Universidade Federal do ABC
Inteligência na Web e Big Data
2019.Q3

Lista de Exercícios – C2

Exercício 1

Dada a matriz binária em que cada linha é um documento e cada coluna um token:

	A	B	C	D	E	F	G
D1	0	0	1	1	0	1	0
D2	1	0	1	0	0	1	0
D3	0	1	0	1	1	0	1
D4	0	0	0	1	1	1	0
D5	0	1	0	1	1	0	0

E as seguintes funções de hash:

$$h1(x) = (3 * x + 1) \mod 13$$
$$h2(x) = (4 * x + 3) \mod 13$$
$$h3(x) = (6 * x + 2) \mod 13$$
$$h4(x) = (10 * x + 4) \mod 13$$

Calcule a assinatura LSH para 2 faixas e 2 linhas.

Exercício 2

Desenvolva um algoritmo MapReduce/Spark para o algoritmo Edit-KNN. O algoritmo consiste em encontrar objetos candidatos a serem desconsiderados para o KNN pois seus K-Vizinhos mais próximos possuem a mesma classe que ele. A entrada do algoritmo consiste nos dados no formato de tupla $(Int, [Double])$, sendo o primeiro item o índice do objeto, o segundo seus valores e o último a classe. O resultado final deve ser no formato $(Int, Bool)$,

sendo o primeiro o índice do objeto e o segundo o valor *True* se o objeto pode ser desconsiderado, e *False* caso contrário.

Exemplo de Entrada (K=3):

```
(1, [0, 0], '+'), (2, [0, 1], '+'), (3, [1, 0], '+'), (4, [1, 1], '+'),
(5, [0.5, 0.5], '+'), (6, [9, 9], '-'), (7, [8, 8], '-'), (8, [9, 8], '-'),
(9, [8, 9], '-'), (10, [8.5, 8.5], '-'),
(11, [1.1, 1.1], '-'), (12, [9.2, 9.2], '+')
```

Exemplo de Saída:

```
(1, True), (2, True), (3, True), (4, False),
(5, True), (6, False), (7, True), (8, True),
(9, True), (10, True), (11, False), (12, False)
```