CS 8803 Social Computing: Term Project Discussion

Munmun De Choudhury

munmund@gatech.edu

Week 4 | September 8, 2014

Analyzing data

Data

- Data collection, DIY
  - ~1 month to midterm, you'll get credit for this step
  - At midterm stage, show basic characteristics of the dataset
- Data collection, use a publicly available source
  - No credit for collecting this data as it is already available
  - Study characteristics of this dataset towards your project accomplishment

- https://snap.stanford.edu/data/
- http://icwsm.org/2013/datasets/datasets/

Possible datasets you can collect

- Twitter: https://dev.twitter.com/
  - REST API, Streaming API
- Using REST API, you can collect:

GET statuses/user_timeline	Returns a collection of the most recent Tweets posted by the user indicated by the screen_name or user_id parameters. User timelines belonging to protected users may only be requested when the authenticated user either "owns" the timeline or is an approved follower of the owner. The timeline...
GET statuses/home_timeline	Returns a collection of the most recent Tweets and retweets posted by the authenticating user and the users they follow. The home timeline is central to how most users interact with the Twitter service. Up to 800 Tweets are obtainable on the home timeline. It is more volatile for users that follow...
GET statuses/retweets_of_me	Returns the most recent tweets authored by the authenticating user that have been retweeted by others. This timeline is a subset of the user's GET statuses/user_timeline. See Working with Timelines for instructions on traversing timelines.
GET search/tweets	Returns a collection of relevant Tweets matching a specified query. Please note that Twitter's search service and, by extension, the Search API is not meant to be an exhaustive source of Tweets. Not all Tweets will be indexed or made available via the search interface. In API v1.1, the response...
GET statuses/sample	Returns a small random sample of all public statuses. The Tweets returned by the default access level are the same, so if two different clients connect to this endpoint, they will see the same Tweets.
GET direct_messages	Returns the 20 most recent direct messages sent to the authenticating user. Includes detailed information about the sender and recipient user. You can request up to 200 direct messages per call, up to a maximum of 800 incoming DMs. Important: This method requires an access token with RWD (read,...

Possible datasets you can collect

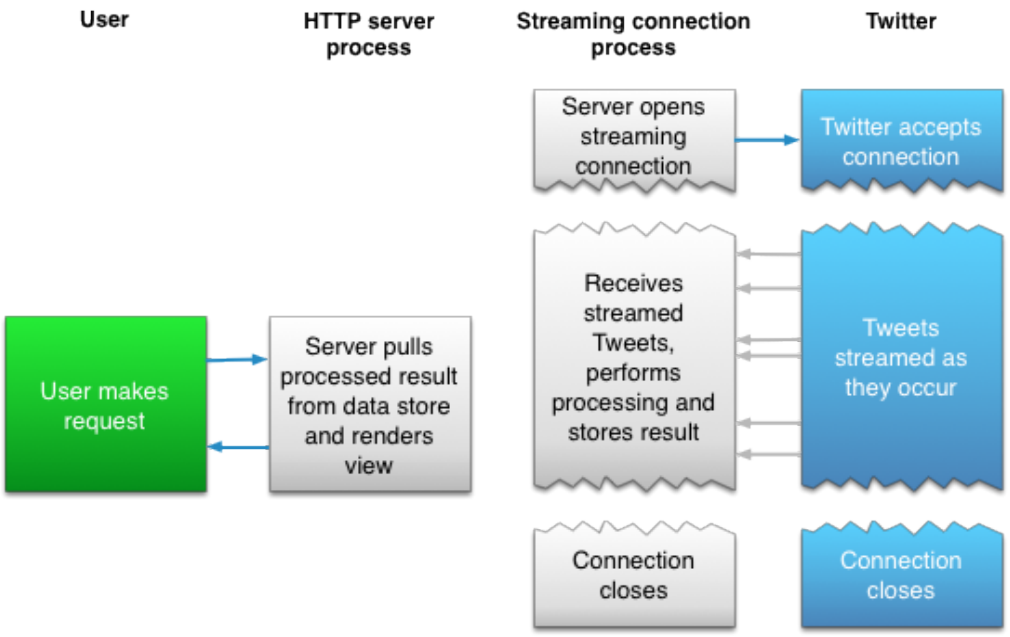
- Twitter: https://dev.twitter.com/
  - REST API, Streaming API
- Using REST API, you can collect:

