|  |  |
| --- | --- |
| Curso: | Jurimetria |
| Disciplina: | Raspagem (webscraping) de tribunais de justiça e administrativos |

**Apresentação**



Fonte: [Imagem criada por meio de inteligência artificial.](https://www.bing.com/images/create/web-scraping-of-lawsuits-from-courts-websites/6433e1e7d5ac4cd99690e1b80976eef9?id=TSM4sKTMDuRzRhrfL6ftbQ%3d%3d&view=detailv2&idpp=genimg&FORM=GCRIDP&mode=overlay)

|  |  |
| --- | --- |
| **Título do vídeo** | **ID** |
| Apresentação da disciplina raspagem (web scraping) de tribunais de justiça e administrativos | A ser informado pelo estúdio |

Prezado(a) aluno(a),

A disciplina raspagem (web scraping) de tribunais de justiça e administrativos é parte do programa Jurimetria – Ciência de Dados Aplicada ao Direito.

Esta disciplina busca oferecer autonomia à participante para ela (ou ele) mesma possa coletar e estruturar dados processuais de vários tribunais do Brasil, tanto judiciais quanto administrativos.

Por meio dessa disciplina, você irá entender como funcionam os navegadores quando digitamos o endereço de uma página web. Isso permitirá reproduzir o comportamento do navegador na linguagem de programação R para baixar várias páginas. Além disso, você aprenderá a estruturar os dados coletados de forma eficiente.

A dificuldade de ensinar essa disciplina é que ela conta com pouco material didático disponível, pois ela tem uma característica heurística, ou seja, não há uma fórmula genérica para raspar páginas, apenas uma série de dicas sobre como fazer isso e aprende-se com a prática. Então, praticar muito é a principal dica.

A disciplina está dividida em 4 unidades semanais, a saber:

Unidade 1 – Noções básicas de transferência de dados pela internet.

Unidade 2 – Requisição.

Unidade 3 – Extração.

Unidade 4 – Iteração.\

**Cronograma das Atividades**

**Prezados Professores,** o quadro abaixo é padrão e apresenta uma sugestão de cronograma de atividades.

A pontuação das atividades das unidades e a quantidade de atividades fica a critério da coordenação do curso, desde que a soma dos pontos seja **60**.

A Atividade Final da disciplina (equivalente à prova) valerá **40** pontos.

Importante seguir com a mesma métrica em todas as disciplinas de um mesmo curso, para que os alunos se habituem com a dinâmica das aulas.

Todas as unidades devem ter, ao menos, uma atividade dentre as disponíveis (questionário, fórum ou envio de arquivo).

Toda disciplina deverá ter, ao menos, uma questão de **Envio de arquivo**.

Qualquer dúvida, converse com o seu Coordenador. Bom trabalho!

Prezado(a) aluno(a),

Sugerimos que você execute as tarefas online de acordo com o cronograma de atividades a seguir. Dessa forma, você poderá organizar seus estudos semanalmente e finalizar sua disciplina com tranquilidade e sem atropelos.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Prazo** | **Unidade** | **Atividade** | **Tipo** | **Pontos** | |
| 1ª Semana | Unidade I | Leitura | Conteúdo | - | |
| Atividade 1 | Questionário | 10 | |
| 2ª Semana | Unidade II | Leitura | Conteúdo | - | |
| Atividade 2 | Fórum Avaliativo | 15 | |
| 3ª Semana | Unidade III | Leitura | Conteúdo | - | |
| Atividade 3 | Questionário | 15 | |
| 4ª Semana | Unidade IV | Leitura | Conteúdo | - | |
| Atividade 4 | Envio de Arquivo | 20 |
| Conteúdo Completo | Atividade Final | Questionário | 40 |
| **TOTAL da Disciplina** | | | | **100** | |

A avaliação da disciplina está dividida em duas etapas:

* + **Atividades das Unidades:** atividades propostas na sala de aula online (**ver cronograma acima**), em que o aluno concorre a 60 pontos da nota final da disciplina (máximo que pode obter);
  + **Atividade Final:** questionário disponível na última semana da disciplina, em que o aluno concorre a 40 pontos da nota final da disciplina (máximo que pode obter).

A **nota final** **da disciplina** será a soma da nota das atividades propostas nas salas de aula online, sendo que, para ser aprovado, o aluno precisa ter, no mínimo, 70 pontos.

Desejamos a você bons estudos!

