

Day 10: Text Mining from Social Media

Paul Nulty

Essex Summer School 2013

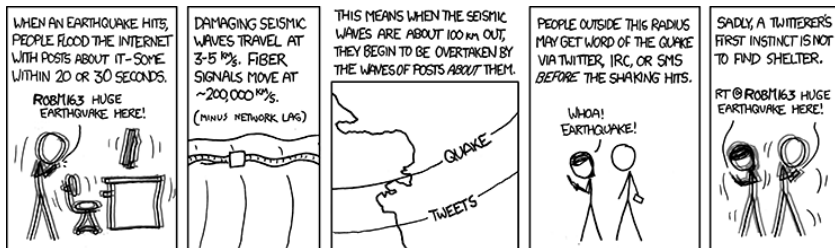
August 2, 2013

Why social media data?

- ▶ Volume: 500M registered users, 400M tweets per day (March 2013), Facebook has 1.15billion users, on average post 36 times a month — coverage and representation
- ▶ Real time — new data is available publicly immediately on current events
- ▶ Metadata — geographic location, user device, profile, timestamp and other metadata is accessible.
- ▶ Tailor-made for machine learning and data mining — lots of data, lots of types of features
- ▶ A text corpus that includes everyone: universally (sort of) accessible public *broadcast* text in electronic form
- ▶ Social network analysis: a graph of social connections

- ▶ Broadcast
 - ▶ simplex (e.g. radio, semaphore, smoke signal)
 - ▶ duplex (e.g. round-table meeting)
- ▶ Point-to-point: sender specifies receivers
- ▶ Social media allow many of these different forms of communication
- ▶ Twitter in particular is a completely new model of communication (social or news?)
- ▶ Every user is a sensor, receiver, and broadcaster — a wireless sensor network

Seismic Waves



Why not?

- ▶ Legal and ethical concerns
 - ▶ twitter is public, facebook private
 - ▶ legal issues need to catch up with the technology
 - ▶ Are EULAs (End-User License Agreement) too complex to allow 'informed consent'?
- ▶ Unconventional language use — slang, txtspk, emoticons :-)
- ▶ Sampling issues and many new methodological headaches: homographs, people tweet about interesting events
- ▶ Technological barriers
- ▶ commercial interfaces are brittle and opaque

Example applications

- ▶ Tracking disease through google search terms and social media
 - ▶ Locate tweets in urban centres
 - ▶ Uses a Porter stemmer and stopwords
 - ▶ Uses regression to learn which words are associated with flu outbreaks: from 1560 to 97 'markers'
 - ▶ Use this association to observe current outbreaks
- ▶ Predicting election outcomes or polls
- ▶ Sentiment: particularly for financial or corporate interests
- ▶ (Vasileios Lampos: www.lampos.net)
- ▶ Espionage, security, and military
- ▶ Social network analysis: a graph of social connections

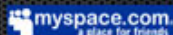
How can we access this data?

- ▶ Acronym Overload Warning.
- ▶ API: Application Programming Interface — a way for two pieces of software to talk to each other
- ▶ Twitter, facebook, google — all expose public web services
- ▶ Your software can receive (and also send) data automatically through these services
- ▶ Data is sent by `http` — the same way your browser does it
- ▶ Most services have helping code (known as a wrapper) to construct `http` requests
- ▶ both the wrapper and the service itself are called APIs
- ▶ `http` service also sometimes known as REST (REpresentational State Transfer, acronym fans) (stateless)

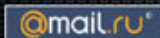
HyperText Transfer Protocol

TOP SECRET//COMINT//REL TO USA, AUS, CAN, GBR, NZL

Why are we interested in HTTP?

The Facebook logo, consisting of the word "facebook" in white lowercase letters on a blue rectangular background.The Yahoo! logo, featuring the word "YAHOO!" in red uppercase letters with a white exclamation point on a white background.The Twitter logo, featuring the word "twitter" in a light blue, lowercase, sans-serif font.The Myspace logo, featuring a blue square with a white icon of three people and the text "myspace.com" in white, with the tagline "a place for friends" in smaller white text below.

Because nearly everything a typical user does on the Internet uses HTTP

The CNN.com logo, featuring the letters "CNN" in red with a white outline, followed by ".com" in white on a black background.The Google Earth logo, featuring the word "Google" in its multi-colored font above the word "Earth" in green.The @mail.ru logo, featuring the text "@mail.ru" in white on a blue background.The Gmail logo, featuring the word "Gmail" in its multi-colored font with "by Google" and "Gmail" in smaller text below.

Anatomy of a http request

```
https://api.twitter.com/1.1/search/tweets.json?  
q=Nick+Clegg%21&since_id=24012619984051000&max_id=250126199
```

Nick Clegg! becomes Nick+Clegg%21

- ▶ Parameters to the API are encoded in the URL
- ▶ you must encode requests — spaces and non ASCII characters are replaced
- ▶ (Uniform Resource Locator; American Standard Code for Information Interchange)

cURL and wget

- ▶ It's not usually necessary to construct these kind of requests yourself
- ▶ R, Python, and other programming languages have libraries to make it easier
- ▶ Usually you will need cURL installed to access an API, wget for downloading a website
- ▶ The documentation for the API will describe the parameters that are available.

Available social media APIs

- ▶ Wikipedia: mediawiki
- ▶ Google
 - ▶ google plus
 - ▶ blogger
- ▶ reddit
- ▶ foursquare
- ▶ twitter: REST, Streaming, firehose, commercial

The twitter APIs: REST

- ▶ This is the most comprehensive API
- ▶ Returns a sample of historical data from the last 8–10 days.
- ▶ Stateless: you send a command and receive a result.
- ▶ http GET requests return information
- ▶ http POST requests upload or alter information (e.g. twitterbots)
- ▶ The manual: <https://dev.twitter.com/docs/api/1.1>
- ▶ R package : `twitter`

The twitter APIs: Streaming

- ▶ Connect to the twitter server and collect tweets as they fly by.
- ▶ The manual: <https://dev.twitter.com/docs/streaming-apis/streams/public>
- ▶ R package: streamR

Authentication

- ▶ Username and Password
- ▶ OAuth (ROauth): share a key without sharing a username and password
- ▶ IP address limitations
- ▶ Rate limitations
- ▶ Per-user and per-application

Other options

- ▶ The firehose: work with twitter and persuade them you can handle it.
- ▶ Commercial options: GNIP and Datasift

The Output: JSON and XML

- ▶ XML: eXtensible Markup Language: encodes documents in a form that is both human-readable and machine readable
- ▶ JSON : JavaScript Object Notation
- ▶ If you have a choice, you probably want JSON
- ▶ JSON uses key:value pairs, XML uses trees
- ▶ JSON is easily read into a programming language
- ▶ Sometimes known as serialization formats

And finally... the text.

- ▶ Full of spam, bots, unicode, and gibberish
- ▶ Homographs are major problem, e.g. Clegg, Cameron, Miliband
- ▶ Human behaviour is difficult to model: why does a user tweet?