TURNING TWEETS INTO KNOWLEDGE
An Introduction to Text Analytics
Twitter

- Twitter is a social networking and communication website founded in 2006
- Users share and send messages that can be no longer than 140 characters long
- One of the Top 10 most-visited sites on the internet
- Initial Public Offering in 2013
- Valuation ~$31 billion
Impact of Twitter

- Use by protestors across the world
- Natural disaster notification, tracking of diseases
- Celebrities, politicians, and companies connect with fans and customers
- Everyone is watching!

Image of Barack Obama by Pete Souza is in the public domain. Source: Wikimedia Commons.
The Associated Press is a major news agency that distributes news stories to other news agencies.

In April 2013 someone tweeted the above message from the main AP verified Twitter account.

S&P500 stock index fell 1% in seconds, but the White House rapidly clarified.
Understanding People

• Many companies maintain online presences
• Managing public perception in age of instant communication essential
• Reacting to changing sentiment, identifying offensive posts, determining topics of interest…
• How can we use analytics to address this?
Using Text as Data

- Until now, our data has typically been
  - Structured
  - Numerical
  - Categorical

- Tweets are
  - Loosely structured
  - Textual
  - Poor spelling, non-traditional grammar
  - Multilingual
Text Analytics

• We have discussed why people care about textual data, but how do we handle it?

• Humans can’t keep up with Internet-scale volumes of data
  • ~500 million tweets per day!

• Even at a small scale, the cost and time required may be prohibitive
How Can Computers Help?

• Computers need to understand text

• This field is called Natural Language Processing

• The goal is to understand and derive meaning from human language

• In 1950, Alan Turing proposes a test of machine intelligence: passes if it can take part in a real-time conversation and cannot be distinguished from a human
History of Natural Language Processing

• Some progress: “chatterbots” like ELIZA

• Initial focus on understanding grammar

• Focus shifting now towards statistical, machine learning techniques that learn from large bodies of text

• Modern “artificial intelligences”: Apple’s Siri and Google Now
Why is it Hard?

• Computers need to understand text

• Ambiguity:
  • “I put my bag in the car. It is large and blue”
  • “It” = bag? “It” = car?

• Context:
  • Homonyms, metaphors
  • Sarcasm

• In this lecture, we’ll see how we can build analytics models using text as our data
Sentiment Mining - Apple

• **Apple** is a computer company known for its laptops, phones, tablets, and personal media players

• Large numbers of fans, large number of “haters”

• Apple wants to monitor how people feel about them over time, and how people receive new announcements.

• **Challenge:** Can we correctly classify tweets as being negative, positive, or neither about Apple?
Creating the Dataset

- Twitter data is publically available
  - Scrape website, or
  - Use special interface for programmers (API)
  - Sender of tweet may be useful, but we will ignore

- Need to construct the outcome variable for tweets
  - Thousands of tweets
  - Two people may disagree over the correct classification
  - One option is to use Amazon Mechanical Turk
Amazon Mechanical Turk

- Break tasks down into small components and distribute online

- People can sign up to perform the tasks for a fee
  - Pay workers, e.g. $0.02 per classified tweet
  - Amazon MTurk serves as a broker, takes small cut

- Many tasks require human intelligence, but may be time consuming or require building otherwise unneeded capacity
Our Human Intelligence Task

• Actual question we used:

Judge the sentiment expressed by the following item toward the software company "Apple"

• Workers could pick from
  • Strongly Negative (-2)
  • Negative (-1)
  • Neutral (0)
  • Positive (+1)
  • Strongly Positive (+2)

• Five workers labeled each tweet
Our Human Intelligence Task

• For each tweet, we take the average of the five scores.
  • “LOVE U @APPLE” (1.8)
  • “@apple @twitter Happy Programmers' Day folks!” (0.4)
  • “So disappointed in @Apple. Sold me a Macbook Air that WON’T run my apps. So I have to drive hours to return it. They won’t let me ship it.” (-1.4)

• We have labels, but how do we build independent variables from the text of a tweet?
A Bag of Words

- Fully understanding text is difficult
- Simpler approach:
  Count the number of times each word appears

- “This course is great. I would recommend this course to my friends.”

