Twitter Data

- **1-2 months’ archive of the complete Twitter firehose**
  - Includes all public tweets sent in that period
  - Approximately **1 billion tweets/week**
  - Tweets include metadata (user, location,...)

- **Data from January and February**
  - #egypt, #bahrain, #libya, #japan, #oscars, #sotu, ...
Project 1: Computing and Tracking Influence

• Define a measure of social influence
  – based on followers? retweets? other signals?

• Sample questions:
  – who are the current influence leaders of the Twitter network in Yemen?
  – Who are the influence leaders for digital cameras?

• To do this, you will have to geotag tweets
  – so that you find all tweets coming from Yemen
Assumptions

- For now, it’s okay if you geotag using the time zone embedded in tweets
  - “time_zone”: “Eastern Time (US & Canada)”
  - if the time zone is used in Yemen, assume the tweet comes from Yemen

- Of course, if tweet has true geo information (a small percentage of tweets has this), use that
Project 2: Build Influence Networks

• Can an influence network be constructed from the Algerian Twitter data?
  – see Project 1 on how to approximate that a tweet is from Algeria

• An influence network specifies who influences whom, which event, etc.
**Project 3: Event Detection**

- Detecting reports of particular kinds of events close to real time (e.g., natural disasters, revolutions, Justin Bieber’s haircut ...)

- Detecting indicators/precursors to particular kinds of events (e.g., forecasting possible protests)

- See Tweetbeat.com for examples of other events
Project 4: Find Failed Planned Events

• What evidence is there of failed Twitter protest communications?
  – i.e., protest calls that went out and failed to reach a “critical mass” for the protest to actually occur

• Hint: if a protest call went out to meet at Square X at time Y, and if this failed, then there usually would be tweets saying that the event at X @ Y has been canceled or something to this effect
  – on the other hand, if the protest actually occurs, there will be a different set of tweets (about the protest), with different pattern, word usage, etc.
Project 5: Tracking Events Over Time

• How to specify an event?
  – one possible solution: use a set of keywords
  – Is this good enough? e.g., “Johns Hopkins shooting”

• Summarize such a message/theme/topic/event
  – first approximation: find all tweets related to this
  – Next: find the “best” tweets, images, videos for the event and come up with an interesting visualization
Project 6

- Distinguish Bots from Human users
  - Turing Test (or Captchas for Twitter)
  - Also spam detection