

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



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.

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

Representação do Algoritmo

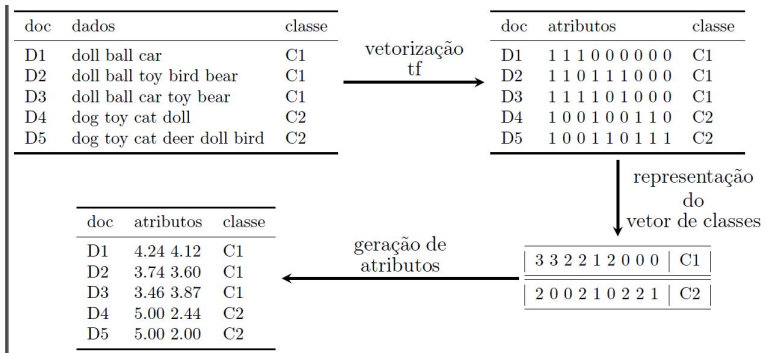


Figure: Exemplo de aplicação do DCDistance.



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.



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