

Module 4: GATE and Social Media

Part 2: Gathering Social Media Data

Social media sites

Twitter, LinkedIn, Facebook

Twitter has varied uptake per country:

- Low in Denmark, Germany (Facebook is preferred)
- Medium in UK, though often complementary to Facebook
- High in USA

Networks have common themes:

- Individuals as nodes in a common graph
- Relations between people
- Sharing and privacy restrictions
- No curation of content
- Multimedia posting and re-posting

Other features: topics, closed groups, moderation, liking, media, groups, person discovery ..

1. Twitter

- Opened in 2006 as a short message blogging service
- Allows 'subscription' to interesting accounts
- Anyone can post, most messages are public
- Messages are <280 characters (used to be <140)
- Posts can come from PC, mobile, SMS, iPad etc
- Specialised markup: #hashtags and @mentions
- Has grown extremely popular
 - 330 million active users per month; over 500 million tweets a day

Public relations

Barack Obama

We just made history. All of this happened because you gave your time, talent and passion. All of this happened because of you. Thanks

Celebrity worship

Kidrauhl ♡

“One day you will forget me. You have a husband and be a mother. But I will never forget you, My Beliebers.” - Justin Bieber ♡

Broadcasting & Activism

Ars Technica

SOPA opponents unveil "Digital Bill of Rights"

[http://arstechnica.com/tech-policy/20 ...](http://arstechnica.com/tech-policy/20...) by @nathanmattise

Social uses

「ジャム」 Jam Gregory

@RyanBibby: lots of people have been talking about it - need to make sure I watch it! Love @ninaconti, got a signed DVD at #EdFringe :D

Conversations/Customer Support

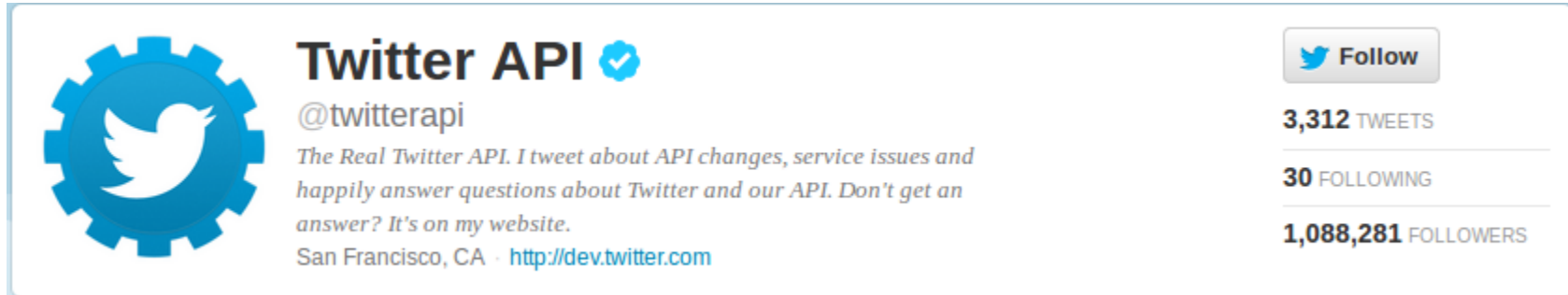


GA **Greater Anglia** @greateranglia 28 May
@adrianmelrose @stephenfry Hi, sorry that the wifi is not working, what service are you on please? GK
Collapse ← Reply ↻ Retweet ★ Favorite
8:55 AM - 28 May 12 via HootSuite · Details



 **Stephen Fry** @stephenfry 28 May
@greateranglia 8:30 to Norwich
Hide conversation ← Reply ↻ Retweet ★ Favorite
8:59 AM - 28 May 12 via Tweetbot for iOS · Details

Twitter User Profiles



The image shows a screenshot of a Twitter profile for the user @twitterapi. The profile picture is a blue gear with a white Twitter bird inside. The name is "Twitter API" with a verified badge. The handle is "@twitterapi". The bio reads: "The Real Twitter API. I tweet about API changes, service issues and happily answer questions about Twitter and our API. Don't get an answer? It's on my website." The location is "San Francisco, CA" and the website is "http://dev.twitter.com". On the right side, there is a "Follow" button, "3,312 TWEETS", "30 FOLLOWING", and "1,088,281 FOLLOWERS".

