COMP9313 Assignment 4

HAO FU Z5102511

Q1 MapReduce

```
def main {
    data = readfromfile(file)
    all_pairs = data.flatMap { line =>
        id, array = line.split(':')
        array_pairs = for {
            ele <- array
        } yield((min(id, ele), max(id, ele)))

        for {
            pair <- array_pairs
        } yield(pair, array_pairs.removeElement(pair))
    }

    result = pirs.reduceByKey{ case( arr_1, arr_2) =>
        arr_1.union(arr_2).distinct.sorted
    }.sortBy{ t => (t._1, t._2)}.map(x => f"(${x._1}: ${x._2})")

    result.writetoFile(output)
}
```

Sample process

input	map	reduce
1: 2, 3, 4	(1,2): (1,3), (1,4) (1,3): (1,2)(1,4) (1,4): (1,2), (1,3)	(1,2): (1,3),(1,4),(1,5) (1,3): (1,2)(1,4) (1,4): (1,2), (1,3) (2,5): (1:2)
2: 1, 5	(1,2): (1,5) (2,5): (1:2)	

Q2 LSH

(i)

ID	input	2-shingles
A	the sky is blue the sun is bright	(the, sky) (sky, is) (is, blue) (blue, the) (the, sun) (sun, is) (is, bright)
B	the sun in the sky is bright	(the, sun) (sun, in) (in, the) (the, sky) (sky, is) (is, bright)

The intersection between two set are `(the, sun), (the, sky),(sky, is), (is, bright)`. And the number of union set is 9.

$$Similar_{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{4}{9} \approx 0.44$$

Thus, the Jaccard similarity should be approximately **0.44** or **44.44%**.

(ii)

The M should be equal to the number of union set which is 9.

$$M = 9$$

Index	Shingles	A	B	5n-1 mod M	2n+1 mod M
1	(the, sky)	1	1	4	3
2	(sky, is)	1	1	9	5
3	(is, blue)	1	0	5	7
4	(blue, the)	1	0	1	9
5	(the, sun)	1	1	6	2
6	(sun, is)	1	0	2	4
7	(is, bright)	1	1	7	6
8	(sun, in)	0	1	3	8
9	(in, the)	0	1	8	1

The final min hash value should be as follow:

	A	B
$H_1 = (5n-1) mod M$	1	3
$H_2 = (2n+1) mod M$	2	1