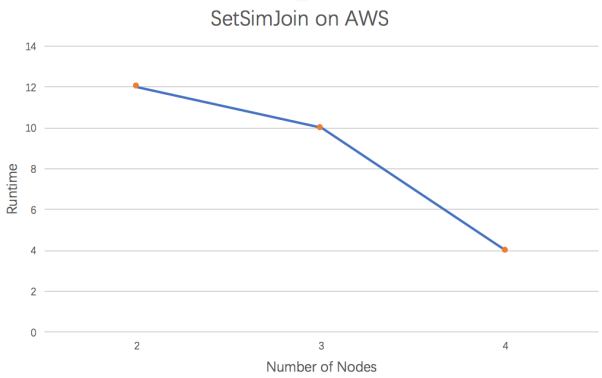
COMP9313 Project3: Optimization Report

Name: Jingjie Jin Znumber: 5085901

1. Input two files as RDD and store records in the format:
   Array[(recordId, Array[record])]
2. Sort the records according to token frequency.
3. Prefix filtering to minimize the number of prefix items emitted from the mappers by the formula:
   p = |record| − ceil(|record| * t) + 1
   And store the prefix records:
   Array[(prefix(i), Array[id + record])]
4. Join prefixArray1 and prefixArray2 of two input files.
5. Compute the Jaccard similarity between two prefixArrays by the formula:
   intersect(a,b) / union(a,b)
6. Filter the result that meets the requirement of threshold.
7. Finally convert the result format to update the output.

Runtime on AWS:



Cluster	Nodes	Runtime
1	2	12
2	3	10
3	4	4