BIG DATA ANALYTICS WITH TWITTER

Course Wrap-Up, Nov 29, 2012
Marti Hearst
TWITTER VISIT

- Thursday December 6
- Meet in the South Hall, first floor Lobby: 4:45pm
- We’ll all take BART together to Civic Center station
- Gather in the Twitter Lobby, go through security at 5:45pm
- Event starts at 6pm
- 5 minutes per presentation, you will get internet access
- I will gather presentations if people are willing, but I said you can use your own computer.
THIS COURSE

• Focused on software methods for handling big data

• **Not** focused on machine learning, **not** data mining, **not** natural language processing, although we will touch on these

• Fast-paced, self-directed learning

• 3 programming assignments

• Class project
COURSE GOALS
(MARTI)

• Learn the latest algorithms and techniques for large data.

• Gain insight on the workings of social media by studying one system in detail.

• Experiment with a mix of formal lectures and in-class exercises.

• Support a range of student backgrounds.

• Let everyone in!
Course Goals (Gilad)

- Work with real data, on real problems
- Learn how it is to work in a place like Twitter
- Build something useful
- Have a good time!
WHAT IS TWITTER?

Sarah Silverman @SarahKSilverman
20 Sep
When ur relatives drive you crazy just close your eyes & pretend it’s dialogue in a woody allen movie
Details

mia farrow @MiaFarrow
@SarahKSilverman tried that. Didn't work
RT When ur relatives drive u crazy just close yr eyes & pretend its dialogue in a woody allen movie
Reply Retweet Favorite
50+ RETWEETS 50+ FAVORITES
1:12 PM - 20 Sep 11 via Twitter for iPhone - Embed this Tweet
TWITTER AS CULTURAL PHENOMENON

The New York Times

Wednesday, August 22, 2012

The Times in Print for Wednesday, August 22, 2012


THE FRONT PAGE

Ignoring Deadline to Quit, G.O.P. Senate Candidate Defies His Party Leaders
By JONATHAN WEIGMAN
Representative Todd Akin said definitively Tuesday that he would not step aside. After his comments on rape, fellow Republicans, including Mitt Romney, asked him to drop out of the Missouri Senate race.

Missouri Controversy May Endanger Republican Chances in the Fall
By JENNIFER STEINHAUER
Representative Todd Akin’s remarks on rape have focused attention on the party’s agenda on restricting abortion rights, a politically volatile topic for Mitt Romney and other candidates.

In Toll of 2,000, New Portrait of Afghan War
By JAMES DAO and ANDREW W. LEHREN
As the war in Afghanistan wears on, an analysis of troop casualties reflects the effect of the 2010 troop surge and points to the growing danger from an ostensible ally.

Dispute Over Islands Reflects Japanese Fear of China’s Rise
By MARTIN FACKLER
A chain of islands at the center of a territorial dispute have Japanese nationalists pushing their country to boldly counter the accelerating growth of China and South Korea.

The New Rise of a Summer Hit: Tweet It Maybe
By BEN SISARIO
This summer’s pop hit, Carly Rae Jepsen’s cheerfully flirtatious “Call Me Maybe,” shows how much the hitmaking machine, and the music industry, has been upended by social media.
Findings

Some 15% of online adults use Twitter as of February 2012, and 8% do so on a typical day. Although overall Twitter usage has nearly doubled since the Pew Research Center’s Internet & American Life Project first asked a stand-alone Twitter question in November 2010, the 15% of online adults who use Twitter as of early 2012 is similar to the 13% of such adults who did so in May 2011. At the same time, the proportion of online adults who use Twitter on a typical day has doubled since May 2011 and has quadrupled since late 2010—at that point just 2% of online adults used Twitter on a typical day.

The rise of smartphones might account for some of the uptick in usage because smartphone users are particularly likely to be using Twitter.

Twitter usage over time

Source: Pew Research Center’s Internet & American Life Project Winter 2012 Tracking Survey, January 20-February 19, 2012. N=2,253 adults age 18 and older, including 901 cell phone interviews. Interviews conducted in English and Spanish. Margin of error is +/-2.7 percentage points for internet users (n=1,729).
Twitter use among 18-24 year olds increased dramatically between May 2011 and February 2012, both overall and on a “typical day” basis. Twitter use within the overall population remained steady over the last year, and usage rates within most major demographic groups changed little over the same time period. The youngest adults (those between the ages of 18 and 24) are the primary exception to this trend—nearly one third of internet users in this age group now use Twitter, up from 18% in May of 2011 and 16% in late 2010.

Twitter use by those in their mid-20s to mid-40s largely leveled off in the last year after roughly doubling between late 2010 and mid 2011.

