# Optimization methods

I take several steps on this spark mission:

1. Get input file from command argument, split them with “\n”to single lines.
2. Map all lines to array that:
    a) A string indicating which number appears In this line
    b) All the array that contains this number, in which:
        i. A RID indicating which line it is
        ii. An array of all number in this line. The head of the line has been dropped because that is RID.
    c) Use one groupBy to reduce processing time.
    d) Do this step on both files, creating two RDD, with structure as follows: RDD[(String, Array[(String, Array[String])])]
    e) Prefix length decided by threshold and number of string taken to be the label of array is decided by prefix length.
3. Merge two RDD as one and try to combine those arrays whose label number are identical. But change the RID of second RDD to “-RID”, to avoid duplicate self-join.
4. Use a map to store those tuples which has been counted, to avoid duplicate calculation.