Exercises on Block1:
Map-Reduce
Retrieval Evaluation
Clustering

Advanced Search Techniques for Large Scale Data Analytics

Pavel Zezula and Jan Sedmidubsky
Masaryk University
http://disa.fi.muni.cz

Map-Reduce (1) – 10min

- 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 pairs (*p, i*) such that *p* is a prime divisor of *i*
    - Example: *map*(12) = [(2,12), (3,12)]
  - The reduce function is addition
    - Example: *reduce*(*p*, [*i* , *i* , ..., *i*]) is (*p*, *i* + *i* + ... + *i*)
- Compute the output, if the input is the set of integers 15, 21, 24, 30, 49

Map-Reduce (2) – 10min

- Suppose we have the following relations R, S:

R

A	B
0	1
1	2
2	3

S

B	C
0	1
1	2
2	3

- Apply the natural join algorithm
  - Apply the Map function to the tuples of relations
  - Construct the elements that are input to the Reduce function

Map-Reduce (3) – 20min

- Design MapReduce algorithms that take a very large file of integers and produce as output:
  1) The largest integer;
  2) The average of all the integers;
  3) The same set of integers, but with each integer appearing only once;
  4) The count of the number of distinct integers in the input.

Retrieval Evaluation (1) – 10min

- The algorithm retrieves the six most relevant documents for each query. We focus on the first relevant document retrieved
  1) Determine a convenient measure for this task
  2) Compute the measure on the following four query rankings with relevant/irrelevant objects:
     - $R_1 = \{d_7, d_5, d_3, d_8, d_1\}$
     - $R_2 = \{d_5, d_6, d_3, d_2, d_4\}$
     - $R_3 = \{d_9, d_3, d_4, d_8, d_5\}$
     - $R_4 = \{d_9, d_3, d_1, d_7, d_5\}$
  3) How can be the result value interpreted?

Retrieval Evaluation (2) – 20min

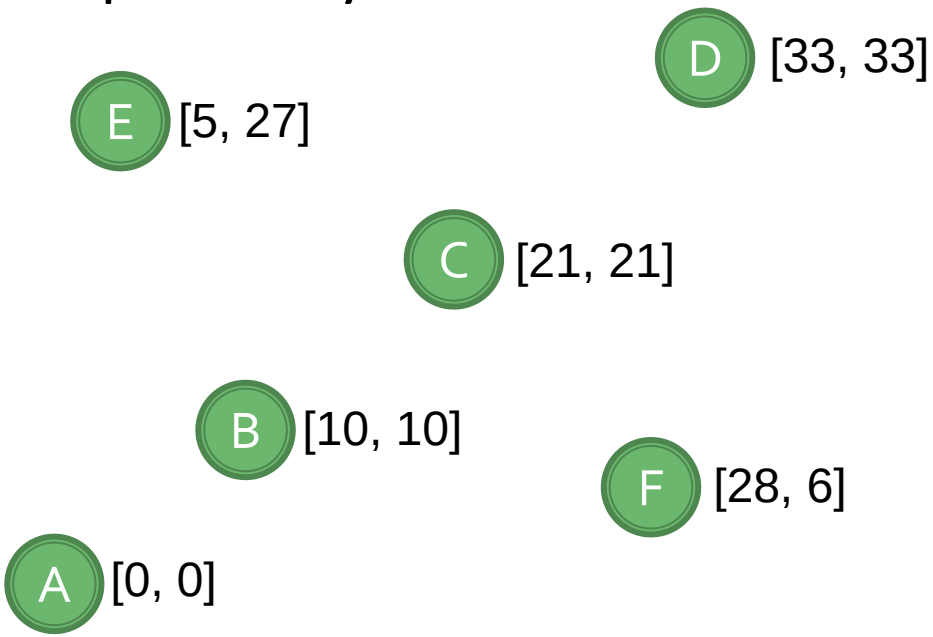
- Assume the following two rankings of documents (for some query):
  - $R_1 = \{d_7, d_5, d_3, d_8, d_1\}$
  - $R_2 = \{d_5, d_8, d_3, d_1, d_7\}$
- Based on these rankings compute:
  - Spearman rank correlation coefficient
  - Kendall Tau coefficient

Clustering (1) – 10min

- The SSE (sum squared error) is a common measure of the quality of a cluster
  - sum of the squares of the distances between each of the points of the cluster and the centroid
- Sometimes, we decide to split a cluster in order to reduce the SSE
  - Suppose a cluster consists of the following three points: (9,5), (2,2) and (4,8)
  - Calculate the reduction in the SSE if we partition the cluster optimally into two clusters

Clustering (2) – 20min

- Perform a hierarchical clustering on points A–F
  - Using the centroid proximity measure



  - There is a tie for which pair of clusters is closest. Follow both choices and identify the clusters