TwitterVane

Administrators Guide

21 February 2013

Introduction

About TwitterVane

TwitterVane is a tool for leveraging the power of the crowd to select websites for web archiving. It enables curators to define web collections and identify trending Tweets and associated URLs that are relevant to those collections.

About this document

This document is the TwitterVane Administrators Guide. It describes how to control, monitor and maintain Tweet Stream and Tweet processing for the TwitterVane web application.

For installation, setup and configuration information, refer to the “TwitterVane System Installation Guide.doc”.

Where to find more information

The public resources for TwitterVane, including the source code, binaries and documentation, are available through the TwitterVane Open Source GitHub project:

https://github.com/ukwa/twittervane

System Overview

Purpose and scope

The application has been designed to stream Tweets directly from Twitter and store them for analysis. The analysis consists of expanding any shortened URLs associated with the Tweets using the Bitly expander service and associates the Tweets with the Web Collections that have been defined.

How does it work?

TwitterVane is divided into three major components

TweetView

TweetView is the curator web interface where users define the collections and search terms that will be used to collect Tweets. A series of reports is also provided to identify popular Tweets and top URLs by collection.

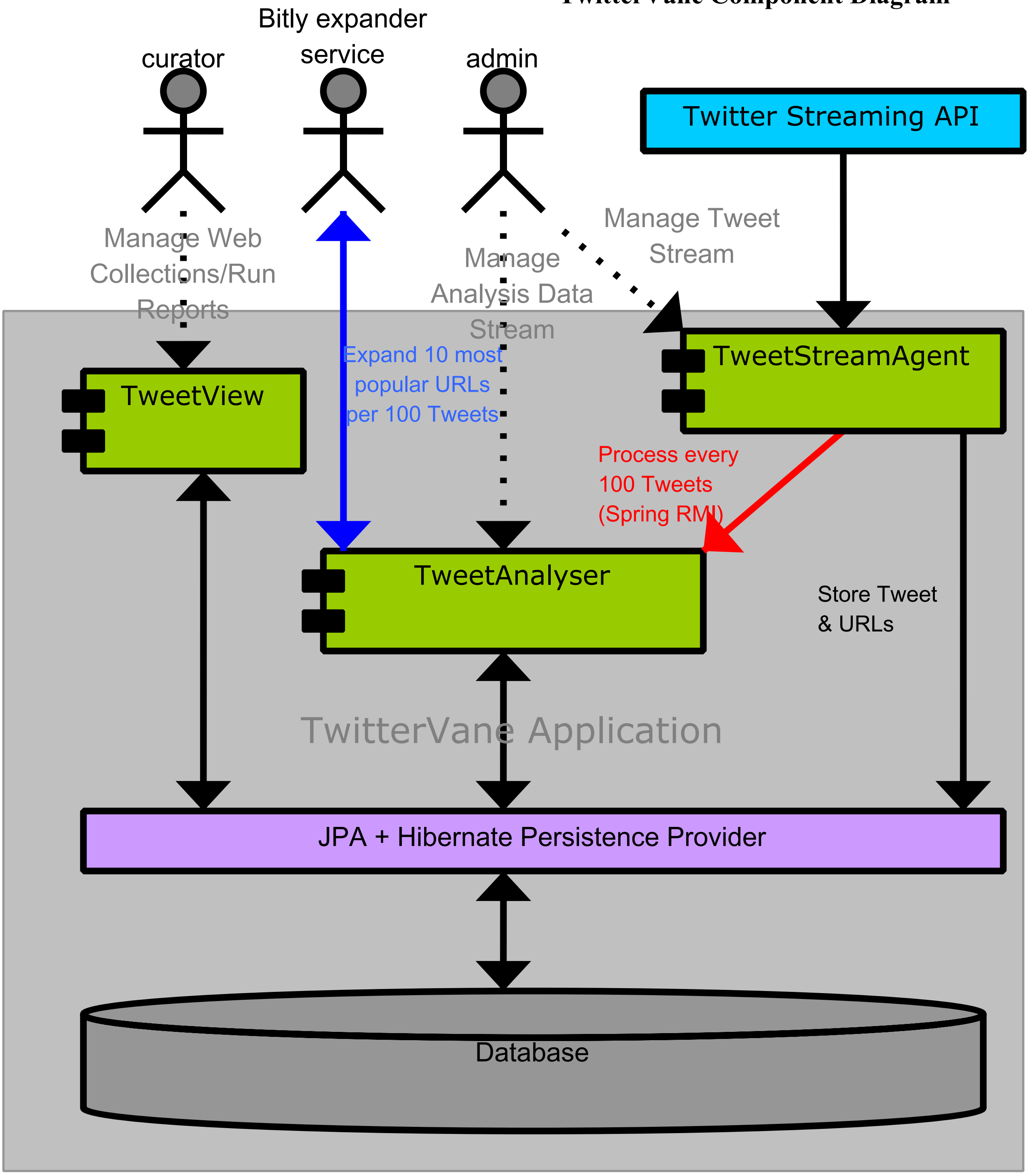
TweetStreamAgent

TweetStreamAgent streams Tweets directly from Twitter and stores them in the TwitterVane database for further analysis. It uses the search terms that have been defined by curators through *TweetView* to filter the Tweet stream. This is a management web interface that enables administrators to control and configure the Tweet stream.

TweetAnalyser

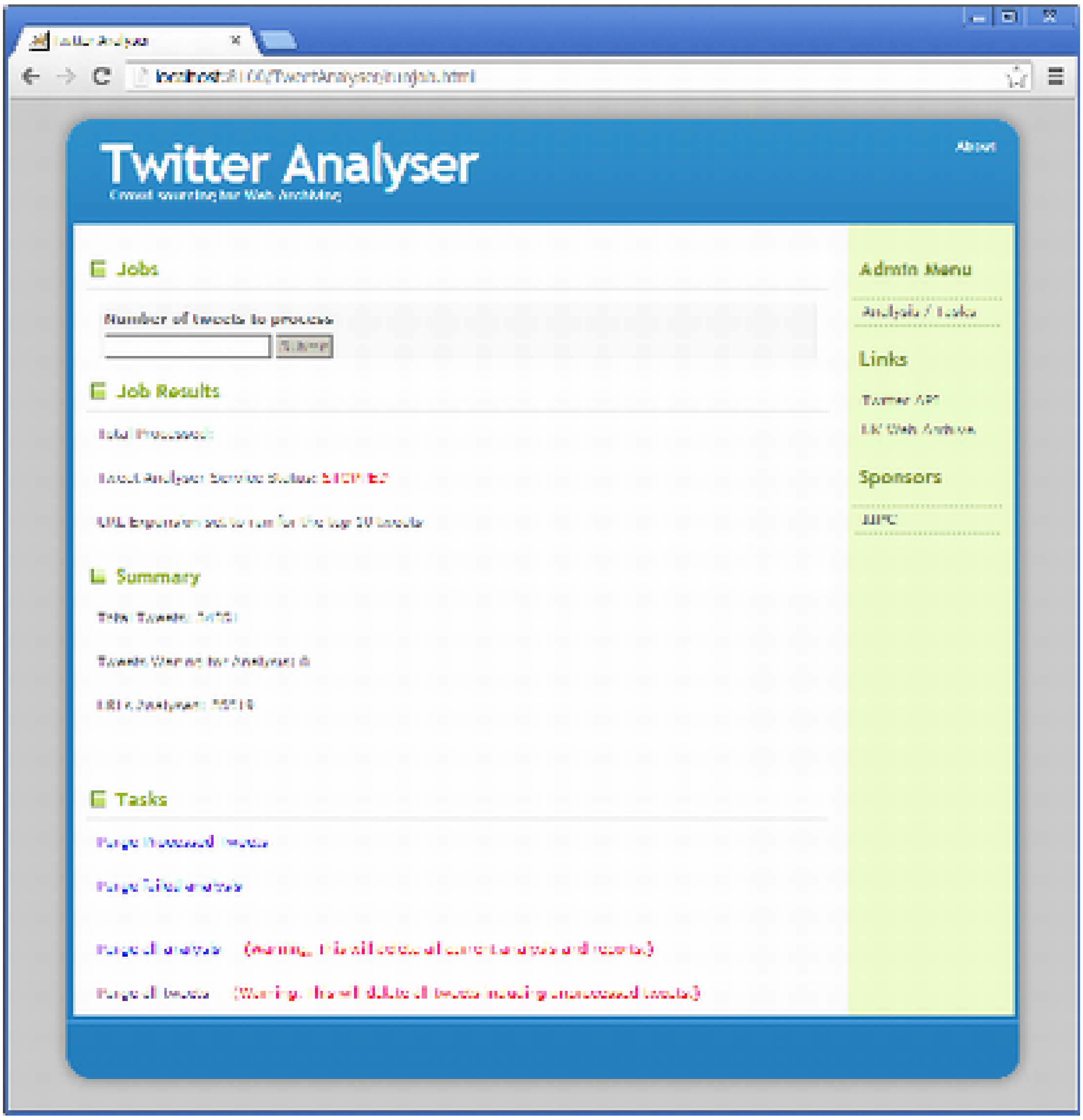
TweetAnalyser runs periodically, after a specific number of Tweets have been received from the Tweet stream, and expands any shortened URLs using Bitly. Any Tweet URLs are then resolved to the web collection using the collection’s search terms.