Minicurrículo do Professor/Tutor

|  |  |  |
| --- | --- | --- |
|  | José de Jesus Filho | Graduação e mestrado em direito, doutorado em administração pública e pós-doc em jurimetria. É professor de jurimetria na PUC-SP, na PUC-PR, Unyleya e UniAnchieta, bem como em cursos de difusão na USP de Ribeirão Preto. É assessor do Ministério Público e São Paulo para jurimetria Possui os três conhecimentos essenciais para a jurimetria: direito, aprendizado estatístico e programação. |

Biblioteca Virtual

Informar quais artigos, slides e referências deverão ser inseridos na biblioteca da disciplina:

Slides Unidade 1.pdf

Slides Unidade 2.pdf

Slides Unidade 3.pdf

Slides Unidade 4.pdf

Nesse caso, foram publicados os pdfs dos slides utilizados durante a gravação dos vídeos.

**Unidade 1 –** Noções básicas de transferência de dados pela internet

Objetivos Específicos

Essa unidade tem como objetivo:

* Familiarizar-se com os principais conceitos da internet;
* Ter um primeiro contato com o protocolo http;
* Inspecionar uma página web;
* Oferecer uma visão geral do processo de webscraping;
* Introduzir o conteúdo das demais unidades.

Playlist

**Prezados Professores, atenção!**

Se sua disciplina é de 6 horas de vídeo, você deverá gravar 1h30 de vídeos por unidade. Ou se a sua disciplina é de 8 horas de vídeo, você deverá gravar 2h de vídeos por unidade.

Na sugestão abaixo, propusemos 5 tópicos por unidade, cada tópico com 2 vídeos de 9 minutos, totalizando 1h30 de vídeo por unidade.

Importante! Os vídeos não devem ultrapassar o tempo máximo de 12 minutos.

Pode haver uma oscilação de quantidade de minutos entre as Unidades, contanto que o total mínimo da disciplina seja de 6 horas de vídeos (já editados) ou 8 horas (já editados), a depender do seu contrato.

Reforçamos que os conteúdos trabalhados nos vídeos devem estar em consonância com o material didático escrito e com as avaliações. Para isso, atentar-se aos demais elementos do Roteiro de Sala.

Qualquer dúvida, converse com o seu Coordenador. Bom trabalho!

Assista aos vídeos do Prof. José de Jesus Filho referentes à Introdução à linguagem de programação R, clicando na playlist abaixo:

|  |  |
| --- | --- |
| **Título do vídeo** | **ID** |
| (1.1) Introdução à internet | A ser informado pelo estúdio |
| (1.2) API e protocolo http |  |
| (1.3) Inspeção de uma página web |  |
| (1.4) As três etapas de webscraping |  |

Material Complementar

[Noções básicas de transferência de dados pela internet](https://direito.consudata.com.br/raspagem/introducao/)

Atividades

Para fixar o conhecimento adquirido até aqui, proceda com as atividades propostas para essa unidade.

UNIDADE 1 - QUESTIONÁRIO

**Questão 01 – Alternativas. Apenas um item está correto.**

Qual a diferença entre Internet e WWW (Web)?

1. Internet é um serviço da Web.
2. Web é um serviço da Internet.
3. São sinônimos.
4. Internet é uma rede local, Web é uma rede global.
5. Web é uma rede local, internet é uma rede global.

**COMENTÁRIO**:

A Web é serviço que utiliza a rede mundial de computadores para disponibilizar conteúdo interativo. Ela utiliza o protocolo http e html para disponibilizar arquivos, textos, imagens e vídeos em um navegador.

**Questão 02 – Alternativas. Apenas um item está correto.**

Uma vantagem de API na transmissão de dados é:

1. API não possui vantagem alguma.
2. APIs são dez vezes mais rápidas na transmissão de dados.
3. APIs facilitam a transmissão de dados entre softwares.
4. Os dados são mais corretos.
5. APIs são mais seguras.

**COMENTÁRIO**:

Quando um tribunal ou qualquer órgão público disponibiliza dados por meio de API, esses dados podem ser transmitidos e obtidos sem o uso de uma interface como uma página html. Isso não torna os dados mais corretos ou a operação mais segura e não necessariamente mais rápida, porém, isso quase sempre ocorre.

**Questão 03 – Alternativas. Apenas um item está correto.**

APIs disponibilizam os dados:

a. Desestruturados.

b. Semiestruturados.

c. Estruturados.

d. Desorganizados.

e. Nenhuma das opções acima.

