

TwitterData

Big Data

Données et traitements

TwitterData

- * Suite logique du travail sur mapReduce
- * twitterData : projet d'analyse des données de Twitter
- * Difficulté : traiter la totalité des données
- * Twitter : mouvance Big Data
- * Twitter génère un flux massif de données : « stream twitter »
- * Un tweet : possède une structure définie.
- * Contenu analysable selon les rubriques

Twitter

- * Société créée le 21 mars 2006, entrée en bourse fin 2013 (novembre 2013)
- * Siège social : San Fransisco (SF Bay)
- * Site de microblogging.

NYSE: TWTR - 23 oct. 10:52 UTC-4

NYSE: TWTR - 23 oct. 10:52 UTC-4

29,95 ↑0,79 (2,71 %)

1 jour

5 jours

1 mois

3 mois

1 an

5 ans

max



Ouverture	29,97
+Haut	30,30
+Bas	29,61

+Haut	30,30
-------	-------

+Bas	29,61
------	-------

Capitalis. 20,26 Md

Cours/bén. -

Rend. div. -

Twitter

- * Particularité :
 - * Mise à disposition programme développeur
 - * Mise à disposition d'API à travers une URL.
 - * Mise à disposition d'un stream publique (limites)
 - * Intérêt : Les tweets (massifs) peuvent donner des tendances
 - * Intérêt : L'analyse peut reposer sur le travail précédent

Analyse de Twitter

- * Deux façon d'analyser le flux :
 - * Outils console : curl (test de connexion), filtres grep, awk, sed (Unix, Linux)
 - * Les outils console ne permettent pas la souplesse d'un développement
- * Programmation suivant le choix d'un langage (extraction du flux)
- * Adaptable sur différentes plateformes
- * Utilisation de librairies (Framework ?)

Analyse du flux

- * Formats de transmission :
 - * XML : balise type HTML, structure rigide, balise de fermeture nécessaire (problème)
 - * JSON : plus souple, moins lourd qu'XML. Produit des fichiers plus légers
 - * Les deux formats autorisent la « sérialization »
 - * S'intègrent bien en POO

Langages

- * Le choix des langages est assez libre :
 - * C/C++ permet de rester sur les concepts du mapReduce
 - * PHP, C#, ObjC, erlang, Python, VB, JAVA, ... aussi !
 - * Choix du C/C++ : application standalone, qui peut fonctionner seule (aucun navigateur, plug-in, extensions ...)
 - * Application qui peut servir des données pour d'autres (mapReduce, tableur, DataViz solution)

Outils

- * Bibliothèques disponibles (open-source) :
 - * Twitcurl
 - * Curl
 - * Oauth
 - * Ofx-twitter (réutilise ofx-dev-master) : intéressant aussi pour d'autres choses (Arduino, kinect, ...)
- * Outils (open-source) :
 - * Elastic Search avec le plug'in twitter peut utiliser le stream public.
 - * Kibana peut servir à faire de la dataViz

Projet

- * Dans le cadre d'un projet PPE
- * Mettre en place une application :
 - * C/C++
 - * Réaliser l'équivalent du projet Windows TwitterClient (C/C++)
 - * Gestion du flux continu !
 - * Collecte : stockage conséquent dans le temps
 - * Analyse : mapReduce avec du volume
 - * Visualisation : présentation des résultats à travers une interface

Project Settings

- * Ouverture d'un compte « twitter » développeur :
 - * <https://dev.twitter.com>
 - * Création d'une application
 - * Remplissage du formulaire de l'application

My applications

Create a new application



getDAtaWithCurl
consuming twitter stream

Project Settings

* Il faut remplir les différents formulaires :

getDataWithCurl

Details

Settings

OAuth tool

@Anywhere domains

Reset keys

Delete



consuming twitter stream

<http://techspeech.fr> 

Organization

Information about the organization or company associated with your application. This information is optional.

Organization	None
--------------	------

Organization website	None
----------------------	------

Project Settings

- * Clés à utiliser dans l'application

OAuth Settings

Consumer key: *

5e01a12h22xgny10fKvw

Consumer secret: *

015_1473_0_01_10N0P158X0860010P0_100_100

Remember this should not be shared.