TwitterVane Component Diagram


Bitly expander service
curator
admin
Twitter Streaming API
Manage Tweet Stream
Manage Web Collections/Run Reports
Manage Analysis Data Stream
Expand 10 most popular URLs per 100 Tweets
TweetStreamAgent
TweetView
Process every 100 Tweets (Spring RMI)
TweetAnalyser
Store Tweet & URLs
TwitterVane Application
JPA + Hibernate Persistence Provider
Database

TweetAnalyser

The main page that controls the **TweetAnalyser** is pictured below:



Total Processed:

Tweet Analyser Service Status: STOPPED

URL Expansion set to run for the top 10 tweets

Total Tweets: [illegible]

Tweets Waiting for Analysis: 6

URLs Analysed: [illegible]

Purge Processed Tweets

Purge failed analysis

Purge all analysis (Warning, this will delete all current analysis and reports)

Purge all tweets (Warning, this will delete all tweets including unprocessed tweets)

Jobs

The **Jobs** section enables administrators to input the number of Tweets to be analysed. Enter a positive number in the **Number of tweets to process** field and click **Submit** to launch the processing. If the field is left blank, the TweetAnalyser will analyse all unprocessed Tweets in the database.

Job Results

This section displays the total number of Tweets that have been processed in the last processing run.

To preserve memory, the Tweets are processed in batches so that only a specified number of Tweets are loaded into memory at any given point in time. This number is specified in the configuration for the TweetAnalyser (refer to the "TwitterVane System Installation Guide.doc").

The **Total Tweets** figure displays the total number of Tweets in the database when the TweetAnalyser processing run was started. This figure is based on summary information gathered during the analysis. The information is collected to remove the need to perform expensive queries during the processing run.

The **Tweets Waiting for Analysis** figure displays the number of unprocessed Tweets.

The **URLs Analysed** figure displays the total number of URLs that have been analysed for all processed Tweets.

Tasks

The **Tasks** section enables the administrator to perform limited housekeeping operations on Tweet and URL data.

Purge Processed Tweets

WARNING: This action will delete all Tweets and associated URLs that have been processed.

Purge failed analysis

WARNING: This action deletes all URLs that encountered an error during analysis. It will not delete any Tweets from the database.

