

# Week10 - BDO Algo

November 2, 2015

```
In [41]: from pprint import pprint
import itertools
import numpy as np

# Our buckets so far
buckets = [ ]

# Stream of data
data_stream = [1, 45, 80, 24, 56, 71, 17, 40, 66, 32, 48, 96, 9, 41, 75, 11, 58, 93, 28, 39, 77]

# Params: size of bucket and k for k-means clustering in bucket
p = k = 3

In [7]: #Group data based on p
data_stream_grouped = [data_stream[n:n + p] for n in range(0, len(data_stream), p)]
data_stream_grouped

Out[7]: [[1, 45, 80],
[24, 56, 71],
[17, 40, 66],
[32, 48, 96],
[9, 41, 75],
[11, 58, 93],
[28, 39, 77]]

In [71]: combos = [
(0,1), ()
]

def do_merge(buckets):
    ''' Check if any buckets need merging and if so, do the merge'''
    curr_size = prev_size = 0
    size_count = 0
    for bucket_idx, bucket in enumerate(buckets):
        prev_size = curr_size
        curr_size = bucket["size"]

        if prev_size != curr_size:
            size_count = 1
        else:
            size_count += 1

    if size_count == 3:
        print "!!!Need to merge!!!"
```

```

        bucket1_idx = bucket_idx - 2
        bucket2_idx = bucket_idx - 1

        bucket1 = buckets[bucket1_idx]
        bucket2 = buckets[bucket2_idx]

        cluster_orderings = [zip([0,1,2], _) for _ in itertools.permutations([0,1,2])]
        distances = []
        for idx, cluster_ordering in enumerate(cluster_orderings):
            sum_of_distances = 0
            for cluster_merge_idx, cluster_merge in enumerate(cluster_ordering):
                cluster1_idx = cluster_merge[0]
                cluster2_idx = cluster_merge[1]

                cluster1 = bucket1["clusters"][cluster1_idx]
                cluster2 = bucket2["clusters"][cluster2_idx]

                distance = np.abs(cluster1["centroid"] - cluster2["centroid"])
                sum_of_distances += distance

            distances.append(sum_of_distances)

        min_distance_idx = np.argmin(distances)

#         new_bucket = {"clusters": [], "num_points": 0}
        new_bucket = {"size": bucket1["size"] + bucket2["size"],
                      "timestamp": np.max([bucket1["timestamp"], bucket2["timestamp"]]),
                      "clusters": []}
        for merge in cluster_orderings[min_distance_idx]:
            cluster1_idx = merge[0]
            cluster2_idx = merge[1]

            cluster1 = bucket1["clusters"][cluster1_idx]
            cluster2 = bucket2["clusters"][cluster2_idx]

            new_centroid = (cluster1["num_points"] * cluster1["centroid"] + cluster2["num_
            new_bucket["clusters"].append({"centroid": new_centroid, "num_points": cluster

        # Remove old buckets
        buckets.pop(0)
        buckets.pop(0)

        # Prepend new merged bucket
        buckets.insert(0, new_bucket)

    return buckets

```

In [ ]:

```

In [72]: buckets = []
        for group_idx, group in enumerate(data_stream_grouped):
            print "=== Adding p more values to the stream ==="

            # Create k clusters of the p points in the new bucket
            clusters = []

```

```

    for point in group:
        clusters.append({"num_points": 1, "centroid": point})

    # create New bucket for these three elements
    bucket = {"size": 3, "timestamp": group_idx, "clusters": clusters}
    buckets.append(bucket)

    print "Pre-merge:"
    pprint(buckets)
    print

    # Check if we need to merge
    buckets = do_merge(buckets)

    print "Post-merge:"
    pprint(buckets)
    print

    #     if group_idx == 2: break

=== Adding p more values to the stream ===
Pre-merge:
[{'clusters': [{'centroid': 1, 'num_points': 1},
                {'centroid': 45, 'num_points': 1},
                {'centroid': 80, 'num_points': 1}],
  'size': 3,
  'timestamp': 0}]

Post-merge:
[{'clusters': [{'centroid': 1, 'num_points': 1},
                {'centroid': 45, 'num_points': 1},
                {'centroid': 80, 'num_points': 1}],
  'size': 3,
  'timestamp': 0}]

=== Adding p more values to the stream ===
Pre-merge:
[{'clusters': [{'centroid': 1, 'num_points': 1},
                {'centroid': 45, 'num_points': 1},
                {'centroid': 80, 'num_points': 1}],
  'size': 3,
  'timestamp': 0},
 {'clusters': [{'centroid': 24, 'num_points': 1},
                {'centroid': 56, 'num_points': 1},
                {'centroid': 71, 'num_points': 1}],
  'size': 3,
  'timestamp': 1}]

Post-merge:
[{'clusters': [{'centroid': 1, 'num_points': 1},
                {'centroid': 45, 'num_points': 1},
                {'centroid': 80, 'num_points': 1}],
  'size': 3,
  'timestamp': 0},

