CS 686: Special Topics in Big Data

Project 3

Lecture 29

Project 3

- P3 is divided into two deliverables: an **individual** deliverable and a **team** deliverable
- Deliverable I: individual
- Deliverable II: group

Deliverable I

- Covers Spark:
  - Converting a few of our old jobs to run in Spark
  - Collecting summary statistics
  - Machine Learning
- Complete these on your own, upload to your individual GitHub repository
- Use any language you wish: Python, Java, Scala
- Fine to use the 30% sample (ask if you need a smaller dataset)

Deliverable II

- Use either Hadoop or Spark – it's up to you.
- This looks a lot like the cancelled Deliverable II from P2, but extended
  - Includes a collaboration plan, implementation, presentation
- You can choose your own dataset (including NAM, but you have to run it by me first)
- Goal: explore interesting datasets and learn something about them!

Collaboration Plan

- Choose a group, decide what dataset you'll analyze and what you hope to learn
- Include your objectives here
  - Your grade is based on whether you achieved these objectives

D2: Examples

- Do sentiment analysis on books
  - Source: project Gutenberg
- …or sentiment analysis on Enron emails
- Analyze housing and census data and fuse it with geographical attributes
- Network analysis over Wikipedia data
- Cool 3D visualizations of cloud cover over the course of a year

Restrictions

- You have to use Spark/Hadoop for the majority of your analyses
  - Maybe the output goes to a Python script that produces visualizations – that's fine
- Dataset should be at least 5 GB *
  - * Or composed of a very large number of small records

Potential Data Sources

- http://academictorrents.com
- https://www.kaggle.com/datasets
- Earth on AWS: https://aws.amazon.com/earth/
- Potentially (if large enough): http://archive.ics.uci.edu/ml/index.php

Today

- Form your groups
- Find a dataset
- Decide what you will analyze