Universidade Federal do ABC
Programa de Pós-Graduação em Ciência da Computação
Disciplina de Mineração na Web e Big Data

Implementação da versão distribuída do algoritmo DCDistance usando a ferramenta Apache Spark

Patrícia Santos
patriciadiassantos@gmail.com

May 11, 2018

Introdução

No projeto foi feita a implementação da versão distribuída do algoritmo *DCDistance: A Supervised Text Document Feature extraction based on class labels*, um algoritmo de extração e redução de atributos supervisionados, utilizando a biblioteca *pyspark*.

Foi utilizada a base *DOHMH New York City Restaurant Inspection Results*[1], com dados em formato CSV, com um tamanho de aproximadamente de 133 megabytes e mais de 370 mil registros.

[1] Disponível em: https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j/data

Representação do Algoritmo

doc	dados	classe
D1	doll ball car	C1
D2	doll ball toy bird bear	C1
D3	doll ball car toy bear	C1
D4	dog toy cat doll	C2
D5	dog toy cat deer doll bird	C2

vetorização tf →

doc	atributos	classe
D1	1 1 1 0 0 0 0 0 0	C1
D2	1 1 0 1 1 1 0 0 0	C1
D3	1 1 1 1 0 1 0 0 0	C1
D4	1 0 0 1 0 0 1 1 0	C2
D5	1 0 0 1 1 0 1 1 1	C2

↓ representação do vetor de classes

3 3 2 2 1 2 0 0 0	C1
2 0 0 2 1 0 2 2 1	C2

← geração de atributos

doc	atributos	classe
D1	4.24 4.12	C1
D2	3.74 3.60	C1
D3	3.46 3.87	C1
D4	5.00 2.44	C2
D5	5.00 2.00	C2

Figure: Exemplo de aplicação do DCDistance.

Representação do Algoritmo

A implementação da versão distribuída do DCDistance foi feita em 4 etapas:

1. ingestão e extração dos dados da base,
2. geração de um Bag-of-Words com ponderação TF-IDF,
3. soma dos vetores dos documentos correspondentes às classes ’critical’ e ’not critical’ de forma a gerar um vetor representativo de cada uma,
4. criação de uma nova representação vetorial para cada documento calculando a distância euclidiana entre este documento e cada vetor representativo.

Resultados

Table: Comparação do tempo de execução entre as versões serial e distribuída.

Tempo	**Serial**	**Distribuída**
CPU time	$0,556s$	$0,465s$
Wall Time	$23m$	$1,84m$