CS224C: NLP for CSS

Project Overview

Diyi Yang
Stanford CS
Pick a question that you’re excited about

✧ What are some pressing issues in society today?
✧ Could you formulate a research question to deeply explore it?
✧ What type of textual data might be available for you to use?
✧ Which softwares or tools could you use to work on it?
✧ How do you evaluate the outcome of your project?
Key Considerations

Availability of data

Be careful in deciding whether to collect and annotate your own data

ML framework

sklearn, keras, pytorch, Tensorflow

Statistical models

R, Stata, etc.

Availability of computation
Literature Review

Conduct a thorough literature survey

A few places to check out:

Google Scholar
ACL Anthology (https://aclanthology.org/)
ICWSM Proceedings (https://www.aaai.org/Library/ICWSM/icwsm-library.php)
SIGCHI (https://dl.acm.org/conference/chi/proceedings)
CSCW (https://dl.acm.org/conference/cscw/proceedings)
PANS (https://www.pnas.org/loi/pnas)
Types of Projects

Case studies of interesting social phenomena (e.g., pandemic, fake news, hate speech, cross-culture difference, social movement, election, …)

Visualization or interpretability study of social phenomena

Applying a computational text-based method to real world problem

Compete in a predefined data science competition

New methodologies tailored to a CSS problem

Empirical survey papers

Position papers or a critic (talk to us first)
Recommendations for Successful Projects

Start early and work on it every week rather than rushing at the end

Get your data first!

Have a clear, well-defined research question (novel/creative ones ++)

Results should teach us something

Visualize results well

Divide the work between team members clearly
Common Issues

Data not available or hard to get access to
No code written for model/data processing
Team starts late
Results/Conclusion don’t say much besides that it didn’t work
   Even if results are negative or unexpected, analyze them
Resources

Computation
  Google Cloud/Google Colab

Discussion
  Come to TAs and Diyi’s Office hours
Project Ideas
Project Datasets

Fake news
https://www.kaggle.com/mrisdal/fake-news

Discourse acts
https://github.com/google-research-datasets/coarse-discourse

Authorship attribution
https://archive.ics.uci.edu/ml/datasets/
Victorian+Era+Authorship+Attribution

Stanford Large Social Network
https://snap.stanford.edu/data/
Project Datasets

Reddit Dataset
Google BigQuery

Election Dataset
https://electionlab.mit.edu/data#tools

ICWSM Datasets
Just search for a conference year like 2023 and Ctrl-F for “Datasets” papers

Misc
LOCO: The 88-million-word language of conspiracy corpus
Sample Project Ideas Part 1 *(Very well defined)*

Identifying and Combating Health Misinformation: https://nnlm.gov/misinfo

Workplace injuries: https://www.topcoder.com/challenges/30103825

Toxic comments

   https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

Modeling interactive affective responses

   https://github.com/kj2013/claff-offmychest/blob/master/README.md
Sample Project Ideas Part 2 *(Relatively open-ended)*

Understanding controversial conversations online
Identify biases in language, online community, fictions, recommender letters…
Detect hate speech or other abusive behaviors
  Who is the target? How does hate evolve over time? How to intervene?
Analyze dark patterns in ads, or campaigns
Track fake news or propaganda online
Build models for fact-checking
CS/NLP for social good (e.g., mental health, depression, climate change…)

Come up with your own idea and talk to us!