Access token: *

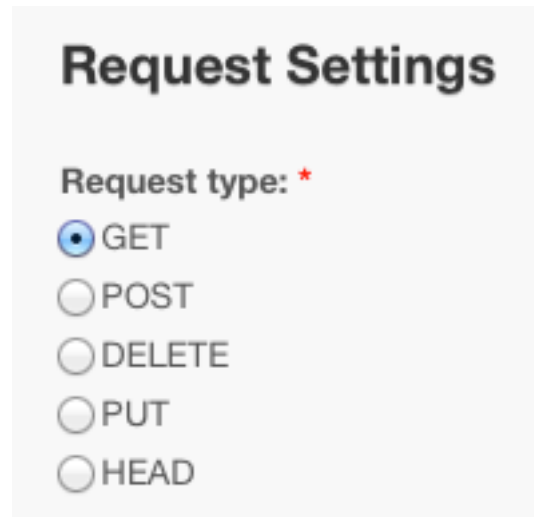
1000270570_0X5_100_T01X_5D1Z57D0_050V14MM110011

Access token secret: *

0X0_1_010_57011_110251PNNV0LTT0W_100_N1P0

Project Settings

- * Il faut mettre en cohérence le code (création de l'URL) et la méthode déclarée :



Request Settings

Request type: *

☒ GET

☐ POST

☐ DELETE

☐ PUT

☐ HEAD

Project Settings

* Une URL permet de faire des tests :

Request URI: *

The full URI, without parameters. For example: *https://api.twitter.com/1/statuses/home_timeline.json*

Request query:

The parameters for your request. For example: *include_entities=true&page=2*. Note these parameters will be sent on the querystring for GET requests, and in the request body for POST requests.

Request URI: *

The full URI, without parameters. For example: *https://api.twitter.com/1/statuses/home_timeline.json*

Request query:

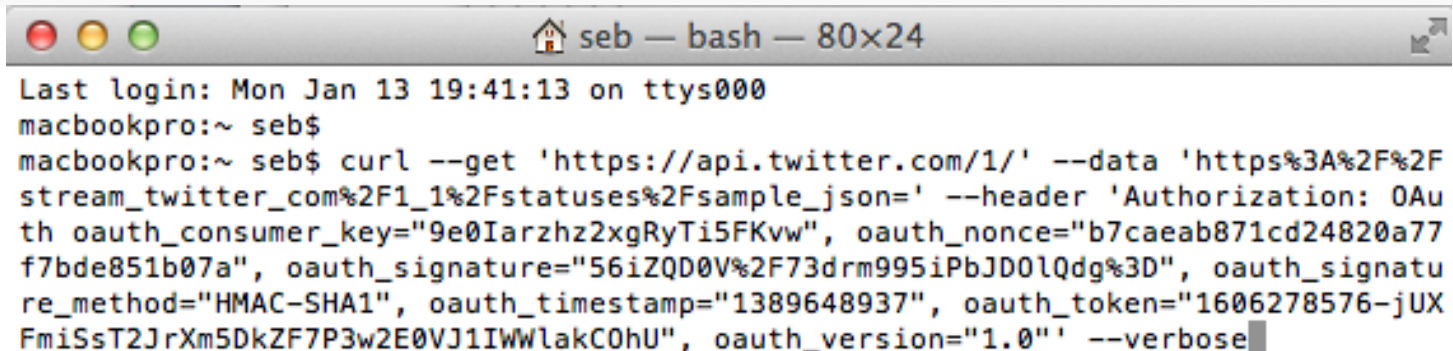
The parameters for your request. For example: *include_entities=true&page=2*. Note these parameters will be sent on the querystring for GET requests, and in the request body for POST requests.

Project Settings

- * L'URL permet de générer une commande CURL en version : 1
- * version actuelle : 1.1

cURL command

```
curl --get 'https://api.twitter.com/1/' --data 'https%3A%2F%2Fstream_twitter_com%2F1_1%2Fstatuses%2Fsample_json=' --header 'Authorization: OAuth oauth_consumer_key="9e0Iarzhz2xgRyTi5FKvw", oauth_nonce="b7caeab871cd24820a77f7bde851b07a", oauth_signature="56iZQD0V%2F73drm995iPbJD0lQdg%3D", oauth_signature_method="HMAC-SHA1", oauth_timestamp="1389648937", oauth_token="1606278576-jUXFmiSsT2JrXm5DkZF7P3w2E0VJ1IWWlakCOhU", oauth_version="1.0"' --verbose
```

