi identificada no período 2015 a 2002.

Para o ano de 2001, identificou-se um par de arquivos para cada semana, sendo cada dupla composta por um arquivo TXT e por um arquivo em formato SGM. E para os períodos anteriores (2000 a 1976), tem-se o armazenamento em arquivos TXT tradicionais.

Todos os arquivos são, essencialmente, arquivos textos, mas com estrutura de dados diferenciadas. Todavia, no caso específico, não foi abordado a forma de leitura de cada

arquivo, para uma avaliação completa das patentes, visto que cada um deles demandaria uma avaliação minuciosa com a busca e identificação da codificação de cada formato de dados.

Feita a análise exploratória do conteúdo da página HTML, foi possível a construção de um código, em Python, para a automatização do processo, o qual será explorado a seguir.

3 CÓDIGO PARA CAPTURA AUTOMÁTICA

Para a construção do código automatizado, optou-se pelo IDE PyCharm, com a elaboração de um script para execução sequencial do código.

Em uma abordagem estruturada, foram construídas 4 funções, a saber:

captura_anos : está, a partir dos cabeçalhos H3, identifica os anos para os quais há arquivos de patentes disponíveis

contar_patentes : após a identificação dos anos, é realizada uma contagem dos links de arquivos disponíveis, mapeando-se a quantidade para cada um dos anos

baixararquivos : esta função, a partir dos links recebidos, faz o download dos arquivos para o diretório indicado. De forma a se otimizar o uso de recursos, a função verifica, para cada arquivos, se o mesmo já existe no diretório, e baixa apenas os arquivos novos.

capturardetalhes: por fim, foi elaborada uma função que, para cada link, captura o tamanho (em bytes) do arquivo associado, salvando a informação em um arquivo indicado. Como no caso da função anterior, a captura, que exige a realização de conexão com o site, só é realizada se o arquivo de salvamento não estiver disponível no diretório de trabalho. A função faz a leitura dos detalhes, salva os dados em arquivo, e retorna um Dataframe com toda a informação.

O código das funções consta do arquivo Patentes.py no repositório Github indicado no início do texto. A execução sequencial das funções construídas consta do arquivo main.py, disponível no mesmo repositório.

A sequência de execução é simples, e contempla os seguintes passos:

- i. Realização da conexão com a página (<https://www.google.com/googlebooks/uspto-patents-grants-text.html>) e captura do código HTML completo
- ii. Identificação dos anos para os quais há arquivos de patentes, mediante a função específica (*captura_anos*)

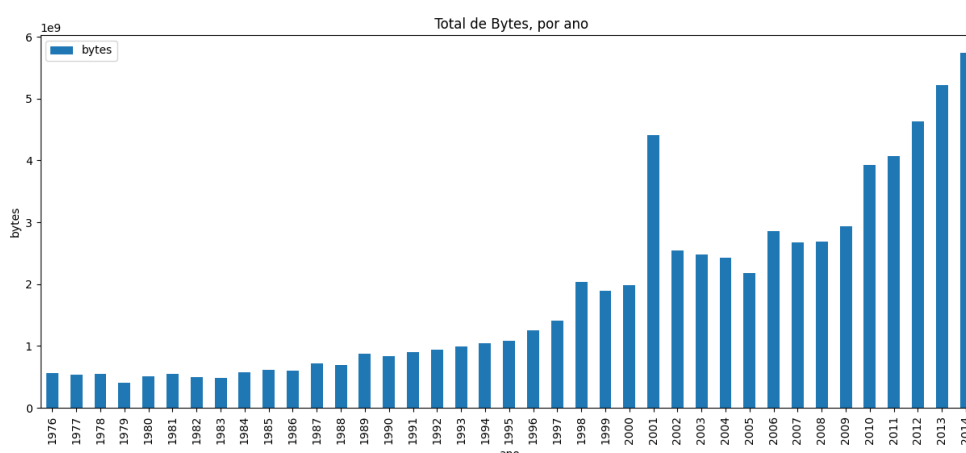
- iii. Seleção de todos os links da página (links1), e filtro dos links vinculados a arquivos .zip (links2)
- iv. Contagem das patentes, ou de forma mais específica, contagem do total de arquivos para cada ano, a partir do chamado da função `contar_patentes`
- v. Download dos arquivos, mediante uso da função `baixararquivos`. Aqui, vale frisar que são cerca de 2.100 arquivos, mas o código foi ajustado para baixar apenas 5, visto que tem fins didáticos. Para download do conjunto completo, deve-se apenas passar o conjunto integral de links, e não apenas os 5 primeiros, como no código exemplo disponibilizado.
- vi. A captura dos detalhes (tamanho dos arquivos de dados) de cada ano é feita a seguir, via função `capturardetalhes`; e
- vii. Por fim, para fins de análise simplificada, o total de dados (bytes) é agrupado por ano, e um gráfico com a evolução temporal dos dados é apresentado.

Como indicado, os arquivos estão disponíveis no repositório <https://github.com/cristianologrado/ExercicioPatentes>.

4 ANÁLISE PRELIMINAR DA EVOLUÇÃO DO NÚMERO DE PATENTES

Considerando os objetivos do trabalho, foi realizada apenas uma análise preliminar do conteúdo da página, partindo-se da premissa de que o total de bytes armazenados a cada ano seja um indicador adequado da quantidade de patentes registradas no ano. Esta abordagem simplificada foi adotada em função da dificuldade associada à realização de um trabalho completo (download e análise de cada arquivo, registrando-se cada patente), que extrapola o objetivo do trabalho em tela.

O gráfico da Figura 4, apresenta o total de bytes para cada ano.



Fonte: própria

Figura 4 – Total de bytes (dados) para cada ano

O gráfico da Figura 4 mostra um período de registro contínuo, com volumes quase uniformes de dados de 1976 a 1987, e a partir deste ano há uma aceleração contínua, com volume cada vez maior de dados.

Para o ano de 2001, há um pico, mas como identificado durante a análise exploratória do repositório, para este ano há uma duplicação, com disponibilização de arquivos em 2 formatos distintos. E, de fato, uma inspeção visual da Figura 4 mostra que um ajuste de 50% no volume de 2001 acomodaria a barra deste ano na tendência histórica. Após um período de estabilidade (entre 2001 e 2009) observa-se nova fase de aceleração, que se manteve até 2014. Os dados de 2015 foram excluídos do gráfico, visto que a página não traz dados integrais para este ano.

5 CONCLUSÕES

Considerando-se o acima exposto é possível concluir que:

- i. O repositório disponibilizado pelo Google para as patentes tem estrutura HTML simples, o que facilita significativamente a automatização do processo de download dos arquivos, para o período em tela (1976 a 2015)
- ii. Todavia, observa-se o armazenamento das patentes em pelo menos três formatos de arquivos distintos, cada um com um padrão próprio de dados. A análise de cada tipo está além do objetivo deste trabalho, e não foi realizada.
- iii. A ferramenta Jupyter Notebook mostrou-se adequada para a realização da análise exploratória da página, visando a identificação de parâmetros que permitissem a captura automática dos dados.
- iv. O código Python desenvolvido, após a análise exploratória, mostrou-se adequado, realizado a captura dos dados e download dos arquivos de forma satisfatória.
- v. Uma análise preliminar dos dados, adotando-se a quantidade de dados (bytes), de cada ano, como um indicado do total de patentes registradas, mostra que há um período de estabilidade até 1987, quando se inicia uma fase de aceleração, que perdura até 2001. Nova fase de estabilidade é observada de 2001 a 2009, quando se tem nova elevação do ritmo de registro de patentes.