Recitation 12

CS435: Introduction to Big Data

GTA: Bibek R. Shrestha
Email: cs435@cs.colostate.edu

April 13, 2018
Today...

- Discussion on Programming Assignment 3
PA3 Objectives

- Estimation of PageRank values under ideal conditions.
- Estimation of the PageRank values while considering dead-end articles.
- Analysis of the above results: Creating a Wikipedia Bomb
Sample Web Graph

Figure 1. Web graph of example
PageRank using RDD

// Read input link dataset as RDD (Load data)
- val lines = spark.read.textFile(args(0)).rdd

// Create new RDD, in the form \{ (A → (B, C, D), B → (A, D), C → A, D → (B, C)) \} (Preprocessing step)
- val links = lines.map{ s => val parts = s.split("\s+")(parts(0), parts(1))).distinct().groupByKey().cache() }

// Initialize ranks of incoming pages to 1.0, to give the form \{ (A → 1.0), (B → 1.0), (C → 1.0), (D → 1.0) \}
- var ranks = links.mapValues(v => 1.0)

// Iterate to update
- for (i <- 1 to iters) {
  // Join; to give form, \{ A → ([B, C], 1.0 ), B → ([A, D], 1.0), ... \} 
  //contribs: \{ (B, _), (C, _), (A, _), (D, _), ... \}
    val contribs = links.join(ranks).values.flatMap{ case (urls, rank) =>
      val size = urls.size
      urls.map(url => (url, rank / size)) }
  //updated ranks: \{ (B, _), (C, _), (A, _), (D, _), ... \}
  ranks = contribs.reduceByKey(_ + _).mapValues(0.15 + 0.85 * _) //With taxation
}

- val output = ranks.collect()
Output

- After the PageRank for the web graph converges (in our case, 25 iterations), show top 10 pages with highest page ranks.

- Titles dataset only have titles. To get the ID, use `zipWithIndex()` function.

- Then, join the updated ranks (top 10) with Titles dataset to get the corresponding titles.
Wikipedia Bomb

- Get a sub-graph containing the word “surfing”. (First filter out the links dataset)
- Run PageRank on the sub-graph.
- At the end, the top 10 links should have “Rocky Mountain National Park” as the first page.
Thank you.