Recuperação de informações

Júlio César Batista
@ejuliobatista | julio.batista@outlook.com

Web scraping

Revisão

HTML: Hypertext Markup Language

- Diagramação de layouts de páginas da web

CSS: Cascading Style Sheets

- Estilo (cores, fontes, alinhamento, outros) de páginas da web

Javascript

- Programação (comportamento) de página web

DevTools

- Ferramentas no navegador que ajudam no desenvolvimento web

Referências úteis

https://github.com/rennerocha/scrapy-tutorial

https://github.com/rennerocha/pybr14-scrapy-tutorial

https://www.youtube.com/watch?v=5LMG4OCoEn0&t=248s

https://www.youtube.com/watch?v=rbiKXQSOWIM

Referenciando elementos na Web

Css Selectors: https://code.tutsplus.com/en/tutorials/the-30-css-selectors-you-must-memorize--net-16048

XPath: https://devhints.io/xpath

Utilidades

https://github.com/scrapinghub/extruct: Extrai informações de data schemas (HTML Microformat, Open Graph)

https://github.com/scrapinghub/dateparser: Extrai datas em vários formatos e idiomas

https://github.com/scrapinghub/spidermon: Monitoramento de spiders/crawlers

https://github.com/scrapinghub/arche: Validação de dados a partir de JSON Schema

https://github.com/scrapinghub/price-parser: Extrai valor e moeda de um preço

https://github.com/scrapinghub/js2xml: Transforma JS em uma árvore XML

Hands-on

https://docs.scrapy.org/en/latest/

Vamos coletar dados de http://quotes.toscrape.com/

- scrapy shell: útil para o desenvolvimento
- scrapy runspider executa apenas um spider
- scrapy startproject: projetos maiores
  - Pipelines
  - Middlewares
  - Extensions
  - Settings
  - Items & ItemLoaders

Exercício

http://books.toscrape.com/index.html

A partir da URL acima, colete as seguintes informações para todos os livros:

- Nome: string
- Preço: float
- Disponivel: bool
- Quantidade: int
- Avaliacao (número de estrelas): float
- Categoria: string
- UPC: string

Exercício - Exemplos

http://books.toscrape.com/catalogue/olio_984/index.html

- Nome: “Olio”
- Preço: 23.88
- Disponivel: True
- Quantidade: 19
- Avaliacao (número de estrelas): 1
- Categoria: “Poetry”
- UPC: “feb7cc7701ecf901”

Exercício - Exemplos

http://books.toscrape.com/catalogue/scott-pilgrims-precious-little-life-scott-pilgrim-1_987/index.html

- Nome: “Scott Pilgrim's Precious Little Life (Scott Pilgrim #1)”
- Preço: 52.29
- Disponivel: True
- Quantidade: 19
- Avaliacao (número de estrelas): 5
- Categoria: “Sequential Art”
- UPC: “3b1c02bac2a429e6”

Exercício - Extras

- Modifique o spider para receber uma parâmetro que é uma categoria disponível no site (“Poetry” por exemplo) e colete apenas os livros dessa categoria.
  - É possível coletar todos os livros e apenas filtrar os que são dessas categorias
  - Faça a alteração para encontrar a categoria no website (este link http://books.toscrape.com/catalogue/category/books/poetry_23/index.html) e coletar apenas estes livros
  - Exemplo do comando para rodar: `scrapy crawl books.toscrape.com -a categoria=”Poetry”`

Exercício - Extras

- Gere os dados do spider em um arquivo (csv, json, jsonlines, xml, outro) e leia no python/R. Alguns exemplos que podem ser investigados:
  - Como é a distribuição de preços
  - Como é a distribuição de quantidade disponível
  - Existe relação entre preço e avaliação (pode gerar um gráfico)
  - Existe relação do preço com a quantidade disponível
  - Existe relação da quantidade disponível com a avaliação
  - Existe diferença na avaliação dos livros entre categorias (categorias apenas com livros bons/ruins)
  - Existe diferença no preço entre as categorias (categorias com livros mais caros/baratos)
  - Quais as palavras mais/menos comuns nos títulos dos livros?
  - Existe relação do título (tamanho, palavras específicas) com o preço/avaliação?
  - Colete a descrição do livro e veja se existe alguma relação das palavras na descrição com a categoria, preço e avaliação)
  - Muito mais...

Experimentos

Processo

- Cada observação pode ser usada apenas para exploração ou confirmação, não ambos
- Uma observação pode ser usada quantas vezes quiser para exploração, mas uma vez que você usa uma observação para confirmação, ela não pode mais ser usada

Fonte: https://r4ds.had.co.nz/model-intro.html

Processo

- ~60% dos dados para exploração **(explorar os dados, gerar hipóteses)**
- ~20% dos dados para validação **(testar ideias e hipóteses construídas pela exploração)**
- ~20% dos dados para teste **(só pode ser usado uma vez, quando achar que tem um bom modelo no conjunto de validação)**

Fonte: https://r4ds.had.co.nz/model-intro.html

As separações podem mudar de acordo com o problema, mas esse é um bom ponto de partida

Processo

- Elaborar uma hipótese
  - Diminuir os passos de compra vai aumentar as vendas em 4%
- Coletar os dados (se não possuir)
  - Teste A/B: 50% dos usuários seguem o fluxo atual e 50% seguem o fluxo com menos passos
  - Quantas amostras são necessárias?
  - Como vamos separar os 50% dos usuários?
- Confirmar ou rejeitar a hipótese
  - Após a coleta, houve uma diferença significativa que representa 4% a mais de vendas no novo fluxo?

Processo

- Elaborar uma hipótese
  - Trocar do algoritmo A para o Algoritmo B, teremos um aumento de 10% de precisão
  - Precisão é medida pelo F1-Score
- Coletar os dados (se não possuir)
  - Executar o Algoritmo A
  - Executar o Algoritmo B
  - Quantas vezes os algoritmos devem ser executados com diferentes conjuntos de treino/validação?
- Confirmar ou rejeitar a hipótese
  - Houve uma melhora de 10% no F1-Score no Algoritmo B em relação ao Algoritmo A?

Processo - Métricas

- A métrica que você usa para avaliar o seu algoritmo é muito importante, visto que ela pode influenciar o resultado
- Por exemplo, existem 100 amostras na base, sendo:
  - 90 do Tipo A
  - 10 do Tipo B
- Se executarmos um algoritmo que sempre retorna o Tipo A (indiferente da situação) e medirmos a acurácia, ele terá 90% de precisão.
- Métricas no scikit-learn

Processo - Métricas

- Para classificação binária, o f1-score é uma boa opção
- Para ranqueamento (1 < 2) o ICC é uma opção
- Para regressão o RMSE é uma opção

Processo - Coleta de dados

Marcio coletou indicadores do mercado de ações a partir do ano de 2015 para criar um modelo para prever os preços futuros.

Ele pegou os dados disponíveis de todas as empresas no período.

Existe algum problema na abordagem de Marcio?

Processo - Coleta de dados

Maria ligou para a residência de 200 pessoas, entre 8h e 18h, para coletar informações sobre o transporte público em sua cidade.

Ela conseguiu falar com 100 das pessoas que ligou.

Existe algum problema nessa abordagem?

Processo

Referência

https://www.geckoboard.com/learn/data-literacy/statistical-fallacies/