GET friends/ids	Returns a cursored collection of user IDs for every user the specified user is following (otherwise known as their "friends"). At this time, results are ordered with the most recent following first — however, this ordering is subject to unannounced change and eventual consistency issues....
GET followers/ids	Returns a cursored collection of user IDs for every user following the specified user. At this time, results are ordered with the most recent following first — however, this ordering is subject to unannounced change and eventual consistency issues. Results are given in groups of 5,000 user...
GET friends/list	Returns a cursored collection of user objects for every user the specified user is following (otherwise known as their "friends"). At this time, results are ordered with the most recent following first — however, this ordering is subject to unannounced change and eventual consistency issues...
GET followers/list	Returns a cursored collection of user objects for users following the specified user. At this time, results are ordered with the most recent following first — however, this ordering is subject to unannounced change and eventual consistency issues. Results are given in groups of 20 users and...

Possible datasets you can collect

- Twitter: https://dev.twitter.com/
  - REST API, Streaming API
- Using the Streaming API, you can collect:

Public streams	Streams of the public data flowing through Twitter. Suitable for following specific users or topics, and data mining.
User streams	Single-user streams, containing roughly all of the data corresponding with a single user's view of Twitter.



Twitter data collection resources

- Python based API wrapper: (no need to parse the JSON!)
  - https://github.com/bear/python-twitter

```
>>> import twitter
>>> api = twitter.Api(consumer_key='consumer_key',
                      consumer_secret='consumer_secret',
                access_token_key='access_token',
                access_token_secret='access_token_secret')
```

```
>>> statuses = api.GetUserTimeline(screen_name=user)
>>> print [s.text for s in statuses]
```

```
>>> users = api.GetFriends()
>>> print [u.name for u in users]
```

Other libraries

https://dev.twitter.com/docs/twitter-libraries

Python

- tweepy *maintained by @applepie & more* — a Python wrapper for the Twitter API (documentation) (examples)
- python-twitter *maintained by @bear* — this library provides a pure Python interface for the Twitter API (documentation)
- TweetPony *by @Mezgrman* — A Python library aimed at simplicity and flexibility.
- Python Twitter Tools *by @sixohsix* — An extensive Python library for interfacing to the Twitter REST and streaming APIs (v1.0 and v1.1). Also features a command line Twitter client. Supports Python 2.6, 2.7, and 3.3+. (documentation)
- twitter-gobject *by @tchx84* — Allows you to access Twitter's 1.1 REST API via a set of GObject based objects for easy integration with your GLib2 based code. (examples)
- TwitterSearch *by @crw_koepp* — Python-based interface to the 1.1 Search API.
- twython *by @ryanmcgrath* — Actively maintained, pure Python wrapper for the Twitter API. Supports both normal and streaming Twitter APIs. Supports all v1.1 endpoints, including dynamic functions so users can make use of endpoints not yet in the library. (docs)
- TwitterAPI *by @boxnumber03* — A REST and Streaming API wrapper that supports python 2.x and python 3.x, TwitterAPI also includes iterators for both API's that are useful for processing streaming results as well as paged results.
- Birdy *by @sect2k* — "a super awesome Twitter API client for Python"

