Big Data Mining

Zhifei Zhang
Outline

• Big data vs. Small data
• Software for big data
• An example
Small Data vs. Big Data
Big Data

Attributes
- Volume
- Variety
- Velocity
- Variability
- Veracity

Types of Datasets
- High dimensional
- Sparse
- Graph
- Infinite/streaming
- Labeled

Compute Models
- Map Reduce
- Streams / Online
- Single machine in-memory
Big Data

- Acquire (data)
- Extract and Clean
- Aggregate and Integrate
- Represent
- Analyze and Model
- Interpret

- Where is the data coming from?
- How to deal with missing/incomplete/noisy data?
- How to bring datasets together?
- How to facilitate and host datasets for analysis?
- How to query, mine and model data?
- How to ensure that end users understand the data?
Algorithms for Big Data

- Classification: Bayes, clustering.
- Regression: polynomial, SVM.
- Dim. Reduction: PCA, SVD.
Example 1: Naïve Bayes

\[ P(y|x) \propto p(x|y)P(y) \]

\[
P(x|y) = \frac{C(X = x, Y = y)}{C(Y = y)}
\]

\[
P(y) = \frac{C(Y = y)}{C(Y = \text{any})}
\]

Given a data point \((x, y)\):

\[
C(Y = \text{ang}) + +
\]

\[
C(Y = y) + +
\]

\[
C(Y = y, X = x) + +
\]
Example 2: SVM

\[
\min_{\alpha} \left( \sum_{i,j} \alpha_i \alpha_j y_i y_j k(x_i, x_j) - \sum_i \alpha_i \right)
\]

Parallel-Hierarchical SVM

Given a partial data set:

- Implement SVM
- Save the local model
- Combine local models
Algorithms for Big Data

- On-line version
- High efficiency
- Less iteration
- Approximation
Software for Big Data
Software for Big Data

- Common Software
- Machine Learning Lib
- Faster Framework
Tracking Drinking Behavior from Twitter Data
Twitter Data

3/27/2014

60,000,000 Tweets

5/1/2014

Drinking

11,255,207 Tweets
LDA + Hadoop $\rightarrow$ Drinking Words

Tweets (60 million) $\rightarrow$ Mapper $\rightarrow$ Reducer

For LDA

English word list (235,886)

Unmeaning word list

Filter settings

Corpus (17,752)

Tweets (60 million)
Hadoop → Drinking Tweets

Tweets (60 million) → Mapper → Reducer

- LDA: $\beta$
- Drunk, wine, hangover, bar, ...
- Sentiment (positive/negative)
- <date, cnt>
- <time-zone, cnt>
- <source, cnt>
- <date, sentiment>
What Happened around Apr. 23\textsuperscript{rd}, 2014?

- Sinking of the South Korean ferry killed 304 passengers
- Car bomb in Kenya, killing 4 people
- Unrest in Egypt
- Unrest in Ukraine
- 6.6M earthquake in Canada
The Truth is ...

There is no data on 4/23
Source of Drunk Related Tweets

- iPhone: 50%
- Android: 18%
- Windows: < 1%
- Others: 31%
Drinkers Love iPhone More?

2014 Worldwide Smartphone Market Share (IDC)
- 82.30% Android
- 13.80% iOS
- 2.70% Windows Phone
- 1.10% Other OS

Source of Drunk Related Tweets
- iPhone: 50%
- Android: 18%
- Windows: < 1%
- Others: 31%
Drinkers Love iPhone More?

2985 Results of “wine” in App Store

248 Results of “wine” in Google play
## Location of Drinking Related Tweets

46 zones have over 10,000 drunk related tweets

<table>
<thead>
<tr>
<th>Zone</th>
<th>Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>'Eastern/Time/(US/&amp;/Canada)'</td>
<td>1349324</td>
</tr>
<tr>
<td>'Central/Time/(US/&amp;/Canada)'</td>
<td>1141305</td>
</tr>
<tr>
<td>'Pacific/Time/(US/&amp;/Canada)'</td>
<td>783803</td>
</tr>
<tr>
<td>'London'</td>
<td>583376</td>
</tr>
<tr>
<td>'Atlantic/Time/(Canada)'</td>
<td>497300</td>
</tr>
<tr>
<td>'Quito'</td>
<td>490727</td>
</tr>
<tr>
<td>'Amsterdam'</td>
<td>325605</td>
</tr>
<tr>
<td>'Arizona'</td>
<td>291043</td>
</tr>
<tr>
<td>'Mountain/Time/(US/&amp;/Canada)'</td>
<td>234895</td>
</tr>
<tr>
<td>'Hawaii'</td>
<td>198108</td>
</tr>
<tr>
<td>'Casablanca'</td>
<td>187121</td>
</tr>
<tr>
<td>'Alaska'</td>
<td>129952</td>
</tr>
<tr>
<td>'Athens'</td>
<td>102358</td>
</tr>
<tr>
<td>'Brasilia'</td>
<td>92126</td>
</tr>
<tr>
<td>'Beijing'</td>
<td>74283</td>
</tr>
<tr>
<td>'Greenland'</td>
<td>52823</td>
</tr>
<tr>
<td>'Chennai'</td>
<td>51933</td>
</tr>
<tr>
<td>'Bangkok'</td>
<td>49231</td>
</tr>
<tr>
<td>'Edinburgh'</td>
<td>44594</td>
</tr>
<tr>
<td>'Buenos/Aires'</td>
<td>43854</td>
</tr>
<tr>
<td>'Dublin'</td>
<td>38960</td>
</tr>
<tr>
<td>'Singapore'</td>
<td>38276</td>
</tr>
<tr>
<td>'Santiago'</td>
<td>37012</td>
</tr>
<tr>
<td>'Sydney'</td>
<td>31784</td>
</tr>
<tr>
<td>'Madrid'</td>
<td>31394</td>
</tr>
<tr>
<td>'Kuala/Lumpur'</td>
<td>29112</td>
</tr>
<tr>
<td>'Paris'</td>
<td>27678</td>
</tr>
<tr>
<td>'Pretoria'</td>
<td>26745</td>
</tr>
<tr>
<td>'Mexico/City'</td>
<td>26023</td>
</tr>
<tr>
<td>'Caracas'</td>
<td>25682</td>
</tr>
</tbody>
</table>
American Drink More?

Twitter user distribution

Drinking Tweet distribution
Pros & Cons

Pros:
• Through big data mining, we may find something hard to be recognized in daily life.
• Identify effect of certain event on the public.

Cons:
• How to interpret the result? (Misleading)
• Only reflect public behavior but not for individuals.
THANK YOU!