$<idBand,Signature>,<idTweet,tf-idfVector>$

$<idBand,Signature>, List[<idTweet,tf-idfVector>]$

b=1

$<01010,<1,v_1>>$
$<01011,<1,v_2>>$
.
$<10011,<n,v_n>>$

b=2

$<01010,<1,v_1>>$
$<01011,<1,v_2>>$
.
$<10011,<n,v_n>>$

groupByKey

b=1

$<01010, [<1,v_1>,<2,v_2>,<3,v_3>] >$
.
$<11001, [<5,v_1>,<7,v_2>,<8,v_3>] >$

b=2

$<01010, [<1,v_1>,<2,v_2>,<3,v_3>] >$
.
$<11001, [<5,v_1>,<7,v_2>,<8,v_3>] >$

flatMapValues

$<idTweet, tf-idf Vector>$

$<1,v_1>$
$<2,v_2>$
.
.
$<n,v_n>$

b=3

$<01010,<1,v_1>>$
$<01011,<1,v_2>>$
.
$<10011,<n,v_n>>$

b=4

$<01010,<1,v_1>>$
$<01011,<1,v_2>>$
.
$<10011,<n,v_n>>$

b=3

$<01010, [<1,v_1>,<2,v_2>,<3,v_3>] >$
.
$<11001, [<5,v_1>,<7,v_2>,<8,v_3>] >$

b=4

$<01010, [<1,v_1>,<2,v_2>,<3,v_3>] >$
.
$<11001, [<5,v_1>,<7,v_2>,<8,v_3>] >$

$<1,3>$
$<5,8>$
.
.
$<9,2>$

b=5

$<01010,<1,v_1>>$
$<01011,<1,v_2>>$
.
$<10011,<n,v_n>>$

b=5

$<01010, [<1,v_1>,<2,v_2>,<3,v_3>] >$
.
$<11001, [<5,v_1>,<7,v_2>,<8,v_3>] >$

Lsh
Partitioning

Distribuited Distance
Calculation

DBSCAN
Clustering