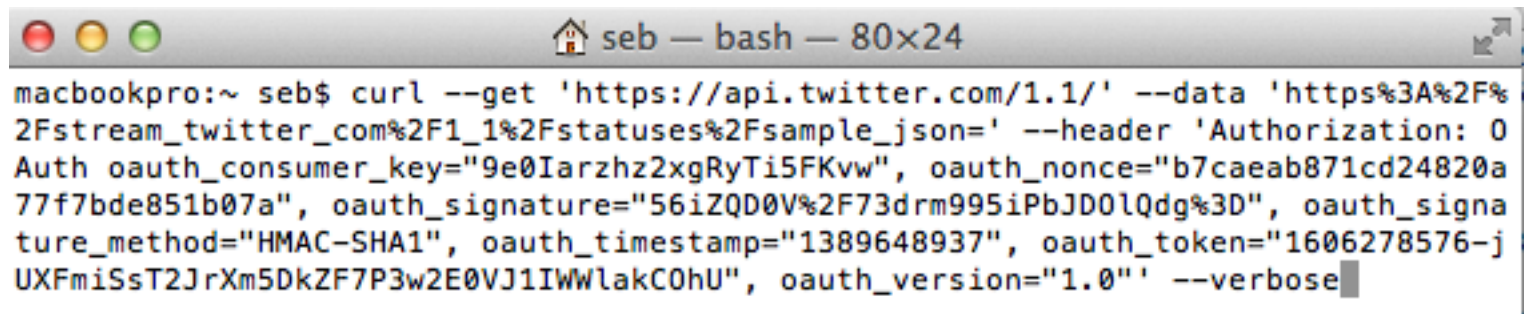


A screenshot of a macOS terminal window. The title bar shows the window name 'seb — bash — 80x24'. The terminal content shows the user 'seb' logging in on 'Mon Jan 13 19:41:13 on ttys000'. The user enters the curl command from the previous block, and the prompt returns to 'macbookpro:~ seb\$'.

```
macbookpro:~ seb$ curl --get 'https://api.twitter.com/1/' --data 'https%3A%2F%2Fstream_twitter_com%2F1_1%2Fstatuses%2Fsample_json=' --header 'Authorization: OAuth oauth_consumer_key="9e0Iarzhz2xgRyTi5FKvw", oauth_nonce="b7caeab871cd24820a77f7bde851b07a", oauth_signature="56iZQD0V%2F73drm995iPbJD0lQdg%3D", oauth_signature_method="HMAC-SHA1", oauth_timestamp="1389648937", oauth_token="1606278576-jUXFmiSsT2JrXm5DkZF7P3w2E0VJ1IWWlakCOhU", oauth_version="1.0"' --verbose
```


Project Settings

- * La requête en 1.1 :



```
macbookpro:~ seb$ curl --get 'https://api.twitter.com/1.1/' --data 'https%3A%2F%2Fstream_twitter_com%2F1_1%2Fstatuses%2Fsample_json=' --header 'Authorization: 0 Auth oauth_consumer_key="9e0Iarzhz2xgRyTi5FKvw", oauth_nonce="b7caeab871cd24820a77f7bde851b07a", oauth_signature="56iZQD0V%2F73drm995iPbJD0lQdg%3D", oauth_signature_method="HMAC-SHA1", oauth_timestamp="1389648937", oauth_token="1606278576-jUXFmiSsT2JrXm5DkZF7P3w2E0VJ1IWWlakC0hU", oauth_version="1.0"' --verbose
```

- * Le résultat de la commande est juste une suite de tests de connexions

Project settings

```
* Adding handle: conn: 0x7fca4a003a00
* Adding handle: send: 0
* Adding handle: recv: 0
* Curl_addHandleToPipeline: length: 1
* - Conn 0 (0x7fca4a003a00) send_pipe: 1, recv_pipe: 0
* About to connect() to api.twitter.com port 443 (#0)
*   Trying 199.16.156.104...
* Connected to api.twitter.com (199.16.156.104) port 443 (#0)
* TLS 1.2 connection using TLS_ECDHE_RSA_WITH_RC4_128_SHA
* Server certificate: api.twitter.com
* Server certificate: VeriSign Class 3 Secure Server CA - G3
* Server certificate: VeriSign Class 3 Public Primary Certification Authority - G5
> GET /1.1/?https%3A%2F%2Fstream_twitter_com%2F1_1%2Fstatuses%2Fsample_json= HTTP/1.1
> User-Agent: curl/7.30.0
> Host: api.twitter.com
> Accept: */*
> Authorization: OAuth oauth_consumer_key="9e0Iarzhz2xgRyTi5FKvw", oauth_nonce="b7caeab871cd24820a77f7bde851b07a", oauth_signature="56iZQD0V%2F73drm995iPbJD0lQdg%3D", oauth_signature_method="HMAC-SHA1", oauth_timestamp="1389648937", oauth_token="1606278576-jUXFmiSsT2JrXm5DkZF7P3w2E0VJ1IWlakoC0hU", oauth_version="1.0"
>
< HTTP/1.1 401 Unauthorized
< date: Mon, 13 Jan 2014 21:40:23 UTC
* Server tfe is not blacklisted
< server: tfe
< set-cookie: guest_id=v1%3A138964922365700904; Domain=.twitter.com; Path=/; Expires=Wed, 13-Jan-2016 21:40:23 UTC
< strict-transport-security: max-age=631138519
< Connection: close
<
* Closing connection 0
```

