Dados, onde vivem? Desbravando o Mundo dos Dados Abertos

Não é mais novidade para ninguém [*a menos que você estivesse morando em uma caverna nos últimos 10 anos*] que dados se tornaram o novo petróleo. E o que isso significa? De forma básica e brusca: dados ganharam um alto valor de mercado, sendo extremamente poderosos e influentes no que diz respeito a moldar o futuro da ciência, tecnologia, educação, mercado e por aí vai.
Aí você se pergunta: "sim, e eu com isso?". Bom, aí é que eu jogo uma réplica pra você, "Por que não aproveitar esse *boom* dos dados a seu favor?". E aqui esqueça papo de coach, de ganhar dinheiro em 2 semanas, de se tornar Cientista de Dados Sênior da Google em 3 meses e qualquer conversa dessas, o que interessa aqui é "se dados têm influenciado tanto assim o mundo, e consequentemente a sua vida, como você pode entender melhor o que está acontecendo, como descobrir quais dados são esses e como eles estão sendo utilizados?".

Aí é que eu faço o meu jabá:

> ***Com diversos tipos, formatos, cores e sabores, os dados estão por aí, permeando o mundo, levando a decisões, alicerçando pesquisas científicas e tornando o seu feed mais agradável para você.***
> ***Porém, onde encontrá-los?***
> ***Neste minicurso iremos introduzir o mundo dos dados abertos, apresentando algumas das bases de dados gratuitas mais famosas, a importância da utilização desses recursos e onde podemos aplicá-los.***

[*Ok, esse texto ficou muito longo, vamos logo ao que interessa.*]

---

1. O que são Dados Abertos?

Uma das melhores formas de explicar o que são dados abertos é da Open Knowledge[1]:

> *Dados abertos são dados que podem ser livremente usados, reutilizados e redistribuídos por qualquer pessoa - sujeitos, no máximo, à exigência de atribuição da fonte e compartilhamento pelas mesmas regras.*

[1] https://opendatahandbook.org/guide/pt_BR/what-is-open-data/

Túlio Sousa de Gois
Vice-presidente da LADATA

Agora que temos a explicação básica desse conceito, vamos aprofundar um pouco mais.
Quando falamos nesse tipo de dado, é necessário também entender o conceito de “aberto” que está sendo utilizado. Mas não se preocupe, isso será entregue de uma forma bem resumida e prática na forma das “três leis dos dados abertos”. Essas “leis” foram sugeridas pelo ativista de dados abertos “David Eaves”[2], e dizem o seguinte:

1. Se o dado não pode ser encontrado ou indexado na Web, ele não existe;
2. Se o dado não está disponível num formato aberto e legível por máquina, ele não pode ser reutilizado;
3. Se dispositivos legais não permitem que ele seja compartilhado, ele não é útil.

Certo, sabemos o que são dados, o que são dados abertos, qual o significado de “aberto” nesse contexto. Agora deixo uma pergunta para você:

> *Qualquer dado pode ser aberto?*
>
> Não! Quando adentramos esse âmbito, muitas vezes estamos olhando para o lado do movimento de Governo Aberto, onde todas as informações públicas governamentais (ou seja, os dados públicos) devem ser acessíveis (no caso, abertas). **Porém, nem todo dado é público**.

2. Dados abertos ao nosso favor

Ao longo da introdução desse texto foi citado que você poderia utilizar essas informações a seu favor. Agora, irei explicar um pouco do por que você não só pode, como deve.

Essa concepção de abertura dos dados, principalmente os governamentais, começou a ocorrer em meados de 2009, quando alguns governos anunciaram a abertura de suas informações públicas. Desde então, outros governos e organizações aderiram ao movimento e esses dados vêm sendo aplicados de diversas formas e gerando resultados surpreendentes, por exemplo:

- O ‘tax tree’ na Finlândia e o ‘where does my money go’ no Reino Unido mostram como o dinheiro dos impostos está sendo gasto pelo governo;

[2] https://eaves.ca/2009/09/30/three-law-of-open-government-data/

- O Google Tradutor usa documentos da União Europeia que aparecem em todos os idiomas europeus para treinar algoritmos de tradução;
- Na Dinamarca, o ‘findtoilet.dk’ mostra todos os sanitários públicos do país;
- O QEdu permite a qualquer pessoa obter informações sobre a qualidade do aprendizado nas escolas brasileiras, com dados sobre escolas públicas e particulares.

