

Residência em Engenharia e Ciência de Dados

Processamento de Dados em Larga Escala

PROJETO

[ETL / treinamento / teste sobre Corpus PT7 multiclasse]

O PT7 Web (<https://ieee-dataport.org/open-access/pt7-web-annotated-portuguese-language-corpus>) é um Corpus anotado em língua portuguesa construído a partir de amostras coletadas de setembro de 2018 a março de 2020 de sete países de língua portuguesa: Angola, Brasil, Portugal, Cabo Verde, Guiné-Bissau, Macau e Moçambique. Os registros foram filtrados do Common Crawl — um conjunto de dados em escala de petabytes de domínio público de páginas da Web em vários idiomas, misturados em instantâneos temporais da Web, disponíveis mensalmente [1]. As páginas brasileiras foram rotuladas como classe positiva e as demais como classe negativa (português não brasileiro). O conjunto de dados totalizou 249,74 GB de texto HTML bruto relacionado a 16.346.693 páginas da web exclusivas. Os dados foram pré-processados para produzir vetores de distribuição de palavras de alta dimensionalidade ($2^{18} = 262.144$ características) como entrada para as fases de treinamento e teste. Uma demonstração do uso desses dados pode ser verificada em um projeto fracionário de dois níveis para investigar o desempenho do cluster no Spark [2].

[1] G.Wenzek, M.A.Lachaux, A.Conneau, V.Chaudhary, F.Guzman, A.Joulin, E.Grave, *arXiv preprint arXiv:1911.00359* (2019).

[2] Rodrigues, J.; Vasconcelos, G.; Maciel, P. *Time and cost prediction models for language classification over a large corpus on spark*.

In: 2020 IEEE Symposium Series on Computational Intelligence (SSCI). [S.l.: s.n.], 2020. p. to appear.

Será utilizado um extrato reduzido do PT7 Web, equivalente 17014 páginas ~ 0.1% do Corpus original. Foram disponibilizados cinco arquivos (pt7-raw.zip), separados pelo domínio de nível superior de cada país (.br, .pt, .mo, .gw, .mz, .ao e .pt).

```
CC-MAIN-2018-39_ao
CC-MAIN-2018-39_br
CC-MAIN-2018-39_gw
CC-MAIN-2018-39_mo
CC-MAIN-2018-39_mz
CC-MAIN-2018-39_pt
```

Os dados se encontram rotulados como 1:pt_BR e 0:pt_OTHERS e estão disponível no formato a seguir, onde:

- label - rótulo
- url - endereço original completo da página
- digest - uma função de hash do conteúdo da página
- raw - os dados brutos do texto da página após limpeza de tags HTML

label	url	digest	raw
0	http://1-1.pt/300...	JJYINTQR7DRFBMWAI...	\r\nEmpresas\r\nP...
0	http://1-1.pt/300...	AUP6PWDHXTGEHMW4M...	\r\nEmpresas\r\nP...
0	http://1-1.pt/300...	TMHH62CWHZRRRQK3...	\r\nEmpresas\r\nP...
0	http://1-1.pt/300...	BZXW5XKSAQUUF3FM...	\r\nEmpresas\r\nP...
0	http://1-1.pt/300...	TR4JYZTG563VKFQBY...	\r\nEmpresas\r\nP...
0	http://1-1.pt/300...	6XKFIZTSVDN4BYZTE...	\r\nEmpresas\r\nP...
0	http://1-1.pt/300...	3Y3MQ7H2ZF4B4Z7ST...	\r\nEmpresas\r\nP...
0	http://100-dj.pt/...	BPW5IV333K0AYNQOP...	\r\n100-DJ - ao r...
0	http://100olhos...	32I2HPTFDWTJ5P677...	\r\nLoading...\r...
0	http://100pavor.pt/	7LEAGIOKBGPYZRGLT...	\r\n \n100PAVOR\r...
0	http://100solucoe...	WAJQ27Y2I4SJ7SLN3...	\r\nInício\n(curr...
0	http://100trilhos...	MIMURTS2WQLS3NLSO...	\r\n \nGaleria\nCa...
0	http://100trilhos...	WSB2HR6434YB3YW3H...	\r\n \nEscalada\r...
0	http://13luas.pt/...	BIRLLHEMAPI7C6OKY...	\r\nContacto: +35...
0	http://13yachtbro...	OE5UZKCTVW3SGHIUM...	Porto de Recreio ...
0	http://13yachtbro...	R2CWSY7HDB5TEH4FV...	Porto de Recreio ...
0	http://13yachtbro...	NQ7YRBNLYIAFASEBN...	Porto de Recreio ...
0	http://13yachtbro...	RDZM2NLWH3P4Q7EZB...	\nKawasaki STX 11...
0	http://13yachtbro...	HHVCWGRAFYXDLS4X...	\nFairline Phanto...
0	http://13yachtbro...	5FWCBW3H6DRTK5EIV...	\nCranchi Atlanti...