### Twitter adoption by age, 2010-2012

% of internet users in each group who use Twitter

<table>
<thead>
<tr>
<th>Age Group</th>
<th>November 2010</th>
<th>May 2011</th>
<th>February 2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>All adults</td>
<td>8%</td>
<td>13%</td>
<td>15%</td>
</tr>
<tr>
<td>18-24</td>
<td>16</td>
<td>18</td>
<td>31</td>
</tr>
<tr>
<td>25-34</td>
<td>9</td>
<td>19</td>
<td>17</td>
</tr>
<tr>
<td>35-44</td>
<td>8</td>
<td>14</td>
<td>16</td>
</tr>
<tr>
<td>45-54</td>
<td>7</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>55-64</td>
<td>4</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>65+</td>
<td>4</td>
<td>6</td>
<td>4</td>
</tr>
</tbody>
</table>

Sources: Pew Research Center’s Internet & American Life Project tracking surveys. 2012 data based on January 20-February 19, 2012 Tracking Survey. N=2,253 adults age 18 and older, including 901 cell phone interviews, margin of error is +/-2.7 percentage points based on internet users (n=1729).
Twitter data: text
Twitter data: social graph

Credit: @isaach
Twitter data: time series

#Euro2012
A summary for the action on Twitter during the European football tournament.

The Streamgraph below shows volume of Tweets during the #Euro2012 period. Click on a team’s name to see details.

Countdown to Super Tuesday
Daily new follower growth for the four leading GOP candidates — January 1, 2012 to March 6, 2012
Follow @gop for other interesting data during the campaign.
Twitter data: interest graph

Credit: @psychemedia
Combined: the pulse of the world
TWITTER: A PLATFORM FOR RESEARCH
TWITTER:
AN INSPIRATION FOR RESEARCH
TWITTER BIG DATA CASE STUDY: MOOD SWINGS

Golder & Macy, "Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures", Science, 30 September 2011: Vol. 333 no. 6051 pp. 1878-1881

• Goal: obtain data about people’s moods throughout the day and across the globe to see if and how they correspond to time of day.

• This is in contrast to self-report:
  • not prompted by an experimenter
  • not reported after the fact
  • not a small sample of undergraduates.
The Speakers:
Focus on Processing Big Data
Fundamental dynamics

- Speed & Reliability
- Top of Funnel
- Activation Rate
- Churn Rate
- Engagement

- Each of these can be captured in 1-3 metrics
- Clarify which lever(s) each project is expected to impact
- Build capability to understand the impact, and an effort on each
Twitter Overview

“Every time you see a fail whale, I get sad.” -- Raffi Krikorian
Hadoop/Pig

Bill Graham
@billgraham
Data Systems Engineer at Twitter. Apache Pig committer.
Christopher Cross fan.
San Francisco, CA, USA

Jonathan Coveney
@jco
Bad Venezuelano, decent engineer, trying to be a good human being.
Data eng @twitter. Pig committer. Beauty and happiness before all else. Monad.
SF
HDFS - writes

MapReduce Flow

Shuffle and Sort: aggregate values by keys
Why do we need Pig?

- Writing native Map/Reduce is hard
  - Difficult to make abstractions
  - Extremely verbose
    - 400 lines of Java becomes < 30 lines of Pig
  - Joins are very difficult
    - A big motivator for Pig
  - Chaining together M/R jobs is tedious
  - Decouples logic from optimization
    - Optimize Pig and everyone benefits

Basically, everything about Java M/R is painful...
Sampling the Streaming API: Twitter4J

```java
TwitterStream twitterStream = new TwitterStreamFactory().getInstance();
StatusListener listener = new StatusListener() {
    public void onStatus(Status status) {
        System.out.println("@" + status.getUser().getScreenName() + " - " +
          status.getText());
    }
};
twitterStream.addListener(listener);
twitterStream.sample();
```

Add per-status business logic to the "onStatus" function; extract tweet text, increment counters, etc.
Using the chi square score

• In a nutshell:
  – If (Observed > Expected) then the trend score is equal to:
    \[
    \frac{(O - E)^2}{E}
    \]
    else 0.

• What if E=0?
  – Add-one smoothing.
    \[
    \frac{((O + 1) - (E + 1))^2}{E + 1} = \frac{(O - E)^2}{E + 1}
    \]

• If low frequencies still dominate, use thresholds or Yates’s correction:
  \[
  \frac{(|O - E| - 0.5)^2}{E}
  \]
Duplicate and near-duplicate detection

“Obama Detours From Campaign Trail to Inspect Hurricane Isaac’s Damage”

240ef1a8 → a329b80a → 62c52e05 → ...

b63cc8e6 → 801f4d59

“Barack Obama suspends Campaign to assess Hurricane Isaac’s Damage”

240ef1a8 → 3f029e17 → 731df983 → ...

