



**THE  
POWER  
TO KNOW<sup>®</sup>**

## Surfing the Social Revolution with SAS

---

The Twit's guide to Tweets and Twitter



**THE  
POWER  
TO KNOW®**

## The Business Context

---

# Social Media: The Rise of the Social Network

- The Internet has changed how we live and interact
  - Social networks are costless to form and disband
  - Communities are dynamic and overlapping
  - Information travels the globe in hours (Blogs lag traditional media by 2.5 hours)
  - We interact locally and globally without thinking twice
  - Our devices are increasing becoming our publishing agents
  - Networks are becoming increasingly interconnected
- Rich media is becoming the norm
  - Geo-tagged information
  - Pictures, art, and photos
  - Lifestyle patterns (e.g. exercise, music, travel, technology)



# Social Media: Current Popular Networks

Facebook (250m)

MySpace (263m)

Orkut (67m)

Last.fm (30m)

Twitter (25m)

Plaxo (15m)

LinkedIn (43m)

Bebo (40m)

Classmates.com (50m)

Badoo (37m)

Flixster (63m)

Habbo (117m)

hi5 (80m)

Mixi (21m)

MyLife.com (51m)

MyHeritage (30m)

Netlog (42m)

Odnoklassniki (37m)

Skyrock (22m)

Tagged.com (70m)

Windows Live Spaces (100m)



**THE  
POWER  
TO KNOW®**

## The Technical Overview

---

# Application of multiple SAS capabilities

- SAS 9.2 Web Services
- Plain-text document parsing (RegEx)
- XML import / parsing
- Formatting and presenting the results
- Stored Processes and process distribution
- Text Mining

# Part 1: Web Services

- Use 9.2 Web Services to receive an XML stream
- Twitter returns search results by page, requiring multiple pages for larger results
- Search string is written to a file where PROC HTTP can then pick it up
- Proxy details are required to navigate SAS's proxies
- Languages and geocodes are extensible by editing the tables

```
DATA WebReq;
    req = "geocode=&location&q=&topic&page=&i&rpp=100";
run;
```

```
DATA _NULL_;
    SET WebReq;
    FILE Request;
    PUT req;
run;
```

```
proc http
    in=Request
    out=TwOut
    headerout=HdrOut
    method="get "
    proxyhost="&prxHost "
    proxyusername="&prxUser "
    proxypassword="&prxPass_PASSWORD"
    proxyport=80
    url="http://search.twitter.com/search.atom? ";
run;
```

# Part 2:

## Plain-text document parsing (RegEx)

- Four potential cases:
  - One page returned: Do nothing
  - First page of multiple pages: Strip the trailing tags
  - Intermediary page of multiple pages: Strip the leading and trailing tags
  - Final page of multiple pages: Strip the leading tags
  
- Use Regular Expressions (RegEx, otherwise known as line noise) to identify the lines to be stripped

```

%IF &pages = 1 %THEN %DO;
  DATA _NULL_;
    INFILE TwtOut recfm=v lrecl=2048 TRUNCOVER;
    FILE MinOut lrecl=2048;
    INPUT line $2048.;
    PUT _INFILE_;
%END;
%ELSE %IF &i = 1 %THEN %DO;
  DATA _NULL_;
    INFILE TwtOut recfm=v lrecl=2048 TRUNCOVER;
    FILE MinOut lrecl=2048;
    INPUT line $2048.;
    _INFILE_ = prxchange('s/<\feed>/', 1, _INFILE_);
    PUT _INFILE_;
%END;
%ELSE %IF &i < &pages %THEN %DO;
  DATA _NULL_;
    INFILE TwtOut recfm=v lrecl=2048 TRUNCOVER;
    FILE MinOut MOD lrecl=2048;
    INPUT line $2048.;
    _INFILE_ = prxchange('s/<feed.+>/', 1, _INFILE_);
    _INFILE_ = prxchange('s/<?xml.+>/', 1, _INFILE_);
    _INFILE_ = prxchange('s/<\feed>/', 1, _INFILE_);
    PUT _INFILE_;
%END;

```

# Part 3:

## XML Import / Parsing

- Create a library that directly references the XML file
- Use a pre-generated XML map (using SAS XML Mapper) to map the tags to fields
- Read in the XML file, dropping all languages other than the desired language

```
LIBNAME XMLTwit XML 'c:\twitter\textMining.xml'  
      XMLMAP='c:\twitter\Twitter.map' ACCESS=READONLY;
```

```
Data Twitter.Tweets;  
      SET XMLTwit.XMLTweet;  
      If lang = "&searchLang";  
run;
```