A realidade é que ainda não se sabe qual o limite do que pode aparecer através da aplicação dos dados abertos, e esse potencial ainda inexplorado é excelente em termos de inovação. Novas combinações de dados podem criar novos conhecimentos e descobertas, que podem levar a campos de aplicação totalmente novos.

E onde você entra na jogada?

Muito simples de responder! A exploração desses dados necessita que alguém vá lá, abra, analise, elabore hipóteses, teste, visualize... Que faça todos os procedimentos possíveis para a criação de novos conhecimentos e aplicações.
E além do ganho público dessas aplicações e conhecimentos, onde esse tipo de projeto pode ser inserido para que o trabalho de quem o fez seja reconhecido? Se você pensou em “**Portfólio**”, parabéns, acertou!

A partir daqui vamos começar um pouco da exploração desse mundo, iniciando pelo fim, ou seja, pelos resultados.

- **Educação em escala global:** https://www.kaggle.com/code/nelgiriyewithana/introduction-to-world-educational-data
- **Tutorial para iniciantes usando um dataset de Pokémon** https://www.kaggle.com/code/kanncaa1/data-sciencetutorial-for-beginners

Tutoriais e textos sobre o que você está aprendendo ou sabe também são portfólio, afinal, de que adianta ter dados se você não souber comunicar o que eles querem dizer?

- **Princípios básicos para a visualização de dados** https://medium.com/datavizbr/princípios-básicos-da-visualização-de-dados-5ebc7150fc81

Túlio Sousa de Gois
Vice-presidente da LADATA

- **Estatística — Dados, Gráficos de Dispersão e Gráficos de Barras**
  https://medium.com/@anwarhermuche/estatística-dados-gráficos-de-dispersão-e-gráficos-de-barras-c68303489cf2

Alguns materiais complementares sobre portfólios:

- **Programação Dinâmica**
  https://medium.com/programacaodinamica/como-começar-um-portfólio-de-cientista-de-dados-dd68ee64c85c
- **Datawars**
  https://www.datawars.io/articles/12-free-data-science-projects-to-practice-python-and-pandas

3. Uma aventura pelo mundo dos dados

Agora chega de tanto texto, é hora de você explorar um pouco mais por sua conta em risco. Irei deixar aqui as indicações, descrições e aí é sua hora de pensar.

Base dos dados

A base dos dados é uma ONG sem fins lucrativos e *open source* que trabalha para universalizar o acesso aos dados de qualidade.
No site você irá encontrar mais de 1000 datasets de diferentes organizações, tanto brasileiras quanto internacionais, além de tutoriais e a documentação.
https://basedosdados.org

Portal Brasileiro de Dados Abertos

O Portal Brasileiro de Dados Abertos é o site oficial do Governo Federal que disponibiliza dados oficiais publicados tanto pela federação, quanto governos locais, ministérios e afins.
O site possui um mecanismo de busca avançada para os conjuntos de dados permitindo aplicar diversos filtros, como tema, formato e palavras-chave. Atualmente, estão disponíveis mais de 12000 conjuntos de dados.
https://dados.gov.br/home

kaggle

O kaggle não é somente uma plataforma que disponibiliza datasets, é uma comunidade completamente voltada para a Ciência de Dados. Lá vocês encontrarão mais de [*não tenho nem ideia de um número mas é bastante*] datasets, fóruns sobre ciência de dados (sobre Inteligência Artificial, Python, Machine Learning, um monte de coisa),

cursos gratuitos, modelos de ML pré-treinados. O Kaggle é literalmente um mundo de informações sobre tudo relacionado à Ciência de Dados.
https://www.kaggle.com/datasets

[Nota do autor: *eu recomendo fortemente que você crie uma conta no kaggle, independente de você querer seguir na área de DS, pois lá você encontrará materiais que te auxiliarão em N diferentes momentos da sua vida caso você se interesse minimamente por programação, dados ou IA.*]