Purge all analysis

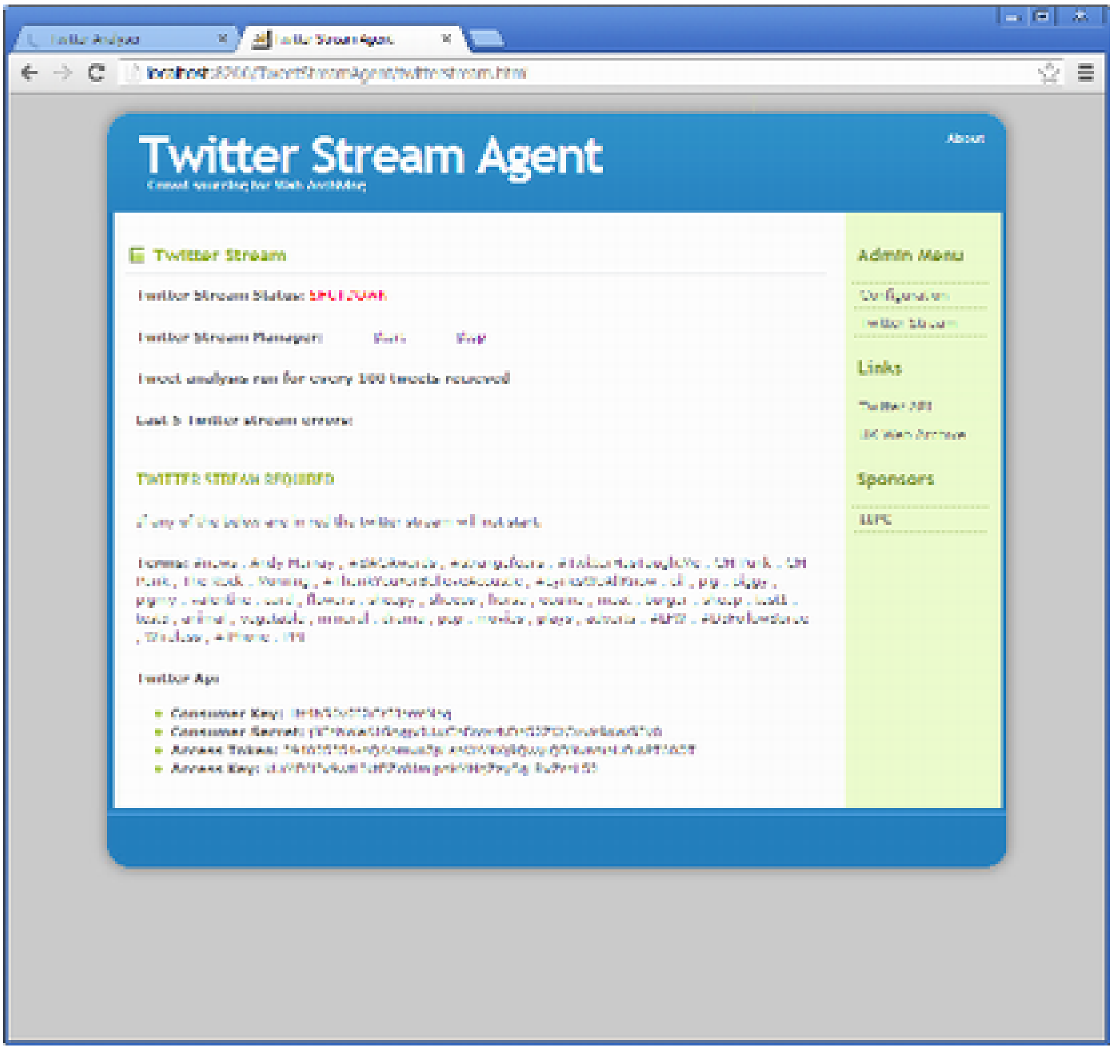
WARNING: This action deletes all URLs within the database. It will not delete any Tweets.

Purge all tweets

WARNING: This action deletes all Tweets and associated URLs from the database.

TweetStreamAgent

The main **TweetStreamAgent** control page is pictured below:



Twitter Analyser

Twitter Stream Agent

localhost:8080/TweetStreamAgent/twitterstream.html

Twitter Stream Agent

Crowd sourcing for Web Archiving

About

Twitter Stream

Twitter Stream Status: SHUTDOWN

Twitter Stream Manager: Start Stop

Tweet analyses run for every 100 tweets received

Last 5 Twitter stream errors:

TWITTER STREAM REQUIRED

If any of the below are in red the twitter stream will not start.

