Setting up Apache Spark

- Refer to last recitation.
PageRank using RDD

Code Snippet [Link]

Complete Code (provide in binary) [../examples/src/main/java/org/apache/spark/examples]
Web Graph

Figure 1. Web graph of example
Running PageRank in IDE and in cluster

- DEMO [Java and Scala]
Wikipedia Bomb

- Filter out the Titles that contains query word (in this case, surfing)
- Get the corresponding Links for the above Titles (this is our new sub-graph).
- If you run PageRank on this sub-graph, the top ten pages will most likely contain pages with surfing word in them.
- But, for Wikipedia bomb, we want Rocky Mountain National Park at the top of that top-10 list.
Wikipedia Bomb cont.

- We need to modify the above sub-graph before running PageRank such that Rocky Mountain National Park is the page with highest page-rank value.
- When will a page will have higher page-rank value?
  - When there are many incoming links to the page, and/or
  - When the incoming links to the page have high page-rank values
Wikipedia Bomb cont.

- Code Snippet and Guide [Link]
Submission

- Source code files and executable jar
- 3 Output files: each for the three cases in PA3 (ideal, taxation and wikibomb)
- The output file should contain only top 10 pages. (Format: ID, Title, PageRank value)
- Due Date: November 6, 2018 By 5:00PM
Thank You