CS341: Project in Mining Massive Datasets

Jure Leskovec
Anand Rajaraman
Jeff Ullman

CS341 Course Staff

- **Mentors:**
  - Anand Rajaraman
  - Jeff Ullman
  - Jure Leskovec
  - Rok Sosic
- **TA:**
  - Mike Chrzanowski

13 Accepted Proposals (1)

- **We accepted 13 out of 22 proposals:**
  - Parkinson Disease Classifier Using Patient Voice Recording Data
  - Automatic Anomaly Diagnosis In Distributed Systems
  - Genre-Specific Breakout Tweets
  - Character profiling and recommendation with movie database
  - Temporal Evolution of Topical Hierarchies in News Articles
  - Scientific Authorship Attribution

13 Accepted Proposals (2)

- **We accepted 13 out of 22 proposals:**
  - Predicting "Breakout Hits"
  - Cookieless Fingerprints Across Devices
  - Generalize Evolution of User Expertise Model and Identify Correlation of Expertise Evolution across Different Product Categories
  - Identifying Breakout Hits in Twitter
  - Cookieless Device Fingerprinting System
  - No Title (OpsClarity Dataset)
  - No Title (Parkinson's Dataset)

Course Logistics (o)

- **Work for the course:**
  - Class meets generally on Wednesdays
    - AWS tutorial
    - Lectures
    - Invited speakers from industry
  - Teams meet their mentors once a week

Course Logistics (1)

- **Schedule:**
  - **Mon 7-Apr: AWS Tutorial 1**
  - **Wed 9-Apr: AWS Tutorial 2**
  - Lectures by us & talks from companies
  - **Mon 5-May:** Progress Presentations
  - **Wed 7-May:** Progress Presentations
  - More lectures by us & talks from companies
  - **End of the quarter:** Final Presentations
    - **During the exam slot: Tue 6/10 12:15-3:15**
  - **Tue 10-Jun:** Final writeup due 11:59pm

Course Logistics (2)

- **Course website:**
  - http://cs341.stanford.edu
  - Lecture slides
  - Schedule/Announcements
- **For questions/clarifications use Piazza**
  - Anything you want to ask, post to Piazza
- **Collaboration using YellowDig**
  - Mike will talk about this more
- **To communicate with the course staff use**
  - cs341-win1314-staff@lists.stanford.edu

Grading

- **The grade for the course is composed of the following parts**
  - Project proposal: **10%**
  - Project midterm presentation: **20%**
  - Final project presentation: **20%**
  - Final project writeup: **50%**

Advice on conducting research

- Make sure you put in the time required (or more), work hard, consistently, independently, but also as a team player
- **Don't be afraid to be innovative and creative in your thoughts**
  - Sometimes the best innovations occur by accident
  - Don't be afraid to modify/shift the project direction

Advice on conducting research

- **Do supplemental reading**
- **Don't be afraid to make a mistake or take a risk**
  - Some of the best innovations occur from people taking risks, making errors, and learning from them
- **Take your work seriously!**

How to prepare for a meeting

- **How to prepare for a research meeting:**
  - **Update on your progress** (max 10 minutes)
    - Prepare a printout or slides with your past progress
      - **Send these out before your meeting**
    - Cover the essential results and findings. Be precise!
    - Results of failed experiments are especially useful
    - Don't try to cover every little thing you did, just focus on important results
  - **Prepare questions/ideas for further directions**
    - Bring a written list of questions or issues to each meeting
      - Mentors cannot fully answer questions that are not asked!
    - Think about what you plan to do next
  - **Take notes!**
    - Keep precise research progress and meeting notes

Next Steps: AWS

- **Each team should create an account with CCN**
- Use one shared account or use IAM for login.
- See http://aws.amazon.com/documentation/iam
- Send Mike your login details and we will give you $3,000 worth of compute time
  - We won't be able to offer more
  - **Please be careful as if you (accidentally) use more, we won't be able to revert the charges**
    - We had a team with a $20k bill!
    - Make sure you power down your instances after using them!

Next Steps: AWS

- **Next week:**

  Class meets both **Mon and Wed**

  - Monday: AWS Basics, Elastic MapReduce
  - Wednesday: EC2, Hive

Useful Resources

- Book **Mining of Massive Datasets** by Anand Rajaraman and Jeff Ullman http://i.stanford.edu/~ullman/mmds.html
  - And also
    - http://i.stanford.edu/~ullman/pub/ch11.pdf
    - http://i.stanford.edu/~ullman/pub/ch12.pdf