[illegible]

Twitter App

- Consumer Key: [illegible]
- Consumer Secret: [illegible]
- Access Token: [illegible]
- Access Key: [illegible]

Admin Menu

Configuration

Twitter Stream

Links

Twitter API

UK Web Archive

Sponsors

[illegible]

Twitter Stream

The **Twitter Stream** section enables the administrator to start and stop the Tweet stream from Twitter.

The **Twitter Stream Status** displays the current status of the Twitter Stream Agent. The status' are:

RUNNING

- the Twitter Stream consumer and Twitter4J Async Dispatcher threads are running

SHUTDOWN

- the Twitter Stream consumer and Twitter4J Async Dispatcher threads have been shutdown and cleanup has been performed

Start

This action will start the Tweet Stream daemon and transition the status from SHUTDOWN to RUNNING.

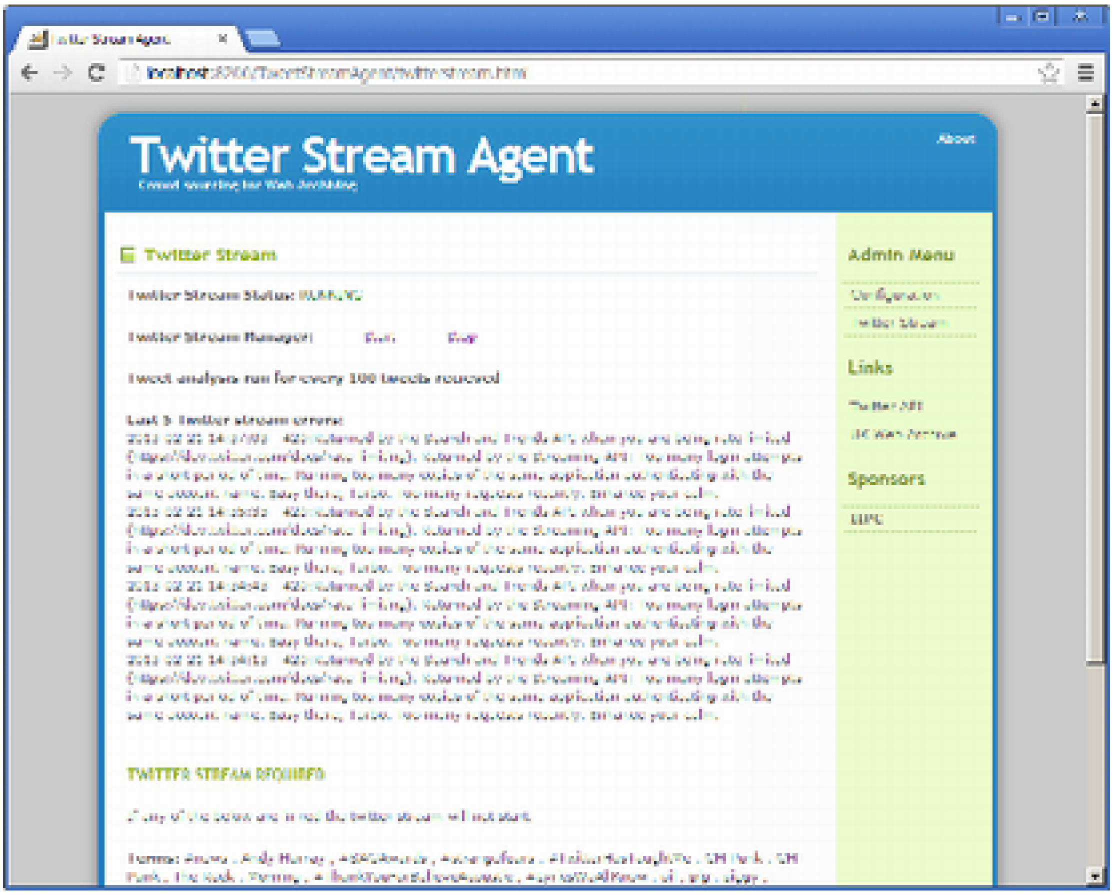
Stop

This action will stop the Tweet Stream daemon and transition the status from RUNNING to SHUTDOWN.

