computational social media

project guidelines & schedule

daniel gatica-perez

20.03.2020
course project

defining your project idea
making progress with your project
evaluation & schedule
defining your project

a topic that you can develop between now and end May

teams of 3 people

options

your own idea: data analysis (qualitative or quantitative), machine learning, visualization, etc.
talk to me immediately if you need ideas

data

use publicly available datasets
collect your own data
checklist

1. **Research questions** (RQs)
   Pose research questions and tasks

2. **Experimental protocol**
   - Define target variables (ground-truth labels)
   - Define independent variables (features)
   - Define subject population
   - Define data collection process (questionnaires, APIs, apps)
   - Address ethics: consent, privacy
   - Address storage, security, sharing
   - Get ethical approval

3. **Data collection**
   - Get datasets according to protocol
     - Manual: typically low-scale data
     - Automated: typically large-data

4. **Processing**
   - Clean, filter, and link raw data

5. **Coding / Annotation**
   - Enrich raw data with labels
   - Get external ratings

6. **Analysis**
   - Check ground-truth quality: reliability
   - Check feature quality
   - Compute descriptive statistics
   - Perform correlation analysis
   - Apply machine learning (regression, classification, clustering)
   - Interpret the results
   - Provide answers to the RQs

7. **Communication**
   - Write report or paper
project schedule & evaluation

1. team building
   email me the list of your team members by Friday 20.03.2020

2. project definition
   next 2 weeks: discussions with each team via zoom
   send 5-slide presentation of your project by Friday 03.04.2020
   structure: title, problem, goals, approach, evaluation

3. project mid-term progress presentation on Friday 01.05.2020
   10-minute presentation per team about project progress
   format to be revisited if needed

4. final project presentation & project report on Friday 05.06.2020
   talk: 20-minute presentation + 20-min questions
   report: ACM conference short paper format (5 pages + references)
   reserve day from 09:00-16:00
examples of research datasets

YFCC100M: Flickr 100 million image dataset (creative commons)
http://www.multimediacommons.org/

ICWSM data repository (mainly Twitter)

Figure Eight (formerly Crowdflower) data for everyone
http://www.figure-eight.com/data-for-everyone

Yelp Dataset Challenge
https://www.yelp.com/dataset/challenge

Research data sharing platforms like Zenodo (https://zenodo.org)
questions?

daniel.gatica-perez@epfl.ch