A tarefa preliminar do projeto consiste em realizar o processo de ETL sobre os dados brutos, transformando o conteúdo de cada página web em um vetor esparsa de características no formato exigido pelo Spark. A base deve separar os dados em novos rótulos, de acordo com cada país, formando uma base rotulada multiclasse.

Tarefa 1 << disponibilizado >> : processamento ETL

Executando os passos descritos, você terá no HDFS dados no formato:

- label - ao, br, pt, mz, mo, gw (no conjunto reduzido há apenas seis países)
- features - vetor esparsa com a representação do texto de cada página

Passo-a-passo da Tarefa 1

- Baixe o arquivo pt7-raw.zip
- Copie a pasta descompactada para user_data/pt7-raw
- Copie os arquivos do PT7 para o HDFS

```
$ hadoop fs -put /user_data/pt7-raw hdfs://master:8020/bigdata/
```

- `Processe o jobs labels-pt7-raw.scala`

```
$ spark-shell --master spark://master:7077 -i /user_data/labels-pt7-raw.scala
```

Como resultado, será obtido um dataframe conforme imagens a seguir.

[illegible]

label	count
.br	7053
.pt	3054
.ao	2122
.gw	1603
.mz	2820
.mo	362

- Processe o job etl-pt7.scala

```
$ spark-shell --master spark://master:7077 -i /user_data/etl-pt7.scala
```

Como resultado, será obtido um dataframe conforme imagens a seguir. Neste ponto, o dataframe multilabel com os vetores esparsos será gravado no seu HDFS no caminho `hdfs://master:8020/bigdata/pt7-hash.parquet`

label	features
.mz	(262144, [3704, 457...
.mz	(262144, [944, 1004...
.mz	(262144, [69, 1004, ...
.mz	(262144, [1004, 200...
.mz	(262144, [452, 2511...
.mz	(262144, [69, 452, 5...
.mz	(262144, [452, 3170...
.mz	(262144, [242, 452, ...
.mz	(262144, [452, 2746...
.mz	(262144, [69, 1004, ...
.mz	(262144, [1004, 370...
.mz	(262144, [839, 1779...
.mz	(262144, [3704, 253...
.mz	(262144, [226, 452, ...
.mz	(262144, [69, 1004, ...
.mz	(262144, [2150, 215...
.mz	(262144, [69, 1004, ...
.mz	(262144, [2746, 370...
.mz	(262144, [2054, 354...
.mz	(262144, [3704, 376...

Este será o conjunto de dados de entrada para o treinamento e teste do modelo.

Tarefa 2 << a implementar >>: treinar e testar um modelo supervisionado.

Investigar as possibilidades de modelos supervisionados em <https://spark.apache.org/docs/latest/ml-classification-regression.html#classification>. Implementar em Scala, Python, R ou Java o processo de treinamento e teste do modelo escolhido.

Dica: Para construção do modelo, baseie-se no slide 12-spark-mllib-kddcup.pdf. E, para ler o dataframe com os vetores esparsos do HDFS, utilize o comando (em Scala) a seguir.

```
val df = {
  spark.read
    .format("parquet")
    .load("hdfs://master:8020/bigdata/pt7-hash.parquet")
}
```

Não se esqueça de salvar o seu modelo treinado no HDFS para posterior uso.

Artefatos da entrega:

- Descrição breve do modelo supervisionado escolhido
- Código-fonte
- Métricas Acurácia, PR e F-Measure do modelo sobre os dados de teste
- Arquivo README.txt com instruções para execução no cluster (para correção)

Dúvidas serão tiradas por e-mail e nos momentos de acompanhamento de projeto definidos no cronograma.

Bons estudos.