



BigML Sources

2023

<lapti>

Programa

- Introdução
 - ~ Propósitos de sources em BigML
- Features básicas
 - ~ Formato de arquivos
 - ~ Métodos de upload
- Features avançadas
 - ~ Source parser
 - ~ Text analysis



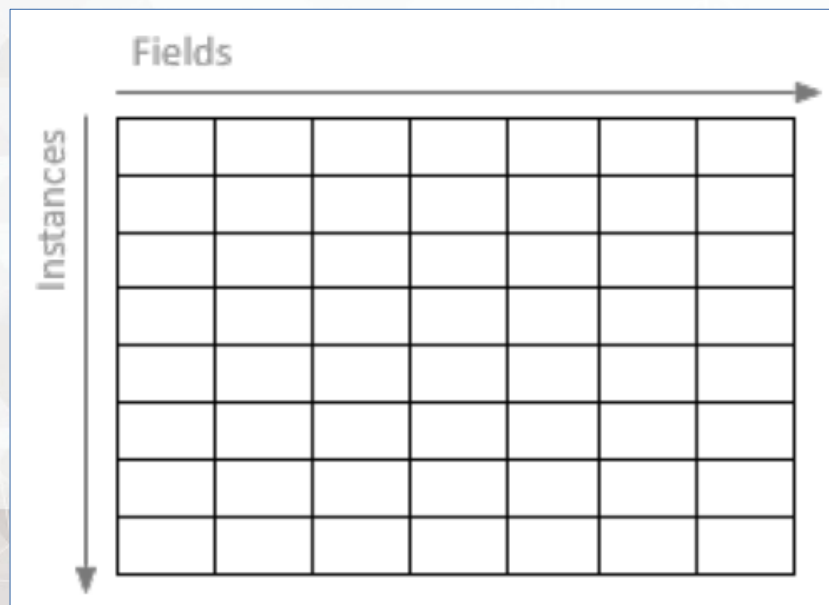
O que é uma Source?

- Sources são o “primeiro passo” no workflow convencional do BigML
 - ~ Source, dataset, modelo, previsão
 - ~ Permitem o upload de dados para o ambiente do BigML
 - Para começar as análises
 - ~ Revisar e corrigir tipos de campo auto-detectados
 - Algoritmo padrão de detecção faz generalizações



O que é uma Source?

- Dados *Machine Learning – Ready Format*
~ Coleção de instâncias em formato tabular



Tipos de campos

- Nas Sources, o tipo do campo auto-detectado é mostrado
- Tipos:
 - ~ Numérico
 - Qualquer valor em número reais
 - ~ Categorical
 - Lista de valores discretos
 - ~ Text
 - Texto free-form
 - ~ Datetime
 - ISO timestamps
 - ~ Items
 - Lista de itens, como os itens em um carrinho de compras

Tipos de campos

1 2 3

1, 2.0, 3, -5.4

numeric

A B C

true / false
yes / no
giraffe / zebra / ape

categorical

DATE-TIME

2013-09-25 10:02

DATE-TIME

text

Be not afraid of greatness:
some are born great, some
achieve greatness, and some have
greatness
thrust upon 'em.

text

YYYY-MM-DD

YEAR

2013

YYYY-MM-DD

MONTH

September

YYYY-MM-DD

DAY-OF-MONTH

25

M-T-W-T-F-S-D

DAY-OF-WEEK

Wednesday

HH:MM:SS

HOUR

10

HH:MM:SS

MINUTE

02

“great” appears 2 times
“afraid” appears 1 time
“born” appears 1 time

items

bread, sugar, coffee,
milk
ice cream, hot fudge

items

Criando uma Source

- Dados precisam estar em um formato tabular
 - ~ CSV, TSV, ARFF (Attribute-Relation File Format - WEKA)
 - ~ JSON (listas de listas ou listas de objetos)
 - ~ MS Excel ou Number for Mac – exportação CSV é preferível
- Suporta compressão
 - ~ gzip, zip, etc
- Criada diretamente usando
 - ~ Drag-n-drop / File browser / Inline
- Criada via cloud-to-cloud
 - ~ URLs: http, s3, hdfs, etc
 - ~ Cloud integration: Google Drive / Dropbox
- Programaticamente
 - ~ API / Bindings



