**Answers\_Assignment 4**

Wanyi Su

Q1.

.cache() saves about few seconds with reddit3 running on local, saves nearly 11s when running with reddit3 on cluster, and saves about 14s when running with reddit4 on cluster.

Cluster – reddit3

w/o cache(): 45.382s

w cache(): 34.794s

Cluster – reddit4

w/o cache(): 57.391s

w cache(): 43.301s

Q2.

If we wrongly cache the RDD before the step where we should cache it (for example, before filtering steps), this can make the cached RDD with unnecessary content to be cached in the memory. Therefore, the RDD cached can be too large in memory. While RDD swaps between disk and memory, the task can run slower.

Also, if we cache RDD that would never be used again can make the task run slower.

Q3.

If the sizes of two RDD joining together are differed a lot (one size is small while another is much bigger), broadcast join can make join process faster than actual join(). We can avoid too much shuffle (communication between nodes) and boost the join time by transforming the smaller RDD to a broadcast object, broadcasting this object to executors, and map on the larger RDD.

Q4.

If we mistake to broadcast the large RDD, which means mapping on the smaller RDD wrongly, broadcast join can be much slower. Large broadcast object transformed from the large RDD can occupy lots of memory and cause slower processing time.