

About the First Problem :

I think there is a problem in the Pipeline.

I recommend you the following pipeline:

////////////////////////////////

Input: sentences

Mapper 1: sentence -> (word, count)

For word in sentence:

Emit (word, 1)

Reducer1: list of (word, count) -> (word, total_count)

(count1, count2) -> count1 + count2(total_count)

Emit (word, total_count)

Mapper2: list of (word, total_count) -> (letter, (word, total_count))

For (word, total_count) in list:

Emit (word[0], (word, total_count))

Reducer2: list of (letter, (word, total_count)) -> (letter, list of (word, total_count))

Reducer3: list of (letter, list of (word, total_count)) -> list of (letter, words)

For letter, list in list of (letter, list of (word, total_count)):

List = SortByValue(list of (word, total_count))

Emit letter

For word in List:

Emit word