Other libraries

https://dev.twitter.com/docs/twitter-libraries

PHP

- tmhOAuth *by @themattharris* (examples)
- twitteroauth *by @abraham* (documentation)
- 140dev Twitter Framework *by @140dev* — The goal of this open source framework is to provide a greatly simplified interface to the Twitter Streaming API. The current version provides a tweet aggregation database, and a plugin for tweet display on any Web page. The framework is written in PHP and Javascript, and uses the MySQL database for storage. The Phirehose library is used for the connection to the Streaming API. The extensive documentation supplied with this library makes it an educational tool for anyone new to tweet aggregation and display on a website.
- Twitter API: Engagement Programming *by @140dev* — PHP Code examples from Adam Green's book Twitter API: Engagement Programming.
- codebird-php *by @myx* — a Twitter library in PHP. (documentation)
- CodeIgniter-Twitter-Search-Library *by @elliottlan* — Search for certain tweets using keywords specified in a mysql database using the search api, streaming api or both at the same time. Written in php for codeigniter
- Zend Framework 1.12.2 *maintained by @zend* — a PHP framework that includes support for Twitter API v1.1
- freebird-php *maintained by @corbanb* — a app-only auth interface for PHP
- PHP OAuth API *maintained by @manuellemos* — This is a PHP class that can implements OAuth authorization and call APIs with OAuth tokens. It also supports two-legged auth. (documentation) (examples)
- Twitter-API-PHP *maintained by @j7mbo* — A very simple and actively maintained wrapper for the Twitter v1.1 REST API that utilises cURL for authenticated requests. Single file include, only a few methods to call - "as simple as you can get." (examples)
- TwitterOAuth *by @Ricard0Per* — a simple PHP library for API v1.1

Other libraries

https://dev.twitter.com/docs/twitter-libraries

.NET

- LINQ2Twitter by *@joemayo* (examples)
- Spring.NET Social extension for Twitter *by SpringSource* — A Spring.NET Social extension with connection support and an API binding for Twitter.
- TweetSharp *by @danielcrenna* — A .net library for Twitter API access
- Tweetinvi *maintained by Linvi* — a Twitter .Net C# API which has for mission to simplify the development of application for Twitter in C#. The streaming API has been used on research projects and collected around 3.2 million tweets a day. The twitter API has been created to be easy to implement new functionality and currently provide access to most of the REST 1.1 functionalities. (documentation a>)
- Crafted.Twitter by *@martbrow* — A caching v1.1 API compatible solution - with implementations for both ASP.Net Web Forms and MVC. Making it easy to include tweets in your website.

Java

- Twitter4J *by @yusuke* — a Twitter API library (Java platform > v1.4.2, Android and GAE ready)

Javascript / node.js

Reminder: It is strongly discouraged to use OAuth 1.0A with client-side Javascript.

- TwitterJSClient *by @BoyCook* — Twitter client library written in Javascript and packaged as a node module
- user-stream *by @AivisSilins* — a simple Node.js User streams client

Possible datasets you can collect

- Tumblr: http://www.tumblr.com/docs/en/api/v2
  - Blog methods
  - User methods
  - Tagged methods
- Reddit: http://www.reddit.com/dev/api
  - Listings: /hot, /new, /random, /top
  - User info
- Flickr: https://www.flickr.com/services/api/
- YouTube: https://developers.google.com/youtube/
- Instagram: http://instagram.com/developer/

Possible datasets you can collect

- Craigslist
- Airbnb: https://www.airbnb.com/
- LiveJournal: http://www.livejournal.com/developer/
- Of course, your own data!
  - Facebook (https://www.facebook.com/help/1311128970284 67)
  - Google talk chat logs
  - Outlook or other email clients that let you download your data

- Other useful resources:
  - https://www.census.gov/
  - http://finance.yahoo.com/q?s=API
  - http://www.google.com/trends/

Things to be careful about…

- Register your "application" or project
- Authentication (OAuth preferred)
- Rate limits
- Navigating and using the API
  - Pagination, multiple calls, going back in time
  - Twitter gives a nice resource on "Working with Timelines": https://dev.twitter.com/docs/working-with-timelines
- Archiving data

Analytic tools

- For text, primarily Python's nltk library
- A text analysis library called Pattern (http://www.clips.ua.ac.be/pages/pattern)

