1. The Mapper

    a. What is the input key and value combination (give the data types for the input key and the input value)

    Answer:

    Input key: LongWritable

    Input value: Text

    b. What should the map function do to each input key value pair. Please be detailed and specific

    Answer:

    Map function first checks if the input record contains “review_id” field which confirms that it is a review record and extracts “business_id” field as the output key. The output value contains the fields namely “user_id”, “stars”, and “text”.

    If the input record does not contain “review_id”, then it confirms it as business record and extracts “business_id” as output key. In this case, output value contains the fields namely, “name”, “address”, “stars”, and “review_count”.

    c. What is the output key value pair that is sent to the reducer (give the data types for the output key and the output value)

    Answer:

    Output key: Text

    Output value: Text

2. The Reducer

    a. What are the datatypes for the key and values submitted by the mapper

    Answer:

    Input key: Text

    Input value: Iterable<Text>

    b. What will the reducer do? What type of aggregation is required here?

    Answer:

    By checking the length of each value, the reducer separates the values into two lists which are business records and review records. Length is 4 for business record and will be stored as an

output key else it is review record and will be added to the list of output values. The reducer will then iterates through the list of output values and write each one out with the output key.

c. What datatypes are needed for the key and value outputted from the reducer

Answer:

Output key: Text

Output value: Text

3. Would a Combiner be useful here? Please explain your reasoning.

Answer:

Yes, a combiner would be useful here. Even though, it performs same function as the reducer, on a smaller scale it reduces the amount of data transferred between the mapper and the reducer. The combiner can take in the same input and output key-value types as the mapper and reducer. Combiner performs a partial aggregation of the sentiment scores and count of reviews for each postal code, thus reducing the amount of data that needs to be processed by the reducer.