CS 2731
Introduction to Natural Language Processing

Session 7: Project match day, CRCD tutorial

Michael Miller Yoder

September 17, 2025

School of Computing and Information

Course logistics

- Homework 1 is **due next Thu Sep 25**
- After your group is formed today:
  - Establish a communication channel (email, Teams through Pitt, Discord, etc)
  - The project proposal, due Oct 16, is the next deliverable
  - I will release instructions for that soon

NLP and culture talk at CMU

- David Bamman from Berkeley is giving an NLP colloquium talk at the Language Technologies Institute at CMU
- **This Fri Sep 19, 12:30-1:50pm**
- Studying movies and songs with data, NLP and computer vision techniques
- Contact Michael if you're interested! We'll be meeting at 12pm on Fri at Michael's office, Sensq 6309, to walk over
- Other interesting NLP speakers: https://www.lti.cs.cmu.edu/misc-pages/lti-colloquium.html

Language Technologies Institute | Carnegie Mellon University School of Computer Science

David Bamman

School of Information, UC Berkeley

David Bamman is an associate professor in the School of Information at UC Berkeley, where he works in the areas of natural language processing and cultural analytics, applying NLP and AI to empirical questions in the humanities and social sciences. His research focuses on improving the performance of computational methods for underserved domains like literature (including LitBank and BookNLP) and developing new empirical approaches for the study of literature, film and culture. Before Berkeley, he received his PhD in the Language Technologies Institute at Carnegie Mellon University and was a senior researcher at the Perseus Project of Tufts University. Bamman's work is supported by the National Endowment for the Humanities, National Science Foundation, Mellon Foundation, and an NSF CAREER award.

Opening Up the Data-Driven Measurement of Contemporary Popular Culture

In this talk, I'll discuss how computational methods (drawing from both NLP and computer vision) can shed light on two of the most influential cultural forms of the past half-century: film and popular music. How do these media sources represent who we are and the stories we tell?

First, I'll describe recent regulatory changes at the U.S. Copyright Office that allow for large-scale text and data mining of film, and chronicle our efforts to build a collection of 2,307 films representing the top 50 movies by U.S. box office over the period 1980 to 2022, along with award nominees. Building this collection allows us to carry out several large-scale computational studies of film, including documenting the changing patterns in the representation of gender and race/ethnicity over the past 43 years (where we see an increase in diversity over the past decade). Second, I'll discuss our efforts designing computational models to measure the stories told in contemporary songs, drawing on both popular songs (from the Billboard charts) and prestigious ones (nominated for Grammy awards) over the period 1960-2024. While we might expect the 1960s (with ballad-driven folk singers like Joan Baez, Bob Dylan and Simon & Garfunkel) to be a high-water mark for narrativity, we find the opposite: narrativity has been steadily increasing over this period, largely due to the rise of the strongly narrative genres of hip hop and rap. This work illustrates a new frontier of the data-driven analysis of culture at a large scale.

Friday, September 19th
DH A302 (Doherty Hall)
12:30PM - 1:50PM
LTI Colloquium Fall 2025

Overview: Project match day

- Project match process
- CRCD resources available for the project

Project match

- Go to the spot in the room with the project printout you are most interested in working on
  - We will likely do this for several rounds
- **Goal: groups of 2-4 on projects**
  - Groups of 3 or 4 students are ideal

CRCD resources for the project

CRCD resources available for the project

- Storage space
  - 5 TB shared space for the whole class at /ix/cs2731_2025f
- CLI for running scripts through the SLURM job scheduler
- Jupyter notebooks
  - Teach cluster
  - OnDemand

Logging into the CRCD with CLI

`ssh <Pitt username>@h2p.crc.pitt.edu`

- You will need to be on the Pitt VPN (GlobalProtect app) if you are not connected to WIRELESS-PITTNET
- Your home directory only has 75 GB of storage!
- Check quota use with crc-quota
- Feel free to store project data, code, etc at class storage space **/ix/cs2731_2025f**
    - 5 TB available

Running scripts with SLURM job scheduler

- You can run scripts (like Python scripts) on the CRCD, just don't do so directly on the nodes that you log into with ssh
- Write a shell script with the commands you want and SLURM options at the top
- See the CRCD documentation: https://crc-pages.pitt.edu/user-manual/slurm/batch-jobs/

Managing Python environments on the CRCD

- See the CRCD Python documentation: https://crc-pages.pitt.edu/user-manual/applications/python/
- First load a pre-installed Python version through Lmod
  - Run module spider python to see options
  - Then module load <module>, e.g. module load python/ondemand-jupyter-python3.11
- Then create a conda environment (recommended over pip)
  - conda create --prefix=/ix/cs2731_2025f/<your_project>
  - source activate /ix/cs2731_2025f/<your_project>
  - conda install <packages>
- You can put source activate /ix/cs2731_2025f/<your_project> in your shell script for SLURM

Jupyter options on CRCD

- There are two!
- The regular JupyterHub on the teach cluster that we've been using in class is fine to use
  - If you need to install additional packages, please use your own Python environment, not the class environment
  - Feel free to select GPU options if needed
- If you need something for longer than 3 hours, see documentation on Open OnDemand (which also has an R portal): https://crc-pages.pitt.edu/user-manual/web-portals/jupyter-ondemand/
  - You request a server and they notify you when it's available
  - You can provide a path to a custom conda environment
  - Email Michael if you can't log in or have other issues