- For network analysis, R has many libraries and functions; also the Stanford toolkit SNAP (https://snap.stanford.edu/)

Sample ideas

- Geo-locations obtained from geotagged tweets in a city throughout the course of a day, and correlating with traffic patterns in the city

Sample ideas

- In a sample of Twitter users, compare the behavior of users who have a non-empty profile photo, and ones who have the "Twitter egg" as display. E.g., see how the two cohorts communicate

Sample ideas

- Seed a twitter social graph crawl starting with your userid. Crawl your friends and followers and their tweets, followed by crawling their friends and followers. Apply community detection (clustering) to see clustered individuals, align communities to topics

Sample ideas

- From Instagram, collect pictures which are tagged with food names. See what food are popular in which geographic regions (based on geo information of the photo), and what food is preferred what time of the day

Sample ideas

- Crawl listings in Atlanta and other major cities in GA from Airbnb and correlate prices of listings to socio-economic/income levels in Georgia given by US Census

Sample ideas

- Opinion (in)consistency in resubmissions on reddit: for a sample of images which have been submitted multiple times on reddit, analyze the similarities and differences in audience reaction across multiple submissions

Sample ideas

- Based on the Enron email dataset, analyze the linguistic tone of "high authority" users i.e., users with high rank in the company

Sample ideas

- Crawl a random sample of videos from YouTube in the "Music" category. Following the "related" link in video pages, and analyze how the community responds to different types of music, e.g., in terms of #views, #comments etc.

Sample ideas

- Pick a few controversial topics such as “gun control”, and crawl the edit history of their corresponding Wikipedia pages. Pick another few non-controversial topics such as “sushi” and compare how their edit history patterns differ

Sample ideas

- Collect tweets with mentions of the Ebola virus, and correlate with Google search trends from Google Trends. Analyze tweets for the nature of content, possibly by handcoding of topical clusters. See if Twitter or search precedes the other

Grading criteria

Identifying an interesting problem.
Addressing it with social data.
Quality of presenting solutions in report.
Use of appropriate/adequate statistical analysis tools.

Building a tool

Sample ideas

- Emotion search of restaurants: A tool that allowed searching for restaurants on Yelp based on a mood type. E.g., “show me sushi restaurants in Atlanta with the mood “exciting”.

Sample ideas

- Social expert recommender: A tool that proposes expert Twitter users on a pre-selected set of topics. Approach may combine how frequently they talk about the topic, how many the user is Rted or mentioned, and the ratio of their followers to followees

Sample ideas

- Event Analyzer: A tool that utilizes streaming tweets on a couple of events of choice, and shows public opinion trend in real-time

Sample ideas

- Social photo collage: A tool that pulls photos shared by people you follow on Twitter, and associates metadata information (e.g., RTs) to them to show dynamics

Sample ideas

- Social circles on Facebook/Twitter: Based on content match, show circles of your friends on Facebook or Twitter in a tool. For each circle, allow the user to see what content is being shared and what content is popular

Grading criteria

- **Actually having something working.**
  Quality of problem statement.
  Thoughtful use of course readings in framing the motivation and in implementing your ideas.
  Insight into social computing issues.
- You need to demo your tool in the presentation(s) during midterm and finals.

