Twitter Challenges

November 2018

Jérôme Thièvre, Lead R&D engineer, jthievre@ina.fr
Legal context: 2011, Decree

Ina:

- Broadcasters’ online communication services
- Online communication services focusing on Radio and TV
- On demand audiovisual Media Services with broadcast content

Archive access restricted to Ina facilities
Twitter Collect at Ina

- tv/radio entities: channels, shows & media personalities
  - official publications and audience reactions
- events with high media echo
  - elections, sport, terrorist attacks, ...
- tweets embedded in archived web pages
- 700,000 tweets / day
Twitter Curation

- integrated with others social and video web platforms
  ⇒ link tv/radio entities to their web presences
    ○ website or part of a website
    ○ social account or keyword(hashtag) (twitter, facebook, instagram, ...)
    ○ video account/channel (youtube, dailymotion, vimeo, ...)

- curation has been refactored the last 2 years

- managed internally in collaboration with a tv/radio documentalist
API motivations & usages

- Canonical form
- Rich structured data
  - easy to index and search
  - opportunity to experiment data visualizations
- 13500 accounts
  - timeline + search (mentions)
- 750 hashtags/keywords
  - streaming + search (7 day window sample retrieval)
- 2500 avg tweets from webpages every day
  - get statuses / hydrate
4 years of collect

- nearly 1 billion tweets collected

- APIs are stable, only one important change

- and reliable, no major collect interruption to report
4 years of collect

- exhaustiveness can’t be obtained, but attention is required to limit data ‘holes’

- 1% streaming cap was rarely reached
  - during mondial events: paris attacks, olympics, world cup, ...

- requires 'smart' use of multiple accounts
Linked materials

- tweet can contain images, videos or links

- images and video entities are collected too
  - declared as media entities by twitter

- but not external links for now
  - we miss pages, and external videos and images
  - but plan to do it
  - already do this kind of collect for linked materials from a webpage
    (embedded video & tweets)
Trends API experimentation

- idea: provide a safety net
- extract France Trends most popular hashtags/keywords
- collect related tweets with streaming API
Trends API experimentation

- idea: provide a safety net
- extract France Trends most popular hashtags/keywords
- collect related tweets with streaming API

- collected tweets are mostly irrelevant to our collection
- and contain a lot of spam/junk/porn messages
- this collection is kept apart from our official collection
Storage

• how to store tweets in DAFF (Digital Archive File Format)
  ○ keep our separation between metadata and data
  ○ new type of metadata web data not identified by an url

• records :
  ○ metadata : urn, status, origin, date, content type, collect method, collection name, content (sha256 link to data record), ...
  ○ data : raw json tweet in standard data DAFF record

Archive storage and preservation is integrated in our current workflow
Access

Specific access application:
- fulltext search + specific fields
- timeline
- facets on relevant fields
- challenge to build
- good feedback from users
Access: social TV
Conclusion

- easy to collect from twitter API
- access complete & canonical data
- scalable & reliable
- storage & preservation integration
- require specific indexing and access
any Question ?
Thank you!

dlweb@ina.fr
Annex : Twitter APIs

- timeline : last 3200 tweets + 200 tweets / request + 1500 requests / 15 min
- streaming : realtime with 1% global cap + 400 filters max
- search : sample of tweets from last 7 days + 100 tweets / requests + 450 requests / min
- get statuses : 100 tweets / request + 900 requests / 15 min
Annex: DAFF

record ==
"R"
<1Byte type flag>
<1Byte compression flag>
<256bit SHA-256 signature>
<64bit content size>
<content>

type flag
0 : format description record
1 : purpose description

record
10 : metadata record in YAML
20 : metadata record in JSON

format
11 : data record
can be extended

compression flag
0 : raw
1 : deflate
2 : gzip
3 : bzip2
can be extended

Metadata record
Type flag: 10 or 20
YAML (10) or JSON (20) encoded.

Mandatory keys
- url, in a normalized form
- date, ISO 8601 string
- content: SHA-256 of the content in hexadecimal format
- status: 'ok' | 'redirection' | 'request_error' | 'server_error' | 'info' | 'unchanged' | 'ignored'
- location: url to redirect to, in case of protocol redirection (status = 'redirection')

Optional keys
- type (MIME Type)
- page, 0/1, reflects if the url was a page
- last_modified, reflecting the Last-Modified HTTP header
- ip, the ip of the peer server (if not a proxy)
- comment, plain message
- ...
Annex: DAFF for http(s) data

**DAFF Data Records**

- **sha_256:** e4ba78b2c0034f
  - **url:** http://prog-tv.fr/images/goldorak.jpg
  - **date:** 2018-01-02 08:51:00Z
  - **sha_256:** e4ba78b2c0034f

- **sha_256:** e4ba78b2c0034f
  - **url:** http://prog-tv.fr/images/goldorak.jpg
  - **date:** 2018-01-13 10:02:00Z
  - **sha_256:** e4ba78b2c0034f

- **url:** http://anime-fan.fr/img/grendizer.jpg
  - **date:** 2018-02-02 18:33:00Z
  - **sha_256:** e4ba78b2c0034f

**DAFF Metadata Records**

- **url:** http://prog-tv.fr/style/main.css
  - **date:** 2018-01-02 08:50:00Z
  - **sha_256:** babc4e3130004def6

- **url:** http://prog-tv.fr/style/main.css
  - **date:** 2018-01-13 10:01:00Z
  - **sha_256:** babc4e3130004def6