- Picture
- Name
- Location
- Website
- Bio

What is Twitter? (2)

- Interest-graph social media
Following/follower relationship is typically not bi-directional
- 77.6% of user connections are not reciprocated (Kwak 2010)
A large graph in which mutual follower/following relationships comprise the edges
Twitterers can 'retweet' one another, so information propagates via the graph quickly
- RTs typically contain links to interesting content
Users can be organised in lists, which introduces groupings

Example Tweet metadata in JSON

```
{  "contributors":null,
  "text":"Automotive RDFa (a horribly researched SEO article on RDFa/Microformats):
http://ow.ly/5JSoS #somanerrorsitsfunny",
  "geo":null,
  "retweeted":false,
  "in_reply_to_screen_name":null,
  "truncated":false, "entities":{"urls":[{"expanded_url":null,"indices":
[74,92],"url":"http://ow.ly/5JSoS"}], "hashtags":
[{"text":"somanerrorsitsfunny","indices":[93,114]}],
"user_mentions":[]},
  "in_reply_to_status_id_str":null,
  "id":94029193863639040,
  "source":"<a href=\"http://www.hootsuite.com\" rel=\"nofollow\">HootSuite</a>",
  "in_reply_to_user_id_str":null,
  "favorited":false,
  "in_reply_to_status_id":null,
  "retweet_count":0,
  "created_at":"Thu Jul 21 13:01:21 +0000 2011",
```

Example Tweet metadata in JSON

```
"user":{"location":"Blacksburg, VA",
...,
"statuses_count":2404,
"lang":"en",
"id":20446311,
...,
"description":"Text from the user profile (max 160 chars)", ...,
"name":"User Name", ...,
"created_at":"Mon Feb 09 16:33:16 +0000 2009",
"followers_count":1239,
"geo_enabled":false, ...,
"url":"The author's URL (optional)",
"utc_offset":-21600,
"time_zone":"Central Time (US & Canada)", ...,
"friends_count":160, ...,
"screen_name":"twitter-user-name", ...,
"listed_count":189, ...
}, ...
```

Embedded user information can become out-of-sync, if the user changes it later

How to get tweets?

The REST API allows access to timelines, tweeting, following, etc.

- REST/JSON based
- Requires registration, and developer / app keys
- Contains access to what was previously the Search API
- Core entities: tweets, users, entities, places
- Heavily rate-limited

The Streaming API streams tweets in real time

- Various strengths available, from 1% to 100% sample (~\$1M p.a.)
- May be filtered by language, location, user view, hashtag, search term
-

See <https://dev.twitter.com/docs> and note that the JSON returned can differ across endpoints and with different parameters

Getting tweets in the cloud

GATE Cloud tools make getting tweets possible without any programming

- Makes use of the streaming twitter API
- Tweets are stored in real time
- Filter by keyword, username, location and language
- Tweets can be downloaded or stored in the cloud