**COMENTÁRIO**:

Um dos propósitos de APIs é disponibilizar os dados de forma estruturada.

**Questão 04 – Alternativas. Apenas um item está correto.**

Que argumentos os tribunais usam para não disponibilizar APIs?

a. APis reduzem a qualidade dos dados.

b. APIs aumentam a qualidade dos dados.

c. APIs são desnecessárias porque as páginas dos tribunais já contêm os dados estruturados.

d. Limitações orçamentárias e de segurança.

e. Nenhuma das opções acima.

**COMENTÁRIO**:

Os principais argumentos contra a criação de APIs utilizados por tribunais são limitações orçamentárias e razões de segurança. No entanto, empresas acabam por criar robôs para acessar os dados, o que limita a disponibilidade e termina por elevar os custos.

Verificando os Objetivos Da Unidade

Verifique se você alcançou os objetivos propostos para essa unidade. Para isso, identifique se você é capaz de:

* Sabe a diferença entre fatores, números e textos.
* Sabe o que são dataframes.
* Sabe criar objetos no R.

**Unidade 2 – Trabalhando com pacotes jurimétricos**

Objetivos Específicos

Essa unidade tem como objetivo:

* Coletar dados do Supremo Tribunal Federal.
* Coletar dados do Superior Tribunal de Justiça.
* Coletar e organizar dados do Tribunal de Justiça de São Paulo;

Playlist

Assista aos vídeos do Prof. José de Jesus Filho referentes ao Trabalho com Pacotes Jurimétricos tratados nessa unidade, clicando na playlist abaixo:

|  |  |
| --- | --- |
| **Título do vídeo** | **ID** |
| (2.1) Coleta e organização de dados do TJSP | A ser informado pelo estúdio |
| (2.2) Coleta de dados do STF |  |
| (2.3) Coleta de dados do STJ |  |

Verificando os Objetivos Da Unidade

Verifique se você alcançou os objetivos propostos para essa unidade. Para isso, identifique se você é capaz de:

* Coletar a jurisprudência do STF;
* Coletar a jurisprudência do STJ;
* Coletar e organizar dados do TJSP;

Caso haja alguma dúvida em algum dos objetivos propostos nessa unidade, reveja os vídeos ou, se a dúvida persistir, entre em contato com seu tutor pelo Fórum de dúvidas, disponível na disciplina.

Material Complementar

[Como baixar dados do TJSP](https://direitoemdados.consudata.com.br/posts/2022-04-29-como-baixar-dados-do-tjsp/)

[Coleta dados do STF](http://stf.consudata.com.br/)

Atividades

Para fixar o conhecimento adquirido até aqui, participe do fórum de discussão e proceda com as atividades propostas para essa unidade.

**FÓRUM DA UNIDADE 2**

Olá, turma!

Para testar nossos conhecimentos, vamos baixar os dados do TJSP sobre infanticídio. Você pode baixar sobre outro tema, mas verifique antes se não serão muitos julgados, o que pode tomar muito tempo.

Para esta atividade, acesse o script com as instruções no link abaixo:

[Coleta jurisprudência do TJSP](https://gist.github.com/jjesusfilho/7b54d2975ce9512a009917f39946c21c)

Perguntas:

Quantas linhas tem o dataframe cjsg?

Quais as colunas do dataframe cjsg?

Qual seria o próximo passo?

Verificando os Objetivos Da Unidade

Verifique se você alcançou os objetivos propostos para essa unidade. Para isso, identifique se você é capaz de:

* Conseguiu baixar dados do STF.
* Conseguiu baixar dados do STJ.
* Realizou efetivamente a atividade de baixar, organizar e salvar os dados do TJSP.

Caso haja alguma dúvida em algum dos objetivos propostos nessa unidade, reveja os vídeos ou, se a dúvida persistir, entre em contato com seu tutor pelo Fórum de dúvidas, disponível na disciplina.

**Unidade 3 – Tidyverse**

Objetivos Específicos

Essa unidade tem como objetivo:

* Introduzir a filosofia tidyverse de transformação de dados;
* Explicar cada um dos pacotes do tidyverse;
* Manusear efetivamente dados dos tribunais brasileiros;
* Ensinar a criar, transformar e remover colunas;
* Mostrar como importar e exportar dados.

