Social Media & Text Analysis
lecture 1 - big data social science

Instructor: Wei Xu
Website: socialmedia-class.org
# Syllabus

<table>
<thead>
<tr>
<th>Part I</th>
<th>Computational Social Science</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part II</td>
<td>Twitter API Tutorial</td>
</tr>
<tr>
<td></td>
<td>Natural Language Processing</td>
</tr>
<tr>
<td></td>
<td>overview, language identification</td>
</tr>
<tr>
<td>Part III</td>
<td>tokenization and normalization</td>
</tr>
<tr>
<td></td>
<td>part-of-speech, chunking, named entity recognition</td>
</tr>
<tr>
<td></td>
<td>summarization, paraphrase, sentiment analysis</td>
</tr>
</tbody>
</table>
What do you expect to learn

• cutting edge research on
  - natural language processing (NLP)
  - computational social science
• popular machine learning algorithms
  - (supervised) Naïve Bayes, Conditional Random Field, Logistic Regression
  - (unsupervised) Brown Cluster, PageRank
• useful NLP tools, especially for Twitter text
• Twitter API for obtaining Twitter data
Vintage Point of View

http://wronghands1.wordpress.com
© John Atkinson, Wrong Hands
Broader Point of View

Source: http://www.conversationprism.com/
Common Features

- Posts and/or user profiles (a lot of data are text)
- Social network (explicit or implicit)
- Cross-post / user linking
- Social tagging
- Comments
- Likes / favorites / starring / …

Source: Tim Baldwin
Big Data

Big Data

- Definition (Doug Laney, 2012):
  - Volume
  - Velocity
  - Variability

- and countless other definitions …
Big Data

• the infamous definition:

“Big data is like teenage sex; everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it”.

Dan Ariely, Duke University
the location of Twitter messages and Flickr photos in New York City
Source: Eric Fischer
Impact

- Politics
- Business
- Socialization
- Journalism
- Cyber Bullying
- Productivity
- Privacy
- Emotions
- …
- and our language (!)
so my plane just crashed...
[Image: pic.twitter.com/X51BLwa5PS]

so yup [Image: pic.twitter.com/2WuLUWzpND]
Breaking News

Table 1: # of topics in each category

<table>
<thead>
<tr>
<th></th>
<th>Subcritical</th>
<th>Critical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exo.</td>
<td>31.5% (1,905)</td>
<td>54.3% (3,290)</td>
</tr>
<tr>
<td>Endo.</td>
<td>6.9% (419)</td>
<td>7.3% (444)</td>
</tr>
</tbody>
</table>

Source: Kwak, Lee, Park and Moon. “What is Twitter, a Social Network or a News Media?” WWW 2010
Military

POLICYWATCH 2186

Foreign Jihadists in Syria: Tracking Recruitment Networks

Aaron Y. Zelin

Also available in العربية

December 19, 2013

Monitoring jihadist social-media networks reveals where fighters are coming from, where in Syria they are fighting, and how best to stem their continued recruitment in countries such as Saudi Arabia, Libya, and Tunisia.
Disaster Relief

2014 Ukrainian Revolution
Social Media Use in Revolution

• to evade detection of protest activity
• spread ideas
• recruit new members
• video the progress of the movement
• organize events

Source: Kyle Jernigan & Richard Skinner
Social Media / Network

Source: André Panisson  https://www.youtube.com/watch?v=2guKJfvq4ul
Research Value

- In contrast to survey/self-report
- A probe to:
  - **real** human behavior
  - **real** human opinion
  - **real** human language use
- Easy to access and aggregate **a lot** of data
- thus **a lot** of information
Research on Social Media

- **Data Science** (part I of this course)
  - marketing, politics, finance, health, military …

- **Text analysis for social media data** (part II)
  - statistical analysis methods, off-the-shelf NLP tools, Twitter API …

- **Social media data for text analysis** (part III)
  - paraphrase, semantics, sentiment …
Data Science

Source: Drew Conway
Data Science

- is the **practice** of:
  - asking question (formulating hypothesis)
  - finding and collecting the data needed (often big data)
  - performing statistical and/or predictive analytics (often machine learning)
  - discovering important information and/or insights
Data Science

• the infamous definition:

"A data scientist is a statistician who lives in San Francisco.
Data Science is statistics on a Mac.
A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician."
Marketing

Source: Twitter Ads  https://www.youtube.com/watch?v=K8KJWoNk_Rg
User Profiling

Delighted I kept my Xmas vouchers - Happy Friday to me 😊 #shopping

Yesterday's look-my new obsession is this Givenchy fur coat! Wolford sheer turtleneck, Proenza skirt & Givenchy boots

We've already tripled wind energy in America, but there's more we can do.