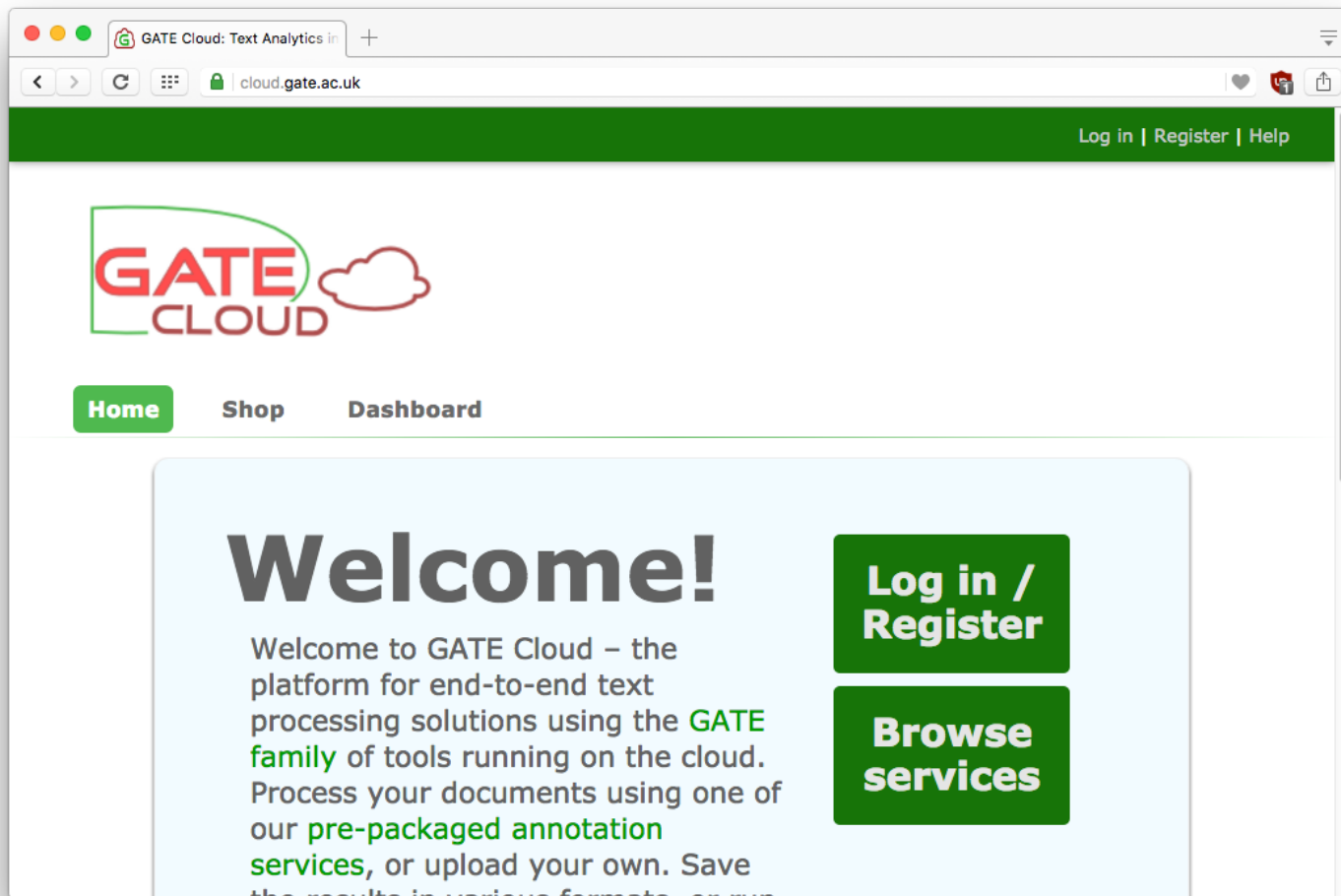
Pay hourly at a very reasonable rate (£0.05 an hour, or about £36 a month)

- First create an account for GATE Cloud
- Load some credit onto your account
- Order the service and wait for your reservation
- Start the machine and configure the collector!

It's recommended to save tweets to S3 or GATE Cloud, as they'll be deleted after a while if not downloaded.

