[Recitation 8]

Paahuni Khandelwal
Email: cs435@cs.colostate.edu

11th October, 2019
Submission

- Submission Deadline for PA2 - 25th October (by 5 pm)
- Demos signup link available - Monday and Tuesday
- Make sure to calculate TF-IDF values on an updated 1gb dataset
- During demo -
  - TF-IDF calculated on 1gb data
  - Summaries of all the articles in 1gb dataset
Programming Assignment 3

- Estimate PageRank values under ideal conditions.
- Estimate PageRank values while considering dead-end articles (Taxation)
- Create Wikipedia Bomb
Datasets

• We have two datasets: Links Dataset and Titles Dataset.

• **Links-Simple-Sorted:**
  • Each line represents the outgoing links from a page.
  • The format of the lines is:
    
    from1: to11 to12 to13
    from2: to21 to22 to23
    ...

• **Titles-Sorted:**
  • Each line represents title of Wikipedia article

• To find the page title that corresponds to integer n, just look up the n-th line in the Titles-Sorted dataset.
Top 10 pages

- After the PageRank for the web graph converges (or after 25 iterations), show top 10 pages with highest page ranks.
- Then join the updated ranks (top 10) with Titles dataset to get the corresponding titles.
Introduction to Apache Spark

- Download latest Apache Spark 2.4.4 binary (pre-build for Hadoop 2.7 and later): [Link]
- Go through its official documentation: [Link]
- Create Maven project and setup Spark on IDE to run WordCount program
  - Refer Recitation 1: We did a demo on running Hadoop job on IDE
- Try to setup Spark cluster on top of your Hadoop cluster. [Link]
Thank you