**Assignment 2**

Wanyi Su

Student # 301445656

Q1:

In WikipediaPopular.java, we got word\_pair (time, view\_counts). To check the popular page title, we can change the word\_pair to (time, (title, view\_counts)). Hence, we can write a new class, which is a writable that represents a pair of a Text value and a long value, and redefine the data type of word\_pair[1] from LongWritable to the new class object.

Then, we need to change the datatype of Mapper’s value output and Reducer’s value input and value output from LongWritable to the new class type. Besides, we need to reset job.setOutputValueClass() from LongWritable.class to the new class.

Also, in Reducer, we need to access values[1] for comparison to get maximum view counts.

Q2:

.map() is one-to-one operation. It applies a function to each element in an RDD and returns an RDD with the same number of output elements as the number of input elements.

.flatMap() is not one-to-one. It returns a list with all elements.

.flatMap() is more like the MapReduce concept of mapping.

Q3:

.reduce() is one-to-one operation as map(). It reduces input elements and put the result into an RDD.

.reduceByKey() merges values for each key. The number of output elements may not the same with that of input elements.

.reduceByKey() is more like the MapReduce concept of reducing.

Q4:

We can change output pair (time, (view\_count, title)) to (time, (view\_count, [title1, title2])), where we replace a single “title” with a list of tied titles.

In Python, we firstly change the datatype of “title” in (time, (view\_count, title)) to list type when splitting text. Then, we need to change the function inside reduceByKey() to return a titles list tied at the maximum view counts.