GATE Cloud

<https://cloud.gate.ac.uk>



Dedicated servers

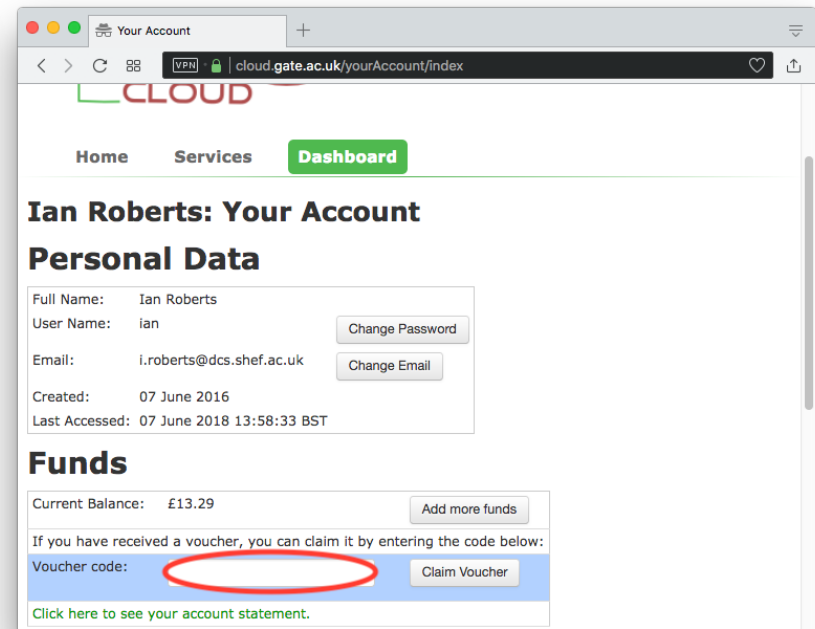
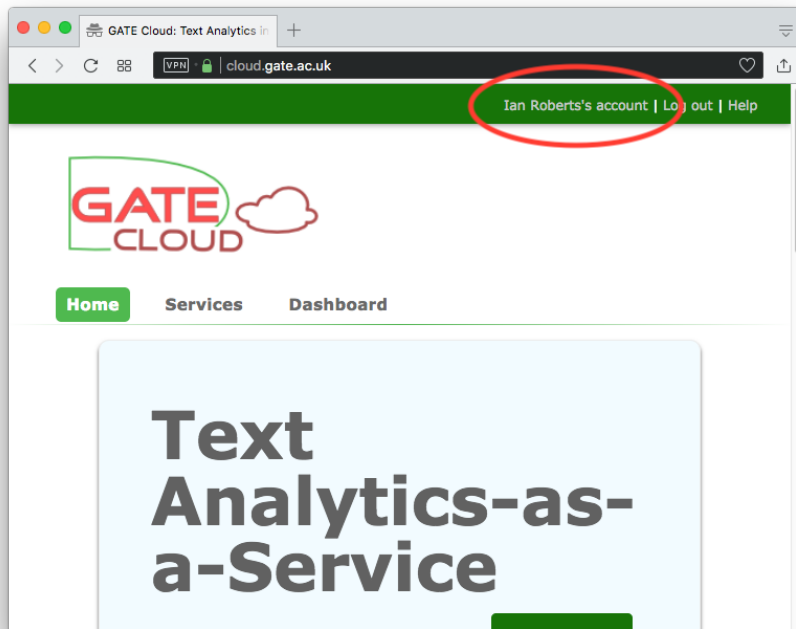
- Twitter collector is provided as a *dedicated server* – you rent a dedicated server for your private use
- Start and stop it as required
- Pay only for the hours it is running (though typically you would leave it running continuously)
- Backup and restore facility available

The screenshot shows the GATE Cloud website interface. At the top, there is a navigation bar with "Log in | Register | Help" links. Below the navigation bar is the GATE Cloud logo. The main content area features a "Services" tab and a list of items tagged with various languages and tasks. The services are displayed in three columns:

