Lab #10: Project Time

Everything Data
CompSci 290.01 Spring 2014
Announcements (Tue. Mar. 25)

• Homework #10 sample solution posted
• **Project mid-term report** due today
  – You should have received email with instructions
Seat assignment

Front of D106

Course staff

Alpha Seekers

The Vets

Ookillem

3NN

Ducks

Karmalytics

DEA

Team L

Retweet!

Happiness

Lore’s Economists

The Middlemen

Climate Change

Back of D106
Format of this lab

• Discussion of Homework #10

• Project time: we are here to answer any questions you have
HW #10, Part 1:

PageRank for undirected graph

- Treat an undirected edge between $u$, $v$ as two directed edges $u \rightarrow v$ and $v \rightarrow u$
  - No dead ends in this case; code could be simplified further

- To find the top ranked nodes
  - Map all nodes + their PageRanks to a single key for one reducer to process

- PageRank ~consistent with degree rank
  - No spammers in Marvel network
HW #10, Part 1:

PageRank on full Stanford Web

My setup

• Local VM: ~20 minutes
• Amazon EMR: ~30 minutes with 8 m1.small + provisioning time

Imagine computing PageRank on the entire Web graph…
HW #10, Part 2:

Two-hop neighbors

Step 1: think as a join

- Joining $u_1 \rightarrow u_2$ and $u_2 \rightarrow u_3$ give us $(u_1, u_3)$
- Map $u \rightarrow v$ to $\langle v, u \rightarrow v \rangle$ and $\langle u, u \rightarrow v \rangle$
  - Encode them as $\langle v, (0, u) \rangle$ and $\langle u, (1, v) \rangle$
- For each $u$, a reducer generates combos of one incoming and one outgoing
  - To eliminate duplicate pairs, emit $\langle (u, v), \text{None} \rangle$ instead of $\langle u, v \rangle$

Step 2: eliminate dups

Step 3: count

961 on the tiny dataset
HW #10, Part 2:

Two-hop on full Stanford Web

My setup

• Local VM: ran out of memory
• Amazon EMR: ~1 hour with 8 m1.small + provisioning time

20613810 on full dataset
Finally

• Remember to **submit team.txt under lab 10 by midnight**