HuggingFace

Sabe tudo o que você acabou de ler sobre o que o kaggle é na Ciência de Dados? Pois dobre e passe pro próximo... Brincadeiras à parte, o HuggingFace é tudo aquilo e um pouco mais só que agora para a Inteligência Artificial. Lá você irá encontrar datasets **imensos** que foram utilizados para treinar diferentes modelos de aprendizado de máquina, os próprios modelos disponíveis para testes, fóruns e discussões, artigos (científicos e informacionais), documentações... [*Chega, abra o link e vá explorar, senão esse texto não acaba*]
https://huggingface.co

[*Vou precisar pedir pra criar conta dessa vez? Eu espero que não*]

4. Seu momento de brilhar

Essa é a melhor parte, a hora de começar a vislumbrar algumas das coisas que os dados podem te fornecer em quesitos de projetos.

[*Não se preocupe, tá tudo escrito bonitinho para você entender cada pedaço do que vamos fazer.*]

Explorando os dados do Spotify

https://www.kaggle.com/code/tuliosg/dados-onde-vivem-explorando-os-dados-do-spotify

*** BÔNUS**: Brincando com modelos no Hugging Face.

Como apareceu em algum lugar nesse texto aqui, o HuggingFace disponibiliza diversos modelos no site, e isso não é tudo, é possível testá-los!

Túlio Sousa de Gois
Vice-presidente da LADATA

Assim, como bônus para você que chegou até aqui, o momento agora é de pura diversão, hora de brincar com modelos de IA. Irei deixar alguns dos que julgo muito interessantes aqui embaixo, mas fica à seu critério:

- **Modelos do HuggingFace**
  https://huggingface.co/runwayml/stable-diffusion-v1-5
- **Análise de sentimentos multilíngue**
  https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment-multilingual?text=Esse+minicurso+de+dados+abertos+é+muito+bom
- **Convertendo texto em imagem**
  https://huggingface.co/runwayml/stable-diffusion-v1-5

5. Finalizando

Chegou o momento triste... A hora da despedida :(

Aqui, eu agradeço imensamente pela atenção ao longo de todo o minicurso (para as pessoas que estiveram presentes) e também ao longo do texto (para você que teve coragem de ler até aqui). A você, o meu **muito obrigado!**

MAS, [quase] todo momento nessa vida é um “até logo”, sendo assim, permita-me deixar algumas formas de você continuar por dentro do mundo dos dados e acompanhando tanto o trabalho da LADATA quanto o meu:

Instagram da LADATA: @ladata.ufs
GitHub da LADATA: @ladata-ufs
GitHub | kaggle : @tuliosg

Referências

Base dos Dados. Disponível em: <https://basedosdados.org>.

Datasets Documentation. Disponível em: <https://www.kaggle.com/docs/datasets#resources-for-starting-a-data-project>.

Guia de Dados Abertos. Disponível em: <https://opendatahandbook.org/guide/pt_BR/>. Acesso em: 18 nov. 2023.

Guia de dados abertos. [s.l: s.n.]. Disponível em: <https://ceweb.br/media/docs/publicacoes/13/Guia_Dados_Abertos.pdf>.

Túlio Sousa de Gois
Vice-presidente da LADATA

Hugging Face – The AI community building the future. Disponível em: <https://huggingface.co>.

KAGGLE. **Kaggle: Your Home for Data Science**. Disponível em: <https://www.kaggle.com>.

MODELO DE REFERÊNCIA DE ABERTURA DE DADOS Documento de referência do Marco 5 do Compromisso 2: Ecossistema de Dados Abertos. [s.l: s.n.]. Disponível em: <https://repositorio.cgu.gov.br/bitstream/1/46701/5/modelo_de_referencia_para_publicacao_de_dados_abertos.pdf>. Acesso em: 18 nov. 2023.

PASQUETTO, I. V.; RANDLES, B. M.; BORGMAN, C. L. On the Reuse of Scientific Data. **Data Science Journal**, v. 16, 22 mar. 2017.

Portal de Dados Abertos. Disponível em: <https://dados.gov.br/home>.

Túlio Sousa de Gois
Vice-presidente da LADATA