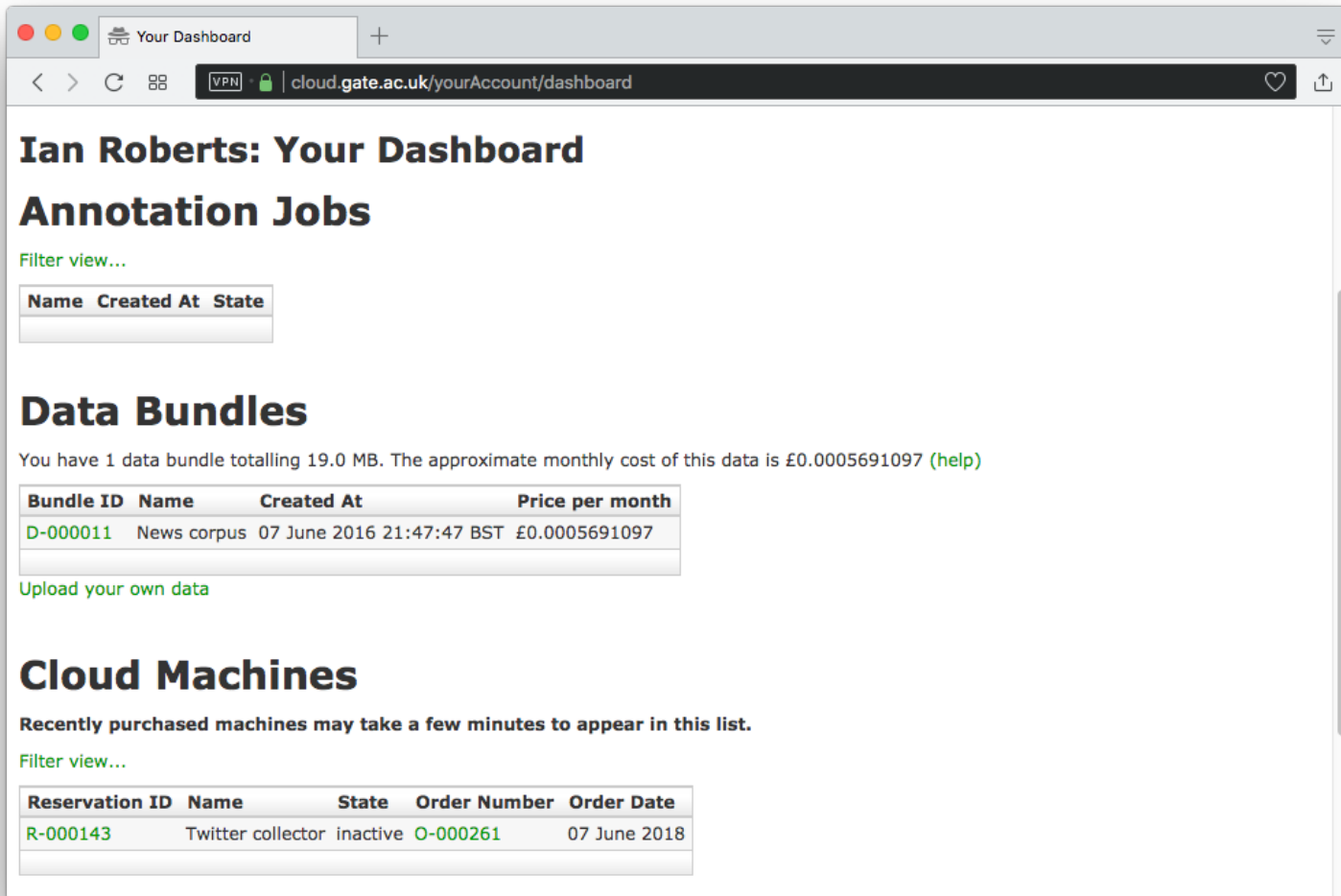
- English Named Entity Recognizer**: Identify names of *persons, locations, organizations*, as well as *money amounts, time and date expressions* in English texts automatically. **1,200 free requests / day**, Larger batches **£0.80 / CPU hour**.
- English Named Entity Recognizer for Tweets**: Analyse tweets for names of *persons, locations, organizations* and other entities. Also performs normalization of abbreviations and common shorthands ("brb", "gr8", "2day", etc.). **1,200 free requests / day**, Larger batches **£0.80 / CPU hour**.
- Twitter Collector**: Collect tweets, view tweet statistics, and store results in your dashboard for further analysis. **£0.05 / CPU hour**.

Reserving a server

- The usual e-commerce experience
 - Sign up for an account
 - Buy a top-up voucher (or use the free one we just gave you)
 - Find the server you want in the shop
 - Press “reserve this machine” and follow the instructions
- Server appears in your *dashboard*
- Behind the scenes, creates a persistent data *volume* for your data



Dashboard



Ian Roberts: Your Dashboard

Annotation Jobs

[Filter view...](#)

Name	Created At	State
------	------------	-------

Data Bundles

You have 1 data bundle totalling 19.0 MB. The approximate monthly cost of this data is £0.0005691097 ([help](#))

Bundle ID	Name	Created At	Price per month
D-000011	News corpus	07 June 2016 21:47:47 BST	£0.0005691097

[Upload your own data](#)

Cloud Machines

Recently purchased machines may take a few minutes to appear in this list.

