Some Sample Questions – BLM4821 Big Data Processing and Analysis

Question: Suppose our input data to a map-reduce operation consists of integer values (the keys are not important). The map function takes an integer *i* and produces the list of pairs (*p*, *i*) such that *p* is a prime divisor of *i*.

For example, map (12) = [(2,12),(3,12)].

The reduce function is addition. That is, reduce (p,[i_1,i_2,...,i_k]) is (p,i_1+i_2+...+i_k).

Compute the output, if the input is the set of integers 15, 21, 24, 30, 49. Then, identify, in the list below, one of the pairs in the output.

a) (5,45)
b) (6,54)
c) (5,49)
d) (3,107)

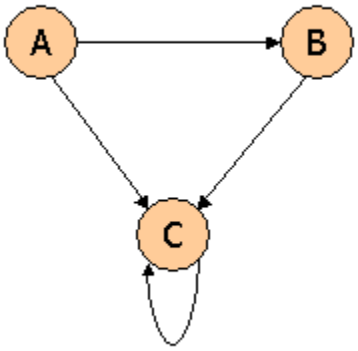
Answer: (5, 45)

Map does the following:
15 -> (3,15), (5,15)
21 -> (3,21), (7,21)
24 -> (2,24), (3,24)
30 -> (2.30), (3,30), (5,30)
49 -> (7.49)

We then group by keys, giving:
(2, [24, 30])
(3, [15, 21, 24, 30])
(5, [15, 30])
(7, [21, 49])

Finally, we add the elements of each list, giving the result (2,54), (3,90), (5,45), (7,70).

Question: Consider three Web pages with the following links:



Suppose we compute PageRank with a β of 0.7, and we introduce the additional constraint that the sum of the PageRanks of the three pages must be 3, to handle the problem that otherwise any multiple of a solution will also be a solution. Compute the PageRanks a, b, and c of the three pages A, B, and C, respectively. Then, identify from the list below, the true statement.

a) a + c = 2.595
b) a + b = 0.655
c) a + c = 2.745
d) a + b = 0.55

Answer: a + c = 2.595

The rules for computing the next value of a, b, or c as we iterate are:

a <- .3
b <- .7(a/2) + .3
c <- .7(a/2+b+c) + .3

The reason is that a splits its PageRank between b and c, while b gives all of its to c, and c keeps all its own. However, all PageRank is multiplied by .7 before distribution, and .3 is then added to each new PageRank. In the limit, the assignments become equalities. That immediately tells us a = .3. We can then use the second equation to discover b = .7*.3/2 + .3 = .405. Finally, the third equation simplifies to c = .7(.555 + c) + .3, or .3c = .6885. From this equation we get c = 2.295. It is now a simple matter to compute the subs of each two of the variables: a+b = .705, a+c = 2.595, and b+c = 2.7.

Question: Here is a matrix representing the signatures of seven columns, C1 through C7.

C1	C2	C3	C4	C5	C6	C7
1	2	1	1	2	5	4
2	3	4	2	3	2	2
3	1	2	3	1	3	2
4	1	3	1	2	4	4
5	2	5	1	1	5	1
6	1	6	4	1	1	4

Suppose we use locality-sensitive hashing with three bands of two rows each. Assume there are enough buckets available that the hash function for each band can be the identity function (i.e., columns hash to the same bucket if and only if they are identical in the band). Find all the candidate pairs, and then identify one of them in the list below.

a) C1 and C4
b) C6 and C7
c) C2 and C6
d) C3 and C6

Answer: C1 and C4. When you divide signatures into b bands of r rows, the first r rows are the first band, the next r rows are the second band, and so on. In the first band (first two rows) C1 and C4 both have (1,2), so they form a candidate pair. Also, C2 and C5 both have (2,3), so that is another candidate pair. In the second band (rows 3 and 4) we find only C1 and C6 agree, and in the third band we find C1-C3 agree and C4-C7 agree. Thus, the five candidate pairs are C1-C4, C2-C5, C1-C6, C1-C3, and C4-C7.

Question: Find the set of 2-shingles for the "document":

```
ABRACADABRA
```

and also for the "document":

```
BRICABRAC
```

Answer the following questions:

1. How many 2-shingles does ABRACADABRA have?
2. How many 2-shingles does BRICABRAC have?
3. How many 2-shingles do they have in common?
4. What is the Jaccard similarity between the two documents"?

Then, find the true statement in the list below.

a) BRICABRAC has 7 2-shingles.
b) The Jaccard similarity is 1/2.
c) BRICABRAC has 4 2-shingles.
d) BRICABRAC has 6 2-shingles.

Answer: a) BRICABRAC has 7 2-shingles.

The 2-shingles for ABRACADABRA: AB, BR, RA, AC, CA, AD, DA.
The 2-shingles for BRICABRAC: BR, RI, IC, CA, AB, RA, AC.
There are 5 shingles in common:AB, BR, RA, AC, CA.
As there are 9 different shingles in all, the Jaccard similarity is 5/9.

Question: Below is a table representing eight transactions and five items: Beer, Coke, Pepsi, Milk, and Juice. The items are represented by their first letters; e.g., "M" = milk. An "x" indicates membership of the item in the transaction.

	B	**C**	**P**	**M**	**J**
1	x		x		
2		x		x	
3	x	x			x
4			x	x	
5	x	x		x	
6				x	x
7			x		x
8	x	x		x	x

Compute the support for each of the 10 pairs of items. If the support threshold is 2, which of the pairs are frequent itemsets? Identify in the list below the pair that is NOT a frequent itemset.

a) {B,M}
b) {B,C}
c) {C,J}
d) {B,P}

Answer: {B,P}