Área de Data Science

Projeto Final - Trainees 2020 - Grupo Turing

Introdução

Olá!! Seja bem-vindx ao projeto final da área de Data Science do Grupo Turing!

Data Science tem uma função muito importante no mundo em que vivemos hoje. Na era da informação, a humanidade produz milhares e milhares de dados a cada momento, sejam eles sobre produtos, textos, informações de clientes, dados climáticos, dentre outros. Sendo assim, Data Science surge como a ciência responsável por coletar dados, entendê-los, processá-los, e, principalmente, conseguir extrair valor para contar uma história com eles.

Até agora, você já deve ter usado várias ferramentas de ciência de dados nas aulinhas do Turing Academy. Nesse projeto, você irá utilizar algumas novas técnicas, além de várias que você já deve conhecer e que poderão ser aprimoradas durante esse processo. Vamos lá!

Enunciado

Você já deve ter trabalhado com vários datasets, sejam eles de cursos do DataCamp ou de aulas que você participou no Turing. Mesmo sendo diferentes, esses datasets tinham uma característica em comum: todos foram previamente coletados e organizados por outra pessoa.

Nesse projeto, queremos que você trabalhe com os dados desde o início, ou seja, você deve ***coletar os dados*** da internet e realizar uma ***análise de dados***. O site que

você deve usar para este projeto é https://scrapethissite.com/pages/forms/?. Ele contém dados sobre o desempenho de diversas equipes de Hockey no Gelo nos últimos anos e, como o próprio nome sugere, foi feito para ser scrapeado, isto é, para fins didáticos.

Além disso, lembre-se que parte do processo de trabalhar com os dados é compreender a sua história, por isso, não se esqueça de também estudar um pouco sobre o problema que você estará analisando. Para isso, procure pesquisar um pouco sobre as regras do Hockey no Gelo, como a estrutura da competição mudou ao longo dos anos, se surgiram times novos nesse período, etc. Isso ajudará muito no processo de extrair insights interessantes.

Web Scraping

O processo de coleta de dados da internet é chamado de ***Web Scraping***. Para isso, temos diferentes ferramentas que podemos utilizar, porém, recomendamos que você utilize duas bibliotecas nesse projeto: **Requests** e **BeautifulSoup.**

As duas têm funções distintas: utilizamos a biblioteca *requests* para navegar pelas páginas e a biblioteca *BeautifulSoup* é usada para extrair os dados das páginas. Mesmo assim, geralmente você sempre usará essas duas juntas.

Para realizar essa tarefa, olhe o notebook de BeautifulSoup do Grupo Turing disponível na árvore de habilidades.

Ao final desta etapa, você deve ter os seus dados coletados e organizados em um arquivo *.csv*.

Análise e Visualização de Dados

Agora que você coletou os dados, chegou a hora de analisá-los. Para tornar isso mais interessante, temos mais alguns dados para você! Eles estão no arquivo *DataSet ESPN.csv*, presente na pasta de Data Science no Drive do Grupo Turing. Ele cobre o período de 2001 a 2011, e contém as seguintes colunas:

- *Team*: O nome da equipe
- *Home Games*: Quantidade de jogos disputados em casa, na temporada regular daquele ano
- *Home Total*: Número total de torcedores para todos os jogos em casa da temporada
- *Home Average*: Média de torcedores por jogo em casa da temporada
- *Road Games*: Quantidade de jogos disputados fora de casa, na temporada regular daquele ano
- *Road Average*: Média de torcedores por jogo fora de casa da temporada
- *Overall Games*: Total de jogos disputados na temporada regular daquele ano
- *Overall Average*: Média de torcedores por jogo da temporada (considerando tanto jogos em casa quanto fora de casa)
- *Year*: O ano em que a temporada terminou (por exemplo, se a temporada em questão for a de 2000-2001, o valor de "*Year*" é 2001)
- *Save Percentage*: Porcentagem de finalizações defendidas
- *Penalty Minutes*: Minutos de penalidade que o time sofreu
- *Penalty Minutes Against*: Minutos de penalidade que o time adversário sofreu

Caso você tenha ficado curioso, os dados foram extraídos desse site: http://www.espn.com/nhl/statistics

Talvez você ainda esteja se perguntando o que exatamente fazer com esses dados, mas a ideia é que isso não seja nada rigoroso mesmo: siga sua curiosidade.

Levante perguntas sobre os dados, e procure respondê-las você mesmo. Algumas ideias possíveis são: Será que times que marcam mais, ganham mais? Ou é o oposto que vale, e uma defesa sólida é mais importante? Como que o desempenho do goleiro afeta o resultado do time como um todo? Algum time tem evoluído muito nas últimas temporadas? A torcida é importante para o desempenho da equipe? Como o número de torcedores evoluiu ao longo do tempo?

Para quaisquer dúvidas sobre o projeto ou sobre onde procurar conteúdos, converse com seu mentor ou com a equipe de Data Science (Victor, Azank e Julia). Esperamos que você se divirta e que aprenda com essa experiência. Quando terminar, não esqueça de subir tudo no seu GitHub.