[Filter view...](#)

Reservation ID	Name	State	Order Number	Order Date
R-000143	Twitter collector	inactive	O-000261	07 June 2018

Reservation control panel

The screenshot shows a web browser window with the URL `cloud.gate.ac.uk/yourAccount/machineReservationDetails/143`. The page header identifies the user as 'Ian Roberts's account' with links for 'Log out' and 'Help'. The GATE Cloud logo is prominently displayed, along with navigation tabs for 'Home', 'Services', and 'Dashboard'. The main content area is titled 'Machine Reservation R-000143' and contains a table of reservation details:

ID	R-000143	Destroy Reservation
Name	Twitter collector	Rename
Machine type	Twitter Collector	
Hourly price	£0.05	
State	inactive	Start Instance
Instance ready	no	
Backups		
Slot 1	<empty>	Create new backup

Below the table, the text 'Reservation Details:' is partially visible.

Controlling the server

- Start and stop instance
 - Startup/shutdown takes a few minutes – system will email you when server is ready
 - You pay the hourly price whenever the instance is running
- Backup and restore
 - Save the state of your data volume so you can roll back later
- Destroy reservation
 - If you no longer need the server, destroy it to discard the data volume and all backups
 - *This cannot be undone*

Other Social Media

- Twitter has historically allowed at least limited access and has recently announced even greater access for academics
- Most other sites actively prohibit access
 - Facebook make it almost impossible to access data (mostly as they've been involved in too many privacy issues in the past)
 - LinkedIn also don't allow access without written permission from each user, and you aren't allowed to store any data at all
- As such we can't recommend trying to access data from either of these



Linked



Storing social media data

What would help us do our science?

- NLP and network analysis tools often data-driven, preferring “as much data as possible”
- Not only do the messages change over time – meta-information also
- A minimum: something that helps others reproduce your work
- Abstract annotations over the raw data != the raw data

What native data can we safely store?

- Twitter: IDs and the freshest seen API call result

Ethical considerations

- We all have something to hide (e.g. from identity thieves)
- Important that personal data cannot proliferate once its owner removes / changes it
- How long to retain for? NSA's minimum 15-year seems excessive

Metadata just as powerful as text data
Text data weaker without metadata

Storing social media data

ELECTRONIC FRONTIER FOUNDATION

30C3 - 30 December 2013

Why Metadata Matters

- They know you rang a phone sex service at 2:24 am and spoke for 18 minutes. But they don't know what you talked about.
- They know you called the suicide prevention hotline from the Golden Gate Bridge. But the topic of the call remains a secret.
- They know you spoke with an HIV testing service, then your doctor, then your health insurance company in the same hour. But they don't know what was discussed.

(fr

Distribution concerns

- Social media corpora are difficult to distribute
- E.g. Twitter does not allow you to give other researchers / companies / anyone tweets you have collected and annotated in bulk
- Instead, distribute the tweet IDs and stand-off markup for the linguistic gold data
- The recipient re-collects all tweets himself, based on the IDs
- Necessary so user-deleted tweets are not propagated – privacy

Corpus completeness

- However, in some cases (e.g. misinformation, smear tweets) messages can be deleted
- Makes re-creating the corpus problematic
- Two classes of deletion:
 - Rapid deletions, usually within first few minutes (e.g. of spam, for editing the text)
 - Slower deletions (Petrovic et al. 2013)
- Our experience is that about 1 in 5 tweets are no longer available a year later.

Increased topic and entity drift: broader range of entities (Eisenstein 2013)

- Corpora age rapidly, and become less useful for some purposes (e.g. NEL)

Hands-on: Loading twitter data

- Open corpora/plain-tweets.json or your own corpus with a text viewer (such as notepad)
- Let's take a more useful view: find an online JSON viewer, and paste one line in. (e.g. "<http://jsonviewer.stack.hu>")
- Note the hierarchical structure of the data, and embedded user profile
- Now, let's load some data into GATE. First, load the Twitter plugin and the Format: JSON plugin
- Create a new GATE corpus called "Raw tweets" and save to DS
- Right-click on the corpus and choose "Populate from JSON"
- Select the JSON file used earlier, and make sure the mime type is set to "text/x-json-twitter"
- Examine the different annotations in the document