

Input  
//key: video ID  
//value: content  
of video

Shingle the video into  
K-shingles

Sort K-shingles and  
delete the recurred  
elements

## Settings

1. number hash functions: **1000**;
2. number of bands: **20**;
3. rows of each band: 50.

Emit(band ID,  
signature matrix,  
video ID)

Hash the k-shingles and  
compute the signature for  
each band

**Mapper**

Shuffling  
& Sorting

Input  
//key: video  
pairs(i, j)  
//value: emit from  
mapper of videos

check signature value  
in each band

Candidate pairs if  
video(i, j) matched in  
any band

**Reducer**

Emit  
(videos ID of  
candidate pairs)

Output combiner  
( whether video pairs  
are emitted before )