Note that **TweetStreamAgent** should be stopped prior to the deployment of the **TweetAnalyser** component.

This section also displays the currently configured Tweet analysis trigger value (refer to the "TwitterVane System Installation Guide.doc" for information on this setting).

A log of the last 5 errors encountered by the **TweetStreamAgent** is also displayed in this section. An example of the errors displayed when Twitter limits the number of stream connections is displayed below:



Twitter Stream

Twitter Stream Status: RUNNING

Twitter Stream Manager: Start Stop

Tweet analysis run for every 100 tweets received

Last 5 Twitter stream errors:
[illegible] 420: Returned by the Search and Trends API when you are being rate limited (https://dev.twitter.com/docs/rate-limiting). Returned by the Streaming API: Too many login attempts in a short period of time. Running too many copies of the same application authenticating with the same account name. Easy there, Turbo. Too many requests recently. Enhance your calm.
[illegible] 420: Returned by the Search and Trends API when you are being rate limited (https://dev.twitter.com/docs/rate-limiting). Returned by the Streaming API: Too many login attempts in a short period of time. Running too many copies of the same application authenticating with the same account name. Easy there, Turbo. Too many requests recently. Enhance your calm.
[illegible] 420: Returned by the Search and Trends API when you are being rate limited (https://dev.twitter.com/docs/rate-limiting). Returned by the Streaming API: Too many login attempts in a short period of time. Running too many copies of the same application authenticating with the same account name. Easy there, Turbo. Too many requests recently. Enhance your calm.
[illegible] 420: Returned by the Search and Trends API when you are being rate limited (https://dev.twitter.com/docs/rate-limiting). Returned by the Streaming API: Too many login attempts in a short period of time. Running too many copies of the same application authenticating with the same account name. Easy there, Turbo. Too many requests recently. Enhance your calm.

TWITTER STREAM REQUIRED

If any of the below are in red the twitter stream will not start.

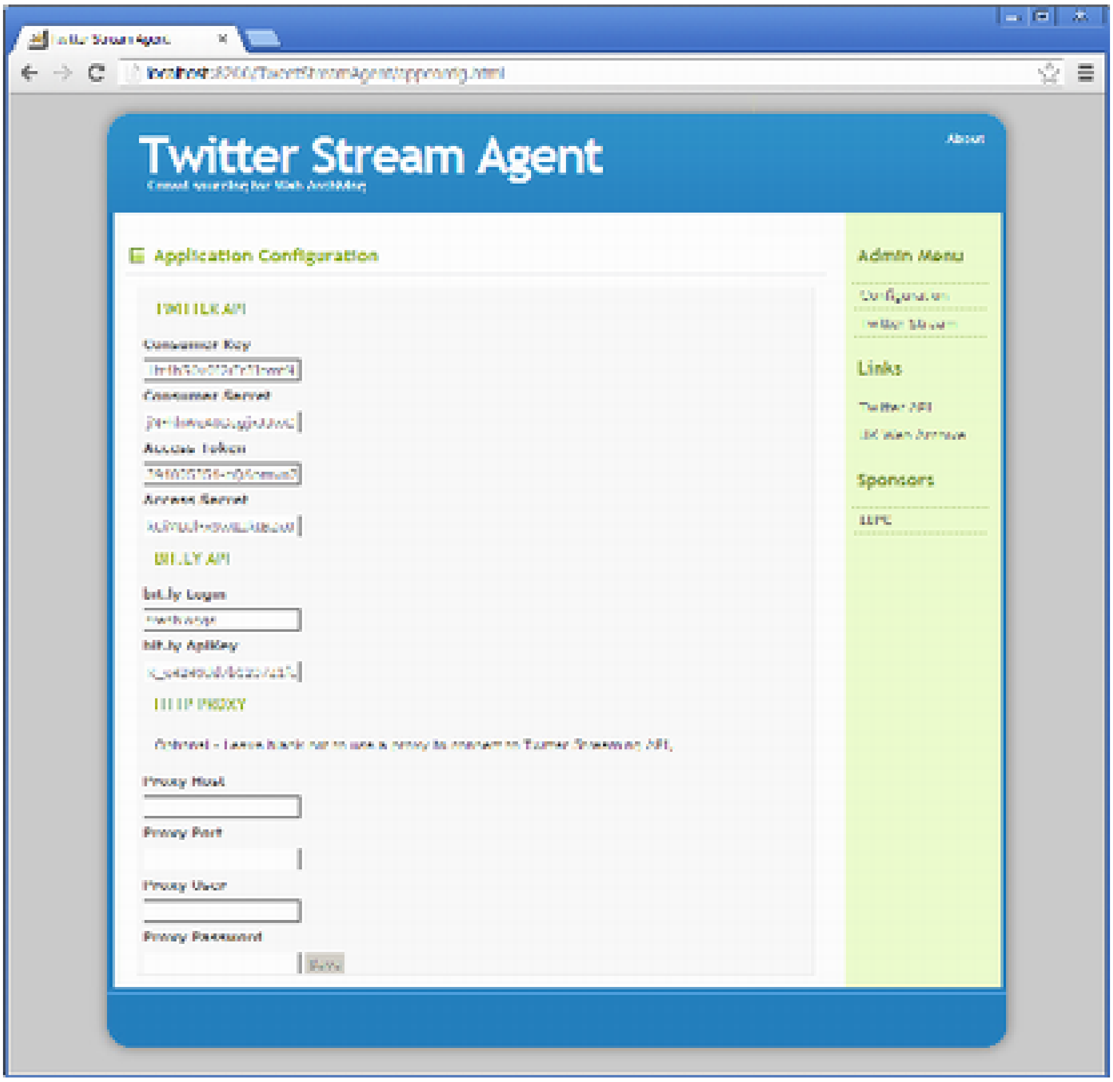
Terms: [illegible] , Andy Murray , [illegible]