9f3e29c2
Search

Brian Larson
@larsonite
Hayes Valley, San Francisco  http://t.co/EBS1iPW

Linking the slices

Dictionary  Parallel arrays

pointer to the last posting in
Integrating Data Sets
Stephen Sorkin, Splunk
Graphs

Link Prediction

- General algorithms:
  - Most popular in a country
  - Popular movie stars, pop stars, etc
- Personalized algorithms
  - Triadic closure
  - Personalized pagerank
  - SALSA
- Data: Social graph + usage (relationship strength, interests)
Strong Triadic closure

Strong Triadic Closure Property: if the node has strong ties to two neighbors, then these neighbors must have at least a weak tie between them.
Big Learning with Graphs

Joseph Gonzalez
jegonzal@cs.cmu.edu

Yucheng Low
Aapo Kyrola
Haijie Gu
Danny Bickson
Arthur Gretton
Carlos Guestrin
Alex Smola
Joe Hellerstein
David O’Hallaron
Guy Blelloch
The Cost of the Wrong Abstraction

Log-Scale!
Assumptions of Graph-Parallel Abstractions

**Idealized Structure**

- *Small* neighborhoods
  - Low degree vertices
- Similar degree
- Easy to partition

**Natural Graph**

- *Large* Neighborhoods
  - High degree vertices
- *Power-Law* degree distribution
- *Difficult to partition*
GraphLab

- Distribute a single vertex-update
  - Move computation to data
  - Parallelize high-degree vertices

- Vertex Partitioning
  - Simple online approach, effectively partitions large power-law graphs
Recommendations, Diffusion
Similarity Measures – contd.

• Cosine similarity: \( \text{sim}(i,j) = \frac{I \cdot J}{\|I\| \cdot \|J\|} \)

• Magnitude-aware measure: \( \text{sim}(i,j) = \frac{U(i) \cap U(j)}{\sqrt{|U(i)| \cdot |U(j)|}} \)

• Jaccard, Pearson correlation …

Harry Potter movie problem?
(b) The process ends after three steps

Security, anomaly detection
Abusive Account – Median Stats

**Lifetime:** 1 day

**Followers:** 2
Users receiving a spammers content

**Tweets:** 13
Total tweets sent

---

**Social Activity**

- Regular Twitter User
- Abusive Account
Spam Value Chain

[Click Trajectories. Kirill et al. 2011]
Scalding

P. Oscar Boykin
@posco
Data Scientist @ Twitter, runner, Programming, Hadoop, Scala, co-author of @scalding
San Francisco, CA - http://pobox.com/~boykin

Argyris Zymnis
@argyris
Engineer at Twitter, Founder of AdGrok
San Francisco, CA
package com.twitter.scalding.examples

import com.twitter.scalding._

class WordCountJob(args : Args) extends Job(args) {
  args("input")
    .flatMap(line => line.getMap("word") { line : String => tokenize(_.size) }
    .groupBy(word => word).map(_._2.size)
    .write(Tsv(args("output")))
}

// Split a piece of text into individual words

def tokenize(text : String) : Array[String] = {
  // Lowercase each word and remove punctuation
  text.toLowerCase.replaceAll("[^a-zA-Z0-9\s]", "")
}
Cosine Similarity

Col, Row types (Int, Int) can be anything comparable. Strings are useful for text indices.
Spark
Making Big Data Analytics Interactive and Real-Time

Matei Zaharia, in collaboration with Mosharaf Chowdhury, Tathagata Das, Timothy Hunter, Ankur Dave, Haoyuan Li, Justin Ma, Murphy McCauley, Michael Franklin, Scott Shenker, Ion Stoica

spark-project.org
Overview

Spark is a parallel framework that provides:
» Efficient primitives for in-memory data sharing
» Simple APIs in Scala, Java, SQL
» High generality (applicable to many emerging problems)
Solution: Resilient Distributed Datasets (RDDs)

Provide an interface based on coarse-grained operations (map, group-by, join, ...)

Efficient fault recovery using lineage
  » Log one operation to apply to many elements
  » Recompute lost partitions on failure
  » No cost if nothing fails
Logistic Regression Performance

Running Time (min)

Number of Iterations

- Hadoop
- Spark

110 s / iteration

first iteration 80 s
further iterations 6 s
Observation

Batch processing models, such as MapReduce, do provide fault tolerance efficiently
  » Divide job into deterministic tasks
  » Rerun failed/slow tasks in parallel on other nodes

Idea: run streaming computations as a series of small, deterministic batch jobs
  » Same recovery schemes at much smaller timescale
  » To make latency low, store state in RDDs
CLASS ASSIGNMENTS

• Learn PIG for big data analysis
  • Analyze web logs
• Learn the Twitter Streaming API
  • Compute Trending topics
• Learn some graphing tools
  • See some Twitter graph phenomena
CLASS PROJECT

• Project Mentor from Twitter

• Presentations at both UCB and Twitter

• Wide range of topics:

  • Infrastructure, apps, analysis ...
YOUR THOUGHTS

• Something:
  • Interesting?
  • Surprising?
  • Useful?
  • Challenging?