



*Twitter Crawl



Edward Garcia

Petre Lukarov

Kina Winoto

Yongxu Zhang



* How to crawl.

There's a lot happening on my twitter feed!

The page shows 2,117 tweets, 10 following, and 4,18,547 followers. The 'Following' section lists several accounts:

- Jesus Christ (@jesus_Carpenter, Healer, God)
- Al Gore (@al Gore)
- Climate Reality (@ClimateReality Readify. It's not an opinion.)
- Betty Gephart (@bettygephart Rainy Day, Rock n' Roll Camp for Girls Los Angeles, person who is from L.A. and also lives in L.A.)
- Mona Tavakoli (@mona_tavakoli Not better, joke maker, kale encoder.)
- Marianne Williamson (@marwilliamson Marianne Williamson is the best selling author of A RETURN TO LOVE)
- Dalai Lama (@DalaiLama Welcome to the official twitter page of the Office of His Holiness the 14th Dalai Lama.)
- Gregory Page (@GregoryPage It's never tomorrow, it's only today - G. Page)
- Todd Hoffman (@toddhoffman Student of life, Teacher of Living, Manager of Integrity, Jogiologist Human, Constant Life Tourist, I Love Living)
- bushwells (@bushwells original gangster of cleveland. hairy, nice guy... slightly disturbed, but mostly in love with you)

I'm going to miss something...



* How to crawl.

I love crawling the web though...



Let's kill two birds...



* How to crawl.

Twitter Crawl will highlight keywords on a page that you're browsing and show you tweets having to do with that keyword.

Screenshot of a Wikipedia page for "United States". A yellow highlighter has been used to draw a large, irregular shape over the left side of the page, covering the sidebar and some of the main content area. The sidebar contains links like Main page, Contents, Featured content, Current events, Random article, Donate to Wikipedia, and Wikipedia Shop. It also lists languages available in the sidebar, including Afrikaans, Armenian, Azerbaijani, Belarusian, Bulgarian, Chinese, Czech, Danish, Dutch, English, French, German, Greek, Hebrew, Italian, Japanese, Korean, Polish, Portuguese, Russian, Spanish, Swedish, Turkish, Ukrainian, Vietnamese, and Welsh. The main content area includes sections for History, Government, Politics, Economy, Society, and Culture, along with a detailed table of statistics at the bottom right.

Relevant tweets:

From: TransWorldNews

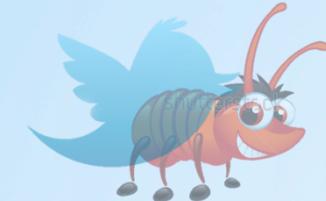
Text: United States freight trucking long distance market: New market research published <http://t.co/cjlN1oI3>



From: RubyGrace

Text: @WTFCrazyFacts: There are more women in China's military than there are people in the United States.



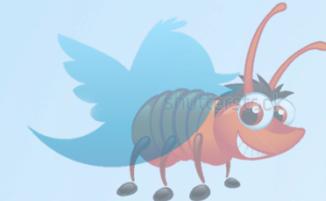


*How we did this.

*Overview:

- * Analyzed user's home timeline
 - * Extracted keywords from each tweet
 - * Created tags from these keywords
 - * Tagged the tweet
- * Parsed web page that a user is viewing
 - * Extracted keywords from the web page
 - * Highlight and display tweets that correspond to that keyword





*Technologies Used

- *Google App Engine
- *Twitter REST API
- *NLTK Python Library (for NLP: keyword extraction)
- *Chrome Extension Tools, Javascript





*Technologies Used

Twitter REST API

- * REST API uses Oauth API to make authorized GET requests
- * We power it using Google App Engine
- * Can retrieve any user's home timeline, controlling
 - * Count
 - * Replies allowed
 - * IDs returned

Screenshot of the Twitter Developers website showing the REST API page.

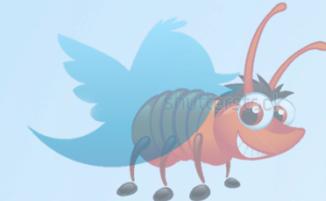
The page has a header with links for Developers, Search, API Health, Blog, Discussions, Documentation, and Sign in.

The main content area shows two sections:

- REST API version 1.1**: Described as the most recent version of the Twitter REST API. It includes links for API v1.1 Resources, Rate Limiting in API v1.1, Authenticating, and Announcement.
- REST API version 1**: Described as now deprecated and will cease functioning in the coming months. It includes a link to Review the deprecated version 1 API.

At the bottom, there are links for Follow @twitterapi, API Terms, API Status, Blog, Discussions, Documentation, and a note that it's a Drupal community site supported by Acquia.





*Technologies Used

NLTK Python Library (for NLP: keyword extraction)

- * NLTK is a natural language toolkit for Python
- * Very powerful
 - * Many corpuses (e.g. dictionaries)
 - * Create tree structures
 - * Cleans text
 - * Calculates various distributions

NLTK 2.0 documentation

[NEXT](#) | [MODULES](#) | [INDEX](#)

Natural Language Toolkit

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning.

Thanks to a hands-on guide introducing programming fundamentals alongside topics in computational linguistics, NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. Best of all, NLTK is a free, open source, community-driven project.

NLTK has been called "a wonderful tool for teaching, and working in, computational linguistics using Python," and "an amazing library to play with natural language."

[Natural Language Processing with Python](#) provides a practical introduction to programming for language processing. Written by the creators of NLTK, it guides the reader through the fundamentals of writing Python programs, working with corpora, categorizing text, analyzing linguistic structure, and more.

Some simple things you can do with NLTK

Tokenize and tag some text:

```
>>> import nltk  
>>> sentence = "At eight o'clock on Thursday morning  
.... Arthur didn't feel very good...."
```

TABLE OF CONTENTS

- NLTK News
- Installing NLTK
- Installing NLTK Data
- nltk Package
- Team NLTK

SEARCH

Enter search terms or a module, class or function name.





*Technologies Used

Chrome Extension Tools, Javascript

- * Uses jQuery to create tooltip highlighting
- * Sends requests to the Cloud
 - * Sends keywords and receives back tweets in json format





*Now for a demo.
..M...M...M...M...