Instagram

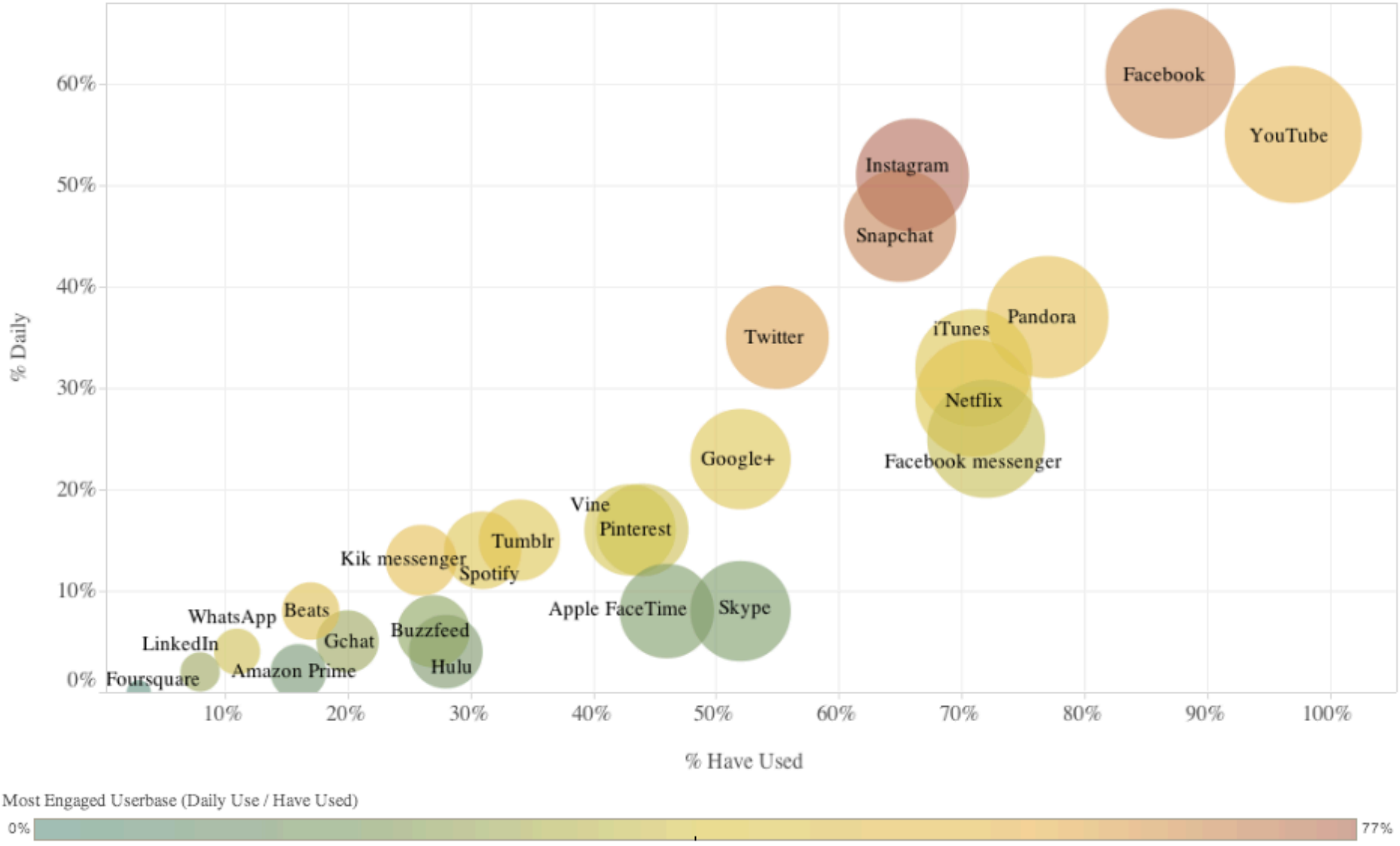
WINNER: Most Engaged Users

Facebook

WINNER: Most Daily Users

YouTube

WINNER: Most Widespread Usage



Preliminary pitches

Next class

- Monday 9/10 (topic: Statistics Review)
- **No assigned readings,** but you need to start discussing ideas on your term project with your teammates if you didn't do a pitch today

Due Dates

- Team project proposal due: September 15
  - Use the time until September 15 to come to me and discuss project ideas
  - Feel free to use/motivate your thoughts about your project based on today's class discussion
  - Once you are final, email me and TA and the project proposal (2-3 sentences)
  - Project proposals are for us to have a archived record of your project; it will NOT be graded
- Assignment I due: September 15, 3:30pm on T-square