```

```

    {'clusters': [{'centroid': 24, 'num_points': 1},
                  {'centroid': 56, 'num_points': 1},
                  {'centroid': 71, 'num_points': 1}],
     'size': 3,
     'timestamp': 1}]

=== Adding p more values to the stream ===
Pre-merge:
[{'clusters': [{'centroid': 1, 'num_points': 1},
                {'centroid': 45, 'num_points': 1},
                {'centroid': 80, 'num_points': 1}],
  'size': 3,
  'timestamp': 0},
 {'clusters': [{'centroid': 24, 'num_points': 1},
                {'centroid': 56, 'num_points': 1},
                {'centroid': 71, 'num_points': 1}],
  'size': 3,
  'timestamp': 1},
 {'clusters': [{'centroid': 17, 'num_points': 1},
                {'centroid': 40, 'num_points': 1},
                {'centroid': 66, 'num_points': 1}],
  'size': 3,
  'timestamp': 2}]

!!!Need to merge!!!
Post-merge:
[{'clusters': [{'centroid': 12.5, 'num_points': 2},
                {'centroid': 50.5, 'num_points': 2},
                {'centroid': 75.5, 'num_points': 2}],
  'size': 6,
  'timestamp': 1},
 {'clusters': [{'centroid': 17, 'num_points': 1},
                {'centroid': 40, 'num_points': 1},
                {'centroid': 66, 'num_points': 1}],
  'size': 3,
  'timestamp': 2}]

=== Adding p more values to the stream ===
Pre-merge:
[{'clusters': [{'centroid': 12.5, 'num_points': 2},
                {'centroid': 50.5, 'num_points': 2},
                {'centroid': 75.5, 'num_points': 2}],
  'size': 6,
  'timestamp': 1},
 {'clusters': [{'centroid': 17, 'num_points': 1},
                {'centroid': 40, 'num_points': 1},
                {'centroid': 66, 'num_points': 1}],
  'size': 3,
  'timestamp': 2},
 {'clusters': [{'centroid': 32, 'num_points': 1},
                {'centroid': 48, 'num_points': 1},
                {'centroid': 96, 'num_points': 1}],
  'size': 3,
  'timestamp': 3}]

```

```

Post-merge:
[{'clusters': [{'centroid': 12.5, 'num_points': 2},
                {'centroid': 50.5, 'num_points': 2},
                {'centroid': 75.5, 'num_points': 2}],
  'size': 6,
  'timestamp': 1},
 {'clusters': [{'centroid': 17, 'num_points': 1},
                {'centroid': 40, 'num_points': 1},
                {'centroid': 66, 'num_points': 1}],
  'size': 3,
  'timestamp': 2},
 {'clusters': [{'centroid': 32, 'num_points': 1},
                {'centroid': 48, 'num_points': 1},
                {'centroid': 96, 'num_points': 1}],
  'size': 3,
  'timestamp': 3}]

```

=== Adding p more values to the stream ===

```

Pre-merge:
[{'clusters': [{'centroid': 12.5, 'num_points': 2},
                {'centroid': 50.5, 'num_points': 2},
                {'centroid': 75.5, 'num_points': 2}],
  'size': 6,
  'timestamp': 1},
 {'clusters': [{'centroid': 17, 'num_points': 1},
                {'centroid': 40, 'num_points': 1},
                {'centroid': 66, 'num_points': 1}],
  'size': 3,
  'timestamp': 2},
 {'clusters': [{'centroid': 32, 'num_points': 1},
                {'centroid': 48, 'num_points': 1},
                {'centroid': 96, 'num_points': 1}],
  'size': 3,
  'timestamp': 3},
 {'clusters': [{'centroid': 9, 'num_points': 1},
                {'centroid': 41, 'num_points': 1},
                {'centroid': 75, 'num_points': 1}],
  'size': 3,
  'timestamp': 4}]