At the end of the section, the currently loaded search terms are displayed for all web collections. These search terms are reloaded after the analysis is run so that any search term changes are incorporated into the filtered stream from Twitter.

Finally, the current Twitter API details are displayed. These are credentials and application keys that are registered with Twitter and Bitly and are used for Tweet stream access and URL expansion (as detailed in the next section).

Application Configuration

The **Application Configuration** page accessed by clicking the **Configuration** link in the right hand navigation menu:



Twitter Stream Agent

Application Configuration

TWITTER API

Consumer Key

Consumer Secret

Access Token

Access Secret

BITLY API

bit.ly Login

bit.ly ApiKey

HTTP PROXY

Optional - Leave blank not to use a proxy to connect to Twitter Streaming API.

Proxy Host

Proxy Port

Proxy User

Proxy Password

The configuration of TwitterVane is divided into three parts:

TWITTER API

Access to the Tweet stream is configured by four items that are generated by registering with the Twitter developer's site.

Consumer Key, Consumer Secret, Access Token, Access Secret

The consumer API key, consumer secret, access token and access secret define the credentials needed to access the Tweet stream and define the way in which the account can interact with the application (eg: permission level). These keys can be generated by following the instructions here:

https://dev.twitter.com/discussions/631

BITLY API

To expand shortened URLs, the **TweetAnalyser** component needs access to the Bitly expander service.

Bitly Login, Bitly API Key

The Bitly Login and Bitly API Key provide the user account and access token needed to submit requests to the Bitly API. Details on how to register and generate the key can be found here:

http://dev.bitly.com/authentication.html

HTTP PROXY

Http proxy authentication is only required when TwitterVane is situated behind a firewall (eg: on a development PC). The usual method to configure HTTP Proxy authentication consists of specifying HTTP proxy credentials as Java system parameters within the application server startup configuration.

Below is an example of the HTTP proxy configuration for a development Tomcat 6 server that is located behind a firewall:

```
-Dhttp.proxyUser=fredbloggs
-Dhttp.proxyPassword=testpassword
```

The above assumes that the system is already configured to route external requests via a proxy.

This configuration should only be applied to the application server that is running the **TweetStreamAgent** component. The remaining components do not require any proxy configuration.