Résultat
de la
Commande
curl :

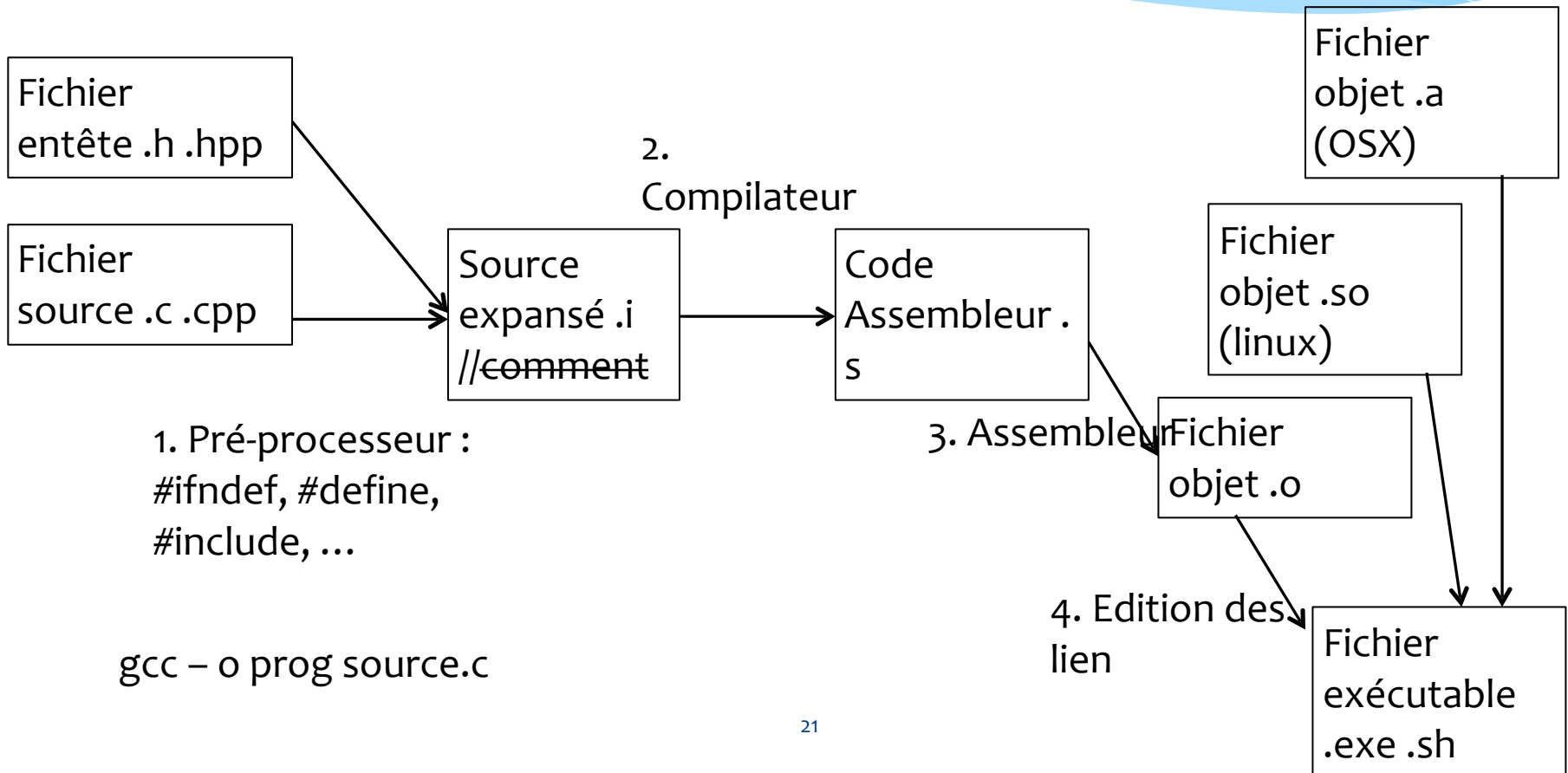
Programmation

- * Utilisation d'un IDE au choix :
 - * Code::Blocks
 - * Xcode
 - * Visual Studio
- * Choix d'un framework : lecture et recherche
- * Recherche et installation du framework :
 - * Tests d'intégration des libraires
 - * Tests de codes

Programmation

- * Configuration du projet :
 - * #include des fichiers nécessaires : .h, .cpp
 - * Disponibilité des librairies :
 - * .dll, .lib : windows
 - * .so : linux
 - * .a, .dy, .dylib : OSX
 - * Édition des liens paramétrés : paramètres de la chaîne de compilation (linkage)

Phase de compilation

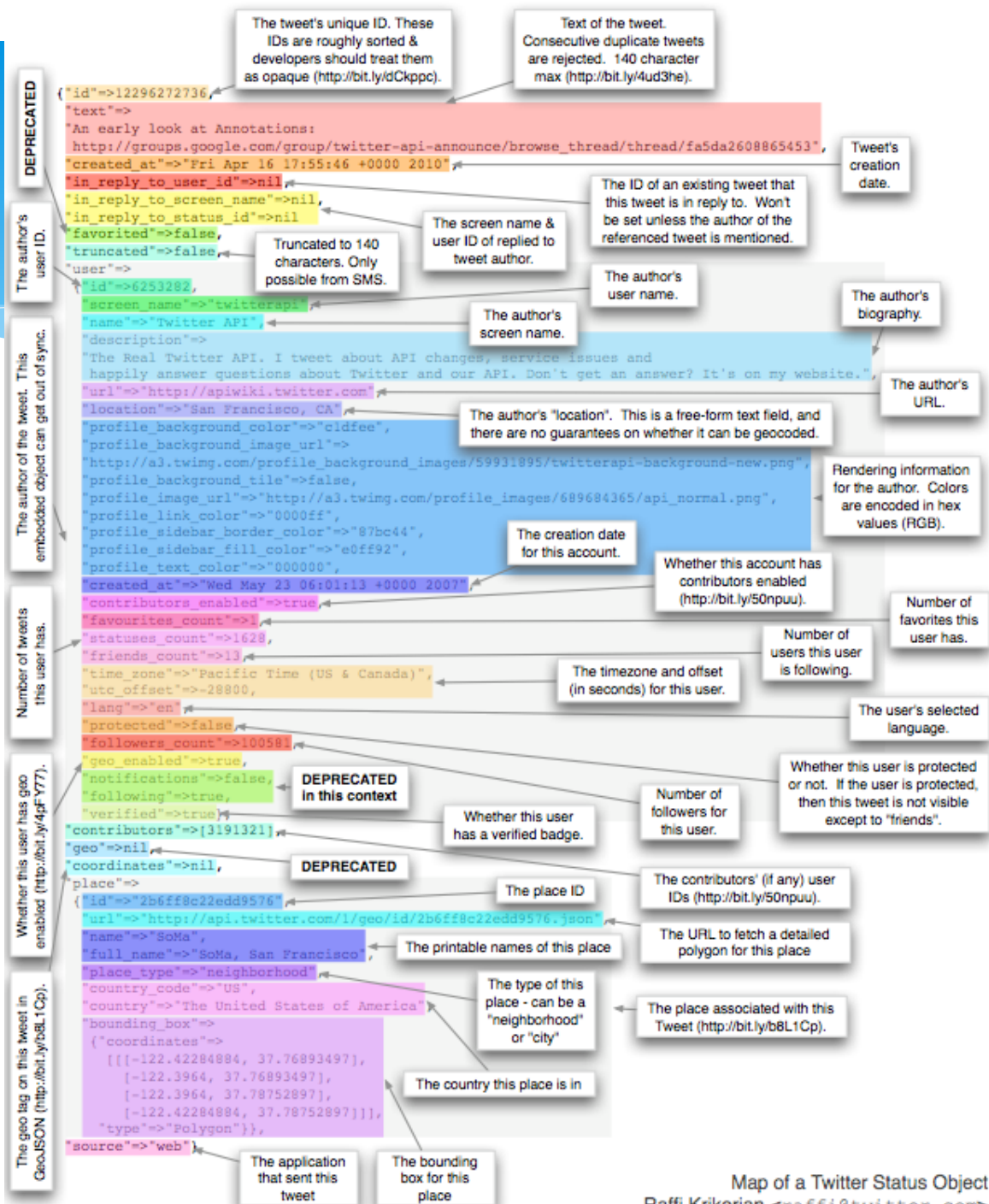


Programmation

- * Objectifs

- * Stocker les tweets dans le temps
- * Récupérer le champ « text » du tweet = structure
- * Analyser le contenu
- * Imaginer d'autres fonctionnalités ...