```

!!!Need to merge!!!

```

Post-merge:
[{'clusters': [{'centroid': 24.5, 'num_points': 2},
                {'centroid': 44.0, 'num_points': 2},
                {'centroid': 81.0, 'num_points': 2}],
  'size': 6,
  'timestamp': 3},
 {'clusters': [{'centroid': 32, 'num_points': 1},
                {'centroid': 48, 'num_points': 1},
                {'centroid': 96, 'num_points': 1}],
  'size': 3,
  'timestamp': 3},
 {'clusters': [{'centroid': 9, 'num_points': 1},

```

```

        {'centroid': 41, 'num_points': 1},
        {'centroid': 75, 'num_points': 1}]],
    'size': 3,
    'timestamp': 4}]

=== Adding p more values to the stream ===
Pre-merge:
[{'clusters': [{'centroid': 24.5, 'num_points': 2},
                {'centroid': 44.0, 'num_points': 2},
                {'centroid': 81.0, 'num_points': 2}],
  'size': 6,
  'timestamp': 3},
 {'clusters': [{'centroid': 32, 'num_points': 1},
                {'centroid': 48, 'num_points': 1},
                {'centroid': 96, 'num_points': 1}],
  'size': 3,
  'timestamp': 3},
 {'clusters': [{'centroid': 9, 'num_points': 1},
                {'centroid': 41, 'num_points': 1},
                {'centroid': 75, 'num_points': 1}],
  'size': 3,
  'timestamp': 4},
 {'clusters': [{'centroid': 11, 'num_points': 1},
                {'centroid': 58, 'num_points': 1},
                {'centroid': 93, 'num_points': 1}],
  'size': 3,
  'timestamp': 5}]

!!!Need to merge!!!
Post-merge:
[{'clusters': [{'centroid': 20.5, 'num_points': 2},
                {'centroid': 44.5, 'num_points': 2},
                {'centroid': 85.5, 'num_points': 2}],
  'size': 6,
  'timestamp': 4},
 {'clusters': [{'centroid': 9, 'num_points': 1},
                {'centroid': 41, 'num_points': 1},
                {'centroid': 75, 'num_points': 1}],
  'size': 3,
  'timestamp': 4},
 {'clusters': [{'centroid': 11, 'num_points': 1},
                {'centroid': 58, 'num_points': 1},
                {'centroid': 93, 'num_points': 1}],
  'size': 3,
  'timestamp': 5}]

=== Adding p more values to the stream ===
Pre-merge:
[{'clusters': [{'centroid': 20.5, 'num_points': 2},
                {'centroid': 44.5, 'num_points': 2},
                {'centroid': 85.5, 'num_points': 2}],
  'size': 6,
  'timestamp': 4},
 {'clusters': [{'centroid': 9, 'num_points': 1}],
  'size': 1,
  'timestamp': 5}]

```

```

        {'centroid': 41, 'num_points': 1},
        {'centroid': 75, 'num_points': 1}],
    'size': 3,
    'timestamp': 4},
    {'clusters': [{'centroid': 11, 'num_points': 1},
                  {'centroid': 58, 'num_points': 1},
                  {'centroid': 93, 'num_points': 1}],
     'size': 3,
     'timestamp': 5},
    {'clusters': [{'centroid': 28, 'num_points': 1},
                  {'centroid': 39, 'num_points': 1},
                  {'centroid': 77, 'num_points': 1}],
     'size': 3,
     'timestamp': 6}]

```

!!!Need to merge!!!

Post-merge:

```

[{'clusters': [{'centroid': 10.0, 'num_points': 2},
               {'centroid': 49.5, 'num_points': 2},
               {'centroid': 84.0, 'num_points': 2}],
 'size': 6,
 'timestamp': 5},
 {'clusters': [{'centroid': 11, 'num_points': 1},
               {'centroid': 58, 'num_points': 1},
               {'centroid': 93, 'num_points': 1}],
 'size': 3,
 'timestamp': 5},
 {'clusters': [{'centroid': 28, 'num_points': 1},
               {'centroid': 39, 'num_points': 1},
               {'centroid': 77, 'num_points': 1}],
 'size': 3,
 'timestamp': 6}]

```

In [ ]: