A Jornada para a Ciência de Dados

Tiago Alves

2020-05-04

Contents

Chapter 1

Uma Jornada Inesperada

Um curso básico e completamente grauito para geeks (e outras pessoas)

Bem vindo! Este é um curso básico de ciências de dados para pessoas que estão começando. Aqui você encontrará muita matemática, estatística, dados e referências à cultura pop.

1.1 Sobre mim

Eu decidi criar este curso para despertar o interesse de mais pessoas para a ciência de dados. Apesar de ser uma área já tão popular hoje em dia, seu conteúdo é denso, repleto de operações matemáticas, álgebra linear e estatística. Não se engane, vamos falar um pouco sobre estes temas aqui, mas espero conseguir trazê-los de uma forma leva e repleta de referências à cultura pop.

O objetivo deste curso não é formar novos cientistas de dados, mas apontá-los na direção certa. O conteúdo aqui presente não tem a intenção de ser o mais completo e detalhado, apenas de mostrar por exemplos como se desenvolve um projeto com dados, buscando dispertar o interesse de pessoas que possam não conhecer a área. Caso você já seja um cientista de dados, fica ~~vai ter bolo~~, quem sabe você aprende algo novo?

Eu me chamo Tiago Alves, sou mestre em Ciências da Computação pela UFMG desde 2018 e trabalho como cientista de Dados desde 2016, já trabalhei com projetos nas áreas … resolvendo problemas como …

http://tiagohca.com/

1.2 Disclaimer

Este é um curso sob construção e em constante mudança. Eu sou responsável apenas pelo conteúdo escrito. Todos os direitos de imagens e propriedades intelectuais devems ser devidamente respeitados.

Chapter 2

Um básico sobre python

Recomendo fortemente que você tenha alguma noção de programação, pelo menos o básico.

Se não tiver, vou tentar te explicar o necessário aqui.

Caso já saiba programar e já tenha utilizado Python, você pode pular para Completando o Pokedex - Um tutorial sobre como obter dados

2.1 Configurando o ambiente

Antes de começar, precisamos instalar o Anaconda.

Acesse este https://www.anaconda.com/products/individual e faça a instalação de acordo com o seu sistema.

Uma das melhores invenções do ser humano foi o Jupyter Notebook, vindo logo depois da internet e do microondas.

2.2 O primeiro programa

Programar é uma tarefa que envolver um forte raciocínio lógico. Não é algo complicado em sua essência, mas também não é algo que se aprenda em alguns minutos. Novamente, eu recomendo que você procure um curso especializado de programação caso esta seja sua primeira vez.

Aqui vou mostrar os primeiros passos para se começar a programar.

2.3 Bibliotecas e outras ferramentas úteis

Nos dias de hoje é raro construirmos programas do zero, graças às extensas comunidades existentes, independente da linguagem que você use. Python possui uma comunidade especialmente grande, com inúmeras bibliotecas públicas que agilizam muito o nosso trabalho. Lembre-se, não há por que reinventar a roda quando você pode comprar um carro.

2.3.1 Pandas

2.3.2 Scikit-Learn

2.3.3 Matplotlib

2.3.4 Plotly

Chapter 3

Completando o Pokedex - Um tutorial sobre como obter dados

O primeiro passo de qualquer projeto de ciência de dados é... conseguir os dados.

Para nossa sorte, existe uma tonelada de dados públicos que podem ser encontrados facilmente na internet. Algumas fontes boas de dados são ...

No nosso primeiro projeto vamos adentrar o mundo Pokémon! Para isto, precisamos de uma base de dados sobre os monstrinhos. A base que usaremos neste projeto pode ser encontrada neste link (você talvez tenha que criar uma conta). O Kaggle, plataforma onde está a base, é um excelente lugar para aprimorar seus conhecimentos em ciência de dados. Lá você encontra bases de dados, códigos e tutoriais feitos pela comunidade, fórums de discussão e competições emocionantes (algumas até chegam a pagar milhares de dólares para os vencedores!).

Assim como no mundo Pokémon encontramos diversas espécies quando entramos no mato alto, no mundo real encontramos diversos tipos de dados. Antes de mais nada vamos falar um pouco sobre estes tipos.

3.1 Tipos de dados

3.1.1 Tabular

3.1.2 Texto

3.1.3 Imagem

3.1.4 Áudio

3.2 Carregando a base de dados

Chapter 4

Como saber tudo o que há sobre os Pokémon - A análise exploratória

Agora que já temos as informações sobre os monstrinhos precisamos descobrir o que elas estão nos dizendo. Esta é uma das partes mais cruciais do projeto, e quando executada com atenção pode te levar muito longe. O que vamos fazer chama-se E.D.A. (Exploratory Data Analysis, ou em português, Análise Exploratória dos Dados)

Chapter 5

Descrevendo os Pokémon - Como funciona a engenharia de features

Agora que já entendemos bastante do dado, está na hora de começar a descrever os Pokémon de uma maneira que uma máquina entenda. Calma! Não vamos escrever zeros e ums! Mas precisamos de fato criar alguns números...

A base de dados consiste de várias informações estruturadas sobre cada Pokémon e embora seja fácil para um ser humano interpretar o que está lá, a máquina pode ter dificuldade com algumas das colunas. Esta etapa chama-se Engenharia de Features (ou de Características). Uma feature nada mais é que uma caracterísica que ajuda a descrever alguma coisa. Por exemplo, nós seres humanos podemos ser descritos com apenas algumas features: idade, gênero, se assistiu todos os filmes do MCU ou não. O que vamos fazer agora é descrever os Pokémon como uma série de números onde cada um representa uma feature que iremos definir.

5.1 One-hot encoding

5.2 Label encoding

5.3 Normalização

Chapter 6

Lendário ou não? - Um primeiro modelo de classificação

Quanto trabalho nos dados, não é mesmo? Não se preocupe, finalmente chegou a hora: vamos criar um modelo!

Um modelo é uma função $y = f(x)$ ...