Playlist

Assista aos vídeos do Prof. José de Jesus referentes ao Tidyverse, clicando na playlist abaixo:

|  |  |
| --- | --- |
| **Título do vídeo** | **ID** |
| (3.1) Importação e exportação com readr, readxl, writexl, readRDS e saveRDS | A ser informado pelo estúdio |
| (3.2) Criação de dataframes e tibbles com tidyr |  |
| (3.3) Transformação de dados com dplyr |  |
| (3.4) Alteração de tibbles com tidyr |  |
| (3.5) Repetição com purrr |  |
| (3.6) Visualização com ggplot2 |  |

Material Complementar

[Introdução ao tidyverse](https://www.youtube.com/watch?v=dvkFWm6Th54)

[Tidyverse: tibble](https://youtu.be/w29CxI3TTgs)

[Dplyr: Select e distinct](https://youtu.be/jTQkMHYxwew)

[Dplyr: filter, arrange, count e pipe](https://youtu.be/A8_xZtyWsSs)

[Mutate e transmute](https://youtu.be/nJSIwzX7h8c)

Atividades

Proceda com as atividades propostas para essa unidade para fixar o conhecimento adquirido até aqui.

UNIDADE 3 - **FÓRUM DA UNIDADE 3**

Olá Turma,

Para testar seus conhecimentos sobre tidyverse, vamos utilizar o mesmo dataframe obtido na atividade da unidade 2 para aplicar algumas funções do dplyr.

Siga os passos do script que pode ser acessado no link abaixo para manipular os dados.

[Exercícios com dplyr](https://gist.github.com/jjesusfilho/c7ea9ab44246a8f4df89e5f4d1907e6b)

Partilhe os arquivos gerados com o tutor.

Verificando os Objetivos Da Unidade

Verifique se você alcançou os objetivos propostos para essa unidade. Para isso, identifique se você é capaz de:

* Selecionar colunas com select do dplyr;
* Filtrar valores de colunas com o filter do dplyr;
* Gerar frequências com o count do dplyr;
* Salvar os resultados em formato excel;

Caso haja alguma dúvida em algum dos objetivos propostos nessa unidade, reveja os vídeos ou, se a dúvida persistir, entre em contato com seu tutor pelo Fórum de dúvidas, disponível na disciplina.

**Unidade 4 – Textos e expressões regulares**

Objetivos Específicos

Essa unidade tem como objetivo:

* Extrair padrões de textos;
* Alterar padrões de textos;
* Remover padrões de texto;
* Extrair padrões por posição;
* Detectar padrões de textos.

Playlist

Assista aos vídeos do Prof. José de jesus Filho referentes aos Textos e Expressões Regulares, clicando na *playlist* abaixo:

|  |  |
| --- | --- |
| **Título do vídeo** | **ID** |
| Funções básicas | A ser informado pelo estúdio |
| Expressões regulares |  |
| Extração de padrões |  |
| Substituição de padrões |  |
| Remoção de padrões |  |
| Detecção de padrões |  |

Material Complementar

[Tutorial completo de regex](https://youtu.be/5Tvrq9G2t68)

[Tutorial sobre stringr](https://direito.consudata.com.br/shiny/stringr)

Atividades

Para fixar o conhecimento sobre Penas, participe do fórum de discussão e proceda com as atividades propostas para essa unidade.

**FÓRUM DA UNIDADE 4**

Olá, turma!

A última atividade procura resumir o que foi aprendido até aqui. Use o script abaixo para acessar o exercício.

[Infanticídio com mutate e stringr](https://gist.github.com/jjesusfilho/0a8f7ab27e080f4913d7ec65c73f68cf)

Quantos casos de homicídio existem?

Verificando os Objetivos Da Unidade

Agora, verifique se você alcançou os objetivos propostos para essa unidade. Para isso, identifique se você é capaz de:

* Usar eficientemente a função mutate para criar novas colunas;
* Usar eficientemente a função str\_detect para identificar padrões de texto;
* Gerar frequências com count;
* Usar o str\_sub para extrair padrão por posição;
* Usar regex para minerar textos jurídicos.

Caso haja alguma dúvida em algum dos objetivos propostos nessa unidade, reveja os vídeos ou, se a dúvida persistir, entre em contato com seu tutor pelo Fórum de dúvidas, disponível na disciplina.

Parabéns, você finalizou a disciplina sobre Mineração de textos jurídicos com R do seu curso. Esperamos que você tenha aproveitado ao máximo os recursos dessa Sala de Aula.

Aguardamos você na próxima etapa!