**SAS XML Mapper**

File Tools Help

Condensed Full Schema

Attributes

- id=tag:search.twitter.com,2005:search/iphone
- link
- link
- title=iphone - Twitter Search
- link
- link
- twitter:warning=adjusted since\_id to 2625758792 (2009-07-14 03:00:00 UTC) for refresh querysince\_id removed for pagination.
- updated=2009-07-21T03:19:39Z
- openSearch.itemsPerPage=100
- link
- entry
  - id=tag:search.twitter.com,2005:2751273488
  - published=2009-07-21T03:19:39Z
  - link
  - title=Suspicious confirmed. #iPhone showing 100% battery power after 24 hours of on time from last recharge. A reboot now shows 20%. Bug must
  - content=Suspicious confirmed. <a href="http://search.twitter.com/search?q=%23iPhone">#<b>iPhone</b></a> showing 100% battery power after 24
  - updated=2009-07-21T03:19:39Z
  - link
  - google.location=Sydney, Australia
  - twitter:source=<a href="http://www.tweetdeck.com">TweetDeck</a>
  - twitter:lang=en
  - author

Properties

Name: XMLTweet

Description:

Path:


End Path: Begin/End

Retain  Replace

XMLTweet

- name
- uri
- published
- updated
- title
- content
- lang
- source

XML source XML Schema source XMLMap SAS Code Example Table view Contents Validate Log

Table: <no data>  SAS formats and informats are not applied to this view.

XML file loaded: xmlTweets.xml

0 0 0 1 1 8

# Part 4: Formatting and Presenting the Results

- Three things left to do:
  - Tell the user if the search didn't return anything
  - Update metadata
  - Present the results

```

%macro NumObs;
data _null_;
    call symputx ('NumObs',put(numobs,14.));
    set Twitter.Tweets noobs=numobs;
    stop;
run;

%if &NumObs = 0 %then %do;
    data NoData;
        length Line $255.;
        label Line='00'x;
        Line = "Sorry, nothing was returned using that search and
        location!";

    run;

    proc print data=NoData noobs label;
    run;
%end;
%mend NumObs;

Proc Metalib;
    omr (library="Twitter" metarepository="Foundation");
    update_rule=(delete);
run;

proc print data=Twitter.Tweets;
    var name published title;

```

# Part 5: Stored Process and Process Distribution

- Create process in Enterprise Guide
- Register process
- Export process

### Properties for Get Tweets

General  
Results  
Prompts  
Summary

#### Prompts

Project prompts used:

| SAS Name        | Display Name         | Data Type |
|-----------------|----------------------|-----------|
| searchTopic     | Search Topic         | Text      |
| searchLocation  | Search Location      | Text      |
| searchLang      | Search Language      | Text      |
| pageCount       | Pages to be Returned | Numeric   |
| prxUser         | Proxy Username       | Text      |
| prxPass_PASS... | Proxy Password       | Text      |
| prxHost         | Proxy Host           | Text      |

### Prompt Manager

| Name        | Displayed text      | Prompt type | Default value          | Used By                   |
|-------------|---------------------|-------------|------------------------|---------------------------|
| pageCount   | Pages to be Retu... | Numeric     | 1                      | Get Tweets (Process Flow) |
| prxHost     | Proxy Host          | Text        | apacproxy01.oz.sas.com | Get Tweets (Process Flow) |
| prxPass_... | Proxy Password      | Text        |                        | Get Tweets (Process Flow) |
| prxUser     | Proxy Username      | Text        |                        | Get Tweets (Process Flow) |
| searchLang  | Search Language     | Text        | en                     | Get Tweets (Process Flow) |
| searchLo... | Search Location     | Text        |                        | Get Tweets (Process Flow) |
| searchTo... | Search Topic        | Text        |                        | Get Tweets (Process Flow) |

# Part 6: Text Mining

- Import table into Enterprise Miner
- Adjust number of clusters until results start to make sense
  - Make sure to look over the actual texts, not just the dominant word characteristics!
- Pick a good search term – quite often, people just aren't talking about certain things ...

# Part 6: Text Mining – Example Results

## Searching for 'Qantas'

- **4%:** TV related discussion, namely the Australian anti-censorship video trying to get on Qantas and the Spirit of Youth awards
- **41%:** Discussion about the Qantas lounge, posts of people in-transit and waiting for the flights / going home
- **22%:** Frequent flier points, Qantas club, Everyday Rewards
- **23%:** Qantas work-related discussion and industry issues (e.g. A380, working at Qantas)
- **8%:** Qantas cargo price fixing

## Searching for 'Virgin'

- **14%:** Delta joint venture, flights to LA
- **11%:** Virgin mobile technical discussion (e.g. 3G, network issues)
- **31%:** Virgin promotions discussion (e.g. iPhone, Virgin Lounge, pricing, tethering, etc)
- **30%:** General 'virgin' catchall (some general Virgin discussion, movies, being new to Twitter, etc)
- **13%:** Virgin Twitter promotion (apparently Virgin's running a Twitter competition)



**THE  
POWER  
TO KNOW®**

## Questions?

---

[Andrew.azzi@sas.com](mailto:Andrew.azzi@sas.com)



**THE  
POWER  
TO KNOW®**