- * Structure d'un tweet
- * Format JSON



Programmation

- * Structure d'un tweet
- * Format JSON

```
"contributors": null,  
"retweeted": false,  
"in_reply_to_user_id_str": null,  
"place": null,  
"retweet_count": 4,  
"created_at": "Sun Apr 03 20:24:49 +0000 201",  
"user": {  
  "notifications": null,  
  "profile_use_background_image": true,  
  "statuses_count": 31,  
  "profile_background_color": "C0DEED",  
  "followers_count": 3066,  
  "profile_image_url":  
twimg.com/profile_images/1285770264/PGP_normal  
  "listed_count": 6,  
  "profile_background_image_url":  
twimg.com/a/1301071706/images/themes/theme1/b  
  "description": "",  
  "screen_name": "PostGradProblem",
```


Programmation

* Format de sortie « raw » de tweets pris aléatoirement :

```
test.txt
{"delete":{"status":{"id":"290197547815927808","user_id":"439490568","id_str":"290197547815927808","user_id_str":"439490568"}}}
{"delete":{"status":{"id":"290197510029459456","user_id":"439490568","id_str":"290197510029459456","user_id_str":"439490568"}}}
{"delete":{"status":{"id":"139463627919982592","user_id":"20159990","id_str":"139463627919982592","user_id_str":"20159990"}}}
{"delete":{"status":{"id":"363513935590195200","user_id":"308968069","id_str":"363513935590195200","user_id_str":"308968069"}}}
{"delete":{"status":{"id":"315646063690280961","user_id":"288340320","id_str":"315646063690280961","user_id_str":"288340320"}}}
{"created_at":"Mon Jan 13 18:28:48 +0000 2014","id":"422797394347294720","id_str":"422797394347294720","text":"Me pone de los nervios que ni ve\u00e1s xd.","source":"\u003ca href=
'http://twitter.com/download/android' rel='nofollow'\u003eTwitter for Android\u003c/a
\u003e","truncated":false,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"us
er":{"id":"171507061","id_str":"171507061","name":"\u003c1\u003b9g\u003b9o\u003b7","screen_name":"DemiMBarrios","location":"BCN","url":null,"description":"Constructor de cohetes,
experto en pizza, amante de los Gigantes, padre.","protected":false,"followers_count":205,"friends_count":55,"listed_count":0,"created_at":"Tue Jul 27 14:18:25 +0000
2010","favourites_count":4456,"utc_offset":3600,"time_zone":"Amsterdam","geo_enabled":true,"verified":false,"statuses_count":
8603,"lang":"es","contributors_enabled":false,"is_translator":false,"profile_background_color":"131516","profile_background_image_url":"http://a0.twimg.com/
profile_background_images/378800000120114183/f196f5dab5ebddc6781162934cf03a78.jpeg","profile_background_image_url_https":"https://s10.twimg.com/profile_background_images/
378800000120114183/f196f5dab5ebddc6781162934cf03a78.jpeg","profile_image_url":"http://pbs.twimg.com/profile_images/413047444054630400/hUvJyI9_normal.png","profile_image_url_https":"https://pbs.twimg.com/profile_images/413047444054630400/hUvJyI9_normal.png","profile_banner_url":"https://pbs.twimg.com/
profile_banners/171507061/
1386443199","profile_link_color":"302D2E","profile_sidebar_border_color":"FFFFFF","profile_sidebar_fill_color":"E6E6E6","profile_text_color":"333333","profile_use_background_image":
true,"default_profile":false,"default_profile_image":false,"following":null,"follow_request_sent":null,"notifications":null,"geo":null,"coordinates":null,"place":null,"contributo
rs":null,"retweet_count":0,"favorite_count":0,"entities":{"hashtags":[],"symbols":[],"urls":[],"user_mentions":
[]},"favorited":false,"retweeted":false,"filter_level":"medium","lang":"es"}
{"created_at":"Mon Jan 13 18:28:48 +0000 2014","id":"422797394347311104","id_str":"422797394347311104","text":"Que grande Cafu. Dos copas del mundo no es f
\u00e1cil","source":"\u003ca href='http://twitter.com/carbonandroid' rel='nofollow'\u003eCarbon for Android\u003c/a
\u003e","truncated":false,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"us
er":{"id":"89851450","id_str":"89851450","name":"Jos\u00e9 Nobrega","screen_name":"nobregad","location":"Granate rules","url":"http://l
lamemoslepatria.blogspot.com","description":"Medicina UC. Poes\u00eda en: Llamemosle Patria","protected":false,"followers_count":475,"listed_count":
3,"created_at":"Sat Nov 14 02:34:11 +0000 2009","favourites_count":10,"utc_offset":-16200,"time_zone":"Caracas","geo_enabled":false,"verified":false,"statuses_count":
16396,"lang":"es","contributors_enabled":false,"is_translator":false,"profile_background_color":"FFF04D","profile_background_image_url":"http://a0.twimg.com/
profile_background_images/394146202/images.jpg","profile_background_image_url_https":"https://s10.twimg.com/profile_background_images/394146202/
images.jpg","profile_image_url":"http://pbs.twimg.com/profile_images/3656262691/
1a26fa28aff68a0fbbdaa9b2ce142da0_normal.jpeg","profile_image_url_https":"https://pbs.twimg.com/profile_images/3656262691/
```