Schema	Description
asv://	Same as azure://
asvs://	Same as azures://
azure://	Microsoft Azure storage
azures://	Same as azure:// but using SSL ¹
dropbox://	Drobox-stored files
gcs://	Google Cloud stores
gdrive://	Google Drive files
hdfs://	The distributed storage used by Hadoop applications
http://	Regular HTTP-accesible files
https://	HTTP secure-accessible files
odata://	Open Data Protocol ² that consumes REST APIs
odatas://	Same as odata:// but using SSL
s3://	Simple Storage Service ³ (S3), the file storage provided by Amazon Web Services (AWS)

Configurações avançadas

- Locale
 - ~ Determina como números e acentos serão interpretados
- Parsing
 - ~ Separador / quotes / missing tokens
- Expansão datetime
- Separador de itens
 - ~ Default é “,”
- Análise de textos

The screenshot displays the 'Sources' configuration interface in WhizzML. The top navigation bar includes 'Sources', 'Datasets', 'Supervised', 'Unsupervised', 'Predictions', and 'Tasks'. A 'Configure source' button is visible. The dataset being configured is 'Movies 2000-2016'.

SOURCE CONFIGURATION

- Locale:** English (United States)
- Separator:** ,(comma) (SINGLE FIELD / MULTIPLE FIELDS toggle)
- Quote:** " (double quote)
- Header:** ml, a,b,c, -, -, - (Expand date-time fields toggle: DISABLED / ENABLED)
- Missing tokens:** "", NaN, NULL, N/A, null, -, #REF!, #VALUE!, ?, #NULL!, #NUM!, #DIV
- Items separator:** Auto detect

TEXT ANALYSIS (DISABLED / ENABLED toggle)

- Language:** Auto detect
- Tokenize:** All
- Stop words removal:** Yes (detected language) / Normal
- Max. n-grams:** five-gram (does runs / A/a)

Filter terms (n=1): Set the terms to be excluded from you dataset

Buttons: Reset, Update

Análise de Textos

Be not afraid of greatness:
some are born great, some
achieve greatness, and
some have greatness
thrust upon 'em

Análise de Textos

Be not afraid of greatness:
some are born great, some
achieve greatness, and
some have greatness
thrust upon 'em

great: aparece 4 vezes

Limites de tamanho

- Local sources
 - ~ Até 64 GB
- Remote sources
 - ~ Até 64 GB
 - ~ Em Amazon S3 (Simple Storage Service) até 5 TB
- Inline sources
 - ~ 16 MB
- Em tamanhos maiores, é possível usar o BigML Private Deployment
- O tamanho da fonte nunca segue limites, porém o Dataset gerado a partir dela é limitado pelo plano selecionado

Então

- Sources
 - ~ Permitem o upload de dados a serem analisados pelo BigML
 - ~ Permite que se corrijam os tipos de campos
- Arquivos de dados
 - ~ Tabulares e podem ser comprimidos
 - ~ Upload usando diversos métodos
 - Direto, URLs remotas, cloud integrations
- Opções avançadas
 - ~ Source parsing
 - Locale, separadores, etc
 - ~ Field parsing
 - Text, Datetime, Items

The background features a light gray geometric pattern of triangles. At the top left, there is a logo consisting of the letters 'UTFPR' in a stylized, blocky font. A thick yellow horizontal bar runs across the top, partially overlapping the logo. The word 'Obrigado' is centered in a large, black, sans-serif font.

Obrigado

leandro@utfpr.edu.br

<http://lapti.ct.utfpr.edu.br>

<lapti>