Giraph Experimental Evaluation

1 Dataset used

1. US Elections
2. Super Tuesday
3. 120M tweets, we will use part of it

2 Algorithms evaluated

1. Native Java implementation
2. Giraph implementation
3. Optimized Giraph implementation
4. Giraph implementation using aggregator
5. Sliding window

3 Number of machines

1. On campus cluster: 3 machines
2. EC2 or SharcNet: 4 machines
3. EC2 or SharcNet: 8 machines
4. EC2 or SharcNet: 16 machines

4 Experiments' objectives

There will be three types of experiments

4.1 Choosing the paramters

1. Choosing the value of l and s for the sliding window. Using 120M dataset, the values of l and s should staisfy the response time constrain.

2. Choosing the optimal value of k, based on the quality of the detected topics using the US Election and Super Tuesday datasets. The following values will be evaluated: 100, 1000, 3000 and 5000

4.2 Validating the accuracy experiment

The algorithms that will be evaluated are

1. Native

2. Giraph implemenmtation

3. Giraph with optimization

4. Giraph with aggregator approximation

It will be performed on a on-campus cluster, which contains three machines. No sliding window will be used as these datasets have non-overlapping timeslots. And the value of k will be changed and measure its effect on the quality of the detected topics.

4.3 Scalability experiment

	US Election	Super Tuesday	120M
Native	✓	✓	✗
Giraph with optimization	✓	✓	✓
Giraph with aggregator	✓	✓	✓
Giraph with sliding window	✗	✓	✓

Cluster size: 4, 8 and 16