Processo Seletivo

10 de janeiro de 2018

1 Qualidade de dados

Não dá para pensar em dados sem pensar em números astronômicos. Você já deve ter ouvido falar que 90% dos dados que há no mundo hoje foram criados apenas nos últimos dois anos e, agora, a cada dois anos, o mundo dobra a taxa em que os dados são produzidos. Estudo recente da Business Software Alliance (BSA), afirma que $2,5$ quintilhões de bytes são criados todos os dias.

Com o advento das Internet das Coisas (em inglês *Internet of Things*), o volume de informações trafegadas e armazenadas tende a aumentar exponencialmente. Hoje em dia, estima-se que apenas 1% de todos os objetos físicos que poderiam estar conectados à internet estão conectados atualmente. Em 2020, a Cisco 50 bilhões de dispositivos estão trocando dados. Para ter uma dimensão do volume, celulares, tablets e computadores serão responsáveis apenas por 17% do tráfego total.

Não resta dúvidas de que a quantidade dos dados é a essência da IoT. No entanto, também é necessário aumentar a qualidade dos dados gerados pelos dispositivos conectados a essa infraestrutura e que, posteriormente, são transmitidos às empresas e aos tomadores de decisão. Eles precisam ser de total confiança.

Contextualização

A Twist recentemente lançou um produto que entrega aos seus clientes um *framework* de monitoramento de qualidade de dados (Twist DQ) para seus clientes. O Framework consiste, simpli-

ficando, um sistema de notas para **cada registro** do *dataset* monitorado. Ou seja, cada registro, recebe notas para diferentes dimensões de qualidade, como, por exemplo, completude, consistência, acurácia, não-duplicidade, confiabilidade, etc.

Exemplo

ID	Nome	Idade	Telefone
A	Fernando	32	
B	José	Vinte	3121-3131
C	Afonso	190	+ 55 (21) 4894-0404

A base de dados fictícia apresenta alguns problemas de qualidade de dados. Vamos considerar apenas três dimensões:

Completude A informação se encontra disponível?

Acurácia A informação pode ser considerada fiel aos fatos que ela representa?

Integridade A informação presente está íntegra, inteira, ou está corrompida, adulterada?

De cara, percebemos que o telefone do item A (nome Fernando) está faltante. Logo, o registro Telefone do Item A tem uma nota 0 para completude.

O item B apresenta alguns problemas. O primeiro é a representação da idade. O número está escrito por extenso e não como um numeral, como esperado. Apesar da informação passar a mensagem corretamente (vinte anos é uma idade provável), porém ela não está integra, pois sua forma está incorreta. Já o telefone está incompleto, pois não possui DDI ou DDD.

O item C não tem idade acurada, já que não há registros de pessoas com 190 anos humanidade.

Sendo assim, as notas dos registros seriam compostas conforme a seguinte tabela:

ID	Nome			Idade			Telefone			**Nota**
	Completude	Acurácia	Integridade	Completude	Acurácia	Integridade	Completude	Acurácia	Integridade	
A	10	10	10	10	10	10	0	0	0	6,3
B	10	10	10	10	10	0	10	0	10	7,8
C	10	10	10	10	0	10	10	10	10	8,9
Nota		10			7,8			5,6		7,8

O item A possui a menor qualidade, pois tem toda uma variável faltante. Os demais itens refletem os números de problemas encontrados. Assim, o item C é o item com mais qualidade. Em relação às variáveis, não é necessário apontar que a variável telefone é a que possui mais problemas, assim, possui a menor qualidade. Há provavelmente um problema de aquisição dessa variável, e a nota obtida mostra isso.

Neste ponto, é interessante reparar que o sistema de notas apresenta 3 visões sobre a qualidade do banco:

Visão Gerencial Um gerente não necessariamente necessita saber todos os detalhes do sistema, apenas uma avaliação generalista. Neste caso, a base de dados avaliada possui nota $7,8$. A partir daí, os especialistas podem ser acionados e possíveis melhorias no processo realizada de maneira que a nota geral do repositório de dados melhore.

Visão Especialista Cada variável pode representar uma etapa do processo de aquisição. Uma variável com nota baixa pode indicar um pedaço do processo com problemas que necessitam a intervenção de um especialista. No caso, a variável Telefone está com problemas.

Visão Analista Entradas com baixa qualidade podem ser descartadas pelos analistas de dados. No

nosso caso, o item A possui qualidade baixa, e talvez poderia ser desconsiderada.

Exercício

Nessa atividade use Python e seu módulo Pandas. Por exemplo, para ler o excel basta:

```
import pandas as pd
df = pd.read_excel("database.xlsx")
```

Utilizar Jupyter notebooks é um bônus desejado (mas não requerido).

Junto com esse texto, estamos encaminhando uma base de dados, que possui 63 registros e 26 variáveis. Em cima dessa base são pedidos

1. As notas:
    a) Visão Gerencial (uma nota)
    b) Visão Especialista (para cada variável - serão 26 notas)
    c) Visão Analista (para cada item - serão 63 notas)

Nota: as dimensões de qualidade serão as mesmas apresentadas na contextualização: Completude, acurácia e integridade. Para mérito de facilidade, para qualquer uma das dimensões, caso um problema seja encontrado, a nota atribuída é 0 (zero). Se nenhum problema é encontrado, a nota é 10 (dez).

Considere os intervalos de validade da Tabela 1.1

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Min	10	1	1	1	2	1	1	0	1	1	2	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
Max	100	2	2	2	7	8	2	7	7	4	8	1	3	3	1	2	1	2	4	4	4	9	13	5	2	2

Tabela 1.1: Limites

2. A base de dados apresentada possui inúmeros dados faltantes. Uma maneira de minimizar esse problema é através da imputação de dados. Realize a imputação de dados (Use o módulo de Python `statsmodels.imputation.mice.MICEData`). Apresente a média para cada variável antes e depois da imputação.

Referências Bibliográficas

[1] Tome nota: 2,5 quintilhões de bytes são criados todos os dias, CIO, http://cio.com.br/noticias/2015/10/27/tome-nota-2-5-quintilhoes-de-bytes-sao-criados-todos-os-dias/

[2] Internet of Things (IoT): O que é (continuação)?, Target Solutions, https://www.targetso.com/2016/07/29/internet-of-things-iot-conceito-continuacao/

[3] A qualidade dos dados importa muito no universo da Internet das Coisas, Informatica Blog, https://blogs.informatica.com/br/2017/03/09/a-qualidade-dos-dados-importa-muito-no-universo-da-internet-das-coisas/

[4] statsmodels.imputation.mice.MICEData http://www.statsmodels.org/dev/generated/statsmodels.imputation.mice.MICEData.html