Data Mining and Analysis

The Project

CSCE 676 :: Fall 2019
Texas A&M University
Department of Computer Science & Engineering
Prof. James Caverlee

Theme: Data Mining for Social Good

Inspiration:

https://www.kdd.org/kdd2014/

http://www.dssgfellowship.org/

https://www.kdd.org/kdd2019/social-impact

Social Good?

Broadly defined

“Real-world problems in education, health, energy, public safety, transportation, economic development, international development, and more”

“harnessed for societal benefit”

Deliverables

Project Proposal (10%)

October 17 by 11:59 (Piazza)

In-Class Pitch (10%)

October 21 and 23 in class

Peer Feedback (5%)

October 27 by 11:59 (Piazza)

Project Showcase (40%)

December 2 and 4 in class

Project Website (35%)

December 4 by 11:59 (Piazza)

Project Proposal (10%)

1-2 pages in PDF; New thread on Piazza:

`Proposal: [project_name]`

Including:

- **The name of your team. Something memorable!**
- **The name of all team members.**
- **What is the need? Who wants or benefits?**
- **What data (or datasets)?**
- **What is your "data mining" toolkit? You should list specific methods you will implement.**
- **Preliminary sketch of what you hope to find.**

In-Class Pitch (10%)

3-5 minute talk

~3-5 slides

Highlight the:

- Problem
- Proposed method
- Expected result

Peer Feedback (5%)

Per individual

Post feedback to Piazza for at least 3 other teams

Use the project proposal thread

Feedback:

- Can be brief, but should be **specific** (e.g., pointer to a helpful library or paper, link to data, clarifying comment)
- Don’t just write “Awesome”

Project Showcase (40%)

In class December 2 and 4 (the week after Thanksgiving)

Either a poster session or PPT presentations (to be determined)

Project Website (35%)

Public website that chronicles your project

Google Sites, Github Pages, …

Include everything from your project showcase + more depth

Links to data, code

Showcase + website details …

Overview and Motivation: Provide an overview of the project goals and the motivation for it. Consider that this will be read by people who did not see your project proposal.

Related Work: Anything that inspired you, such as a paper, a web site, or something we discussed in class.

Initial Questions: What questions are you trying to answer? How did these questions evolve over the course of the project? What new questions did you consider in the course of your analysis?

Data: Source, scraping method, cleanup, storage, etc.

Exploratory Data Analysis: What visualizations did you use to look at your data in different ways? What are the different statistical methods you considered? Justify the decisions you made, and show any major changes to your ideas. How did you reach these conclusions?

Final Analysis: What did you learn about the data? How did you answer the questions? How can you justify your answers?

Presentation: Present your final results in a compelling and engaging way using text, visualizations, images, and videos on your project web site.

FAQ

Data? How much? Trade-off between downloading and crawling

Can we use existing libraries? YES, but … we expect you to do something beyond just calling an API “technical depth”

Ideas? Check recent editions of WSDM, KDD, ICDM, SIGIR conferences

Pretty much no constraints

…