* Avec application C/C++

Programmation

- * L'exemple précédent est basé sur :
- * <https://stream.twitter.com/1/statuses/sample.json>
- * Delivers a random sampling of all Tweets at a volume equal to the public streaming cap.

Programmation

- * Isoler le contenu des champs « text » :

```
"text": "#ActiononSugar
{"text": "zuccheri"
"text": "5 dicas para ser mais positivo na sua empresa\nhttp://t.co/dfGbuqltvm"
"text": "5. @MaffewRagazino \u2013 Brownsville\u2019s Jesus | http://t.co/gZX2tQE8xl #FakeDog
"text": "@Mai_Tetsune_bot \u3010\u65b0\u7cfb\u3011\u79c1\u306e\u3053\u3068\u3067\u3059\u306d\u306e\u5728\u306f\u591a\u6469\u6e56\u7dda\u3011\u591a\u6469\u5ddd\u7dda\u3067\u8d70\u3063\u3066\u306e\u3067
"text": "Send emails urging @BizBash @CyberCoders @KaptureVision & Majestic Hills not to rec
"text": "Ambeh naon atuh mun di foto mata teh sok di bolotot2
"text": "And pastors needing sermon illustrations say
"text": "\"Temos que parar de acreditar no mito da igualdade de g\u00eanero\"
"text": "\u00a1Top 10 de #bandas de #metal 2013! Cu\u00e9ntanos \u00bfCu\u00e9l es tu favorita?\"
{"text": "metal"
"text": "Gamer's Scoop - The future is upon us
"text": "It was a mild sunny morning in the first week of June. I sat up on the coping of the br
"text": "Happy Founders' Day to the ladies of Delta Sigma Theta Sorority
"text": "@TheseAreThings on the blog today! Check them out
"text": "Mesabi Daily News: New job creation fund to aid hirings
"text": "Learn to play the Ukulele with @mytimesplus - 'no previous musical knowledge required;
"text": "New and hands-on Advanced Television and Film - Script to Screen program @StoryArtsCent
"text": "RT @TesDroleToi: @ShaaIdjazair @RomaneOvide MDRDR J'ESP\u00c8RE POUR TOI MESKINA"
"text": "@ShaaIdjazair @RomaneOvide MDRDR J'ESP\u00c8RE POUR TOI MESKINA"
"text": "\u00bfSe acuerdan de la S10? Capturaron una en Maracay http://t.co/gM0tXVVbr5"
```

Programmation

- * Pour isoler les champs « text » :
- * Console : pour évaluer rapidement les choses

```
cat test.txt | awk -F=":" -v RS="," '$1~/ "text"/ {print}' | grep -v -e entities|
```

- * Code :
 - * algorithme de lecture du fichier
 - * Personnalisation du travail

Programmation

- * Problème à gérer :
 - * Stockage continu : lancement de la tâche programmée
 - * Volume de stockage : disques, machines ?
 - * Processus d'exploitation :
 - * temps réel
 - * différé
 - * problème d'accès mutuels (conflits).