Two giant planets may cruise unseen beyond Pluto - space - June 2014 - New Scientist: newsscientist.com/article/dn2571

Source: Volkova, Van Durme, Yarowsky, Bachrach
“Tutorial on Social Media Predictive Analytics” NAACL 2015
Countdown to Super Tuesday

Tweets referencing the four leading GOP candidates – January 1, 2012 to March 4, 2012

Follow @gov for other interesting data during the campaign.

January 3, 2012
Iowa Caucuses

January 10, 2012
New Hampshire Primary

January 16, 2012
Fox News SC Debate

January 21, 2012
South Carolina Primary

February 7, 2012
Colorado Caucuses
Minnesota Caucuses
Missouri Primary

February 4, 2012
Nevada Caucuses

February 11, 2012
Maine Caucuses

February 22, 2012
CNN Arizona Debate

February 28, 2012
Arizona Primary
Michigan Primary

Politics

Source: Twitter Blog
Political Polarization

Figure 1: The political retweet (left) and mention (right) networks, laid out using a force-directed algorithm. Node colors reflect cluster assignments (see § 3.1). Community structure is evident in the retweet network, but less so in the mention network. We show in § 3.3 that in the retweet network, the red cluster A is made of 93% right-leaning users, while the blue cluster B is made of 80% left-leaning users.

Source: Conover, Francisco, Flammini, Flammini. “Political Polarization on Twitter” ICWSM 2011
Movie Sales

Figure 2: Per-movie comparison of income per screen (blue, continuous line) and positive references (green, dashed line), sorted by degree of correlation. For space reasons, the X-axis shows only the movie IMDB ID.

Source: Mishne G, Glance N. “Predicting Movie Sales from Blogger Sentiment” AAAI 2006
Finance

Public Sentiment on Twitter vs Facebook Stock Price
Average Sentiment over time & market price
18 May: 10am – 1pm ET

Sentiment starts to go negative ahead of opening
Social Sentiment recovers
Sentiment declines, then bounces back
Sharp end of day sentiment drop

25 minutes later Stock starts to decline
8 mins after sentiment rebounds, stock rebounds
4 mins later, stock declines then bounces back
10 minutes later the stock declines sharply towards end of day

Wei Xu ◦ socialmedia-class.org
Socialization

The chains progress from the starting position (Omaha) to the target area (Boston) with each remove. Diagram shows the number of miles from the target area, with the distance of each remove averaged over completed and uncompleted chains.

Socialization

- “six degrees of separation” theory
  - originally presented by Hungarian author, Frigyes Karinthy, in 1929
  - first tested by psychologist Stanley Milgram in 1967 (a self-selected group of about 300)
- Twitter shows it’s 4.67 degrees of separation in 2010 (5.2m relationships)
- Facebook proves it’s 4.74 steps (4.37 for US) in 2011 (721m users)

Source: http://barnraisersllc.com/2012/04/studies-social-media-6-degrees-of-separation/
Socialization

Degrees of separation on the Twitter graph

- 4 friends away (36.92 %)
- 3 friends away (6.52 %)
- 2 friends away (0.24 %)
- 1 friend away (0.00 %)
- 7+ friends away (2.08 %)
- 6 friends away (13.06 %)
- 5 friends away (41.16 %)

Source: http://barnraisersllc.com/2012/04/studies-social-media-6-degrees-of-separation/
Politeness

Figure 3: Successful and failed candidates before and after elections. Editors that will eventually succeed (diamond marker) are significantly more polite than those that will fail (circle markers). Following the elections, successful editors become less polite while unsuccessful editors become more polite.

Table 6: Politeness and dependence. Requests made in comments posted by the question-asker are significantly more polite than the other requests. Analysis conducted on 181k requests (106k for question-askers, 75k for answer-givers).

Mood

Source: Golder & Macy. “Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures” Science 2011
Mood

Source: Golder & Macy. “Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures” Science 2011
Mood

Source: Golder & Macy. “Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures” Science 2011
Big Data $\rightarrow$ Real Data

- marketing, finance, politics
  - targeted users (customers, voters …)
  - polling
- human behavior, health, emotions
  - not reported after the fact
  - not prompted by an experimenter
  - not a small sample
- human language
  - not limited to edited text or small amount
Privacy & Ethics

- Facebook's controversial "emotional contagion" study
However

Most research to date on social media has used very shallow text processing.

(We will discuss this point in more details later in the course.)
Instructor: Wei Xu
www.cis.upenn.edu/~xwe/

Course Website: socialmedia-class.org

Thank You!