<table>
<thead>
<tr>
<th>THIS</th>
<th>COURSE</th>
<th>GREAT</th>
<th>...</th>
<th>WOULD</th>
<th>FRIENDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>...</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
A Simple but Effective Approach

- One feature for each word - a simple approach, but effective

- Used as a baseline in text analytics projects and natural language processing

- Not the whole story though - preprocessing can dramatically improve performance!
Cleaning Up Irregularities

- Text data often has many inconsistencies that will cause algorithms trouble.

- Computers are very literal by default – Apple, APPLE, and ApPLe will all be counted separately.

- Change all words to either lower-case or upper-case.

<table>
<thead>
<tr>
<th>Apple</th>
<th>APPLE</th>
<th>ApPLe</th>
<th>apple</th>
</tr>
</thead>
<tbody>
<tr>
<td>apple</td>
<td>apple</td>
<td>apple</td>
<td>3</td>
</tr>
</tbody>
</table>
Cleaning Up Irregularities

- Punctuation also causes problems – basic approach is to remove everything that isn’t a,b,…,z

- Sometimes punctuation is meaningful
  - Twitter: @apple is a message to Apple, #apple is about Apple
  - Web addresses: www.website.com/somepage.html

- Should tailor approach to the specific problem

<table>
<thead>
<tr>
<th>@Apple</th>
<th>APPLE!</th>
<th>--apple--</th>
</tr>
</thead>
<tbody>
<tr>
<td>apple</td>
<td>apple</td>
<td>apple</td>
</tr>
</tbody>
</table>

apple 3
Removing Unhelpful Terms

- Many words are frequently used but are only meaningful in a sentence - "stop words"
  - Examples: the, is, at, which…
  - Unlikely to improve machine learning prediction quality
  - Remove to reduce size of data

- Two words at a time?
  - “The Who” → “ ”
  - “Take That” → “Take”
Stemming

• Do we need to draw a distinction between the following words?
  
  argue  argued  argues  arguing

• Could all be represented by a common stem, argu

• Algorithmic process of performing this reduction is called stemming

• Many ways to approach the problem
Stemming

- Could build a **database of words** and their stems
  - **Pro**: handles exceptions
  - **Con**: won’t handle new words, bad for the Internet!

- Can write a **rule-based** algorithm
  - e.g. if word ends in “ed”, “ing”, or “ly”, remove it
  - **Pro**: handles new/unknown words well
  - **Con**: many exceptions, misses words like *child* and *children* (but would get other plurals: *dog* and *dogs*)
Stemming

- The second option is widely popular
  - "Porter Stemmer" by Martin Porter in 1980, still used!
  - Stemmers have been written for many languages

- Other options include machine learning (train algorithms to recognize the roots of words) and combinations of the above

Real example from data:

“by far the best customer care service I have ever received”

“by far the best custom care servic I have ever receiv”
Sentiment Analysis Today

• Over 7,000 research articles have been written on this topic

• Hundreds of start-ups are developing sentiment analysis solutions

• Many websites perform real-time analysis of tweets
  • “tweetfeel” shows trends given any term
  • “The Stock Sonar” shows sentiment and stock prices
Text Analytics in General

- Selecting the specific features that are relevant in the application
- Applying problem specific knowledge can get better results
  - Meaning of symbols
  - Features like number of words
The Analytics Edge

- Analytical sentiment analysis can replace more labor-intensive methods like polling

- Text analytics can deal with the massive amounts of unstructured data being generated on the internet

- Computers are becoming more and more capable of interacting with humans and performing human tasks

- In the next lecture, we’ll discuss IBM Watson, an impressive feat in the area of Text Analytics