Programmation

* Twitter impose des limites :

Rate limit window duration is currently **15 minutes** long. Curious how rate limiting works in API v1.1? Read [REST API Rate Limiting in v1.1](#).

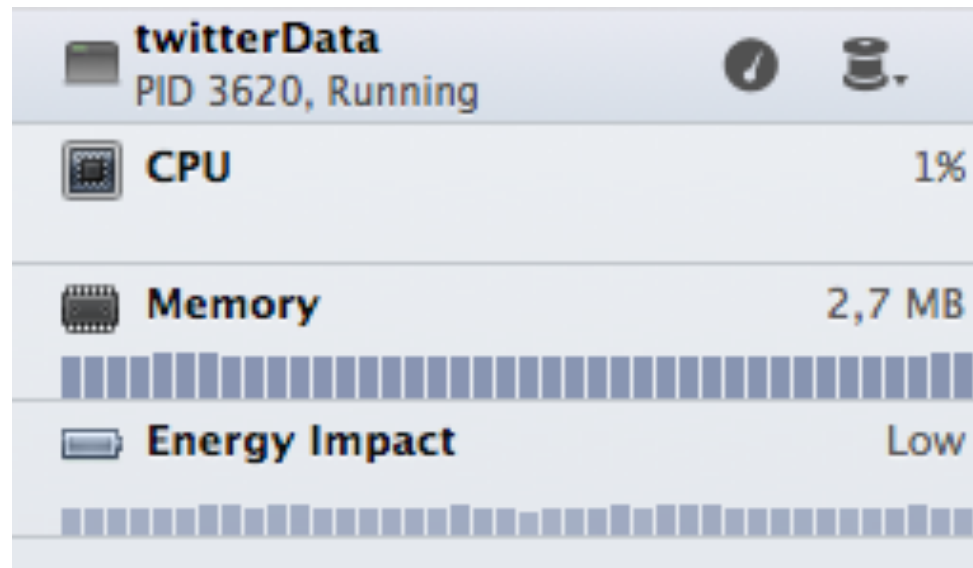
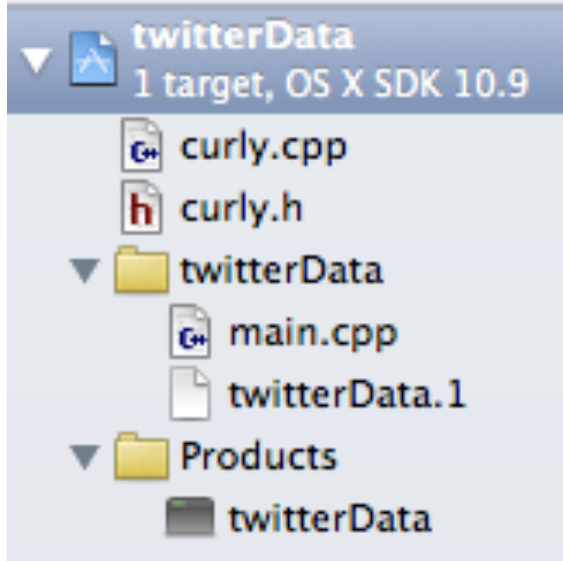
Resource family	Resource	Requests allotted per user	Requests allotted via application-only auth
account	GET account/settings	15	
account	GET account/verify_credentials	15	
application	GET application/rate_limit_status	180	180
blocks	GET blocks/ids	15	
blocks	GET blocks/list	15	
direct_messages	GET direct_messages	15	
direct_messages	GET direct_messages/sent	15	
direct_messages	GET direct_messages/show	15	
favorites	GET favorites/list	15	15
followers	GET followers/ids	15	15

Programmation

- * Limites :
 - * API version 1.1
 - * **15 Minute Windows**
 - * Rate limits in version 1.1 of the API are divided into 15 minute intervals, which is a change from the 60 minute blocks in version 1.0. Additionally, all 1.1 endpoints require authentication, so no longer will there be a concept of unauthenticated calls and rate limits
- * <https://dev.twitter.com/docs/rate-limiting/1.1>

Programmation

- * Structure du projet : (version de base)



Gestion du fichier

- * Test sur 16 minutes :
 - * Taille du fichier : 162 Mo
 - * Réduction le champ « text » : 8Mo
- * Actions à réaliser :
 - * Faire du nettoyage pour enlever les parasites
 - * Sélectionner une langue (au démarrage)
 - * Cibler des comptes (leaders)
 - * Prendre en compte les Followers ...

Schéma/infra

