*Twitter Crawl

Edward Garcia
Petre Lukarov
Kina Winoto
Yongxu Zhang

Columbia University
Cloud Computing, Fall 2012
Final Project

*How to crawl.

There’s a lot happening on my twitter feed!

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.

Relevant tweets:

From: TransWorldNews
Text: United States freight trucking long distance market: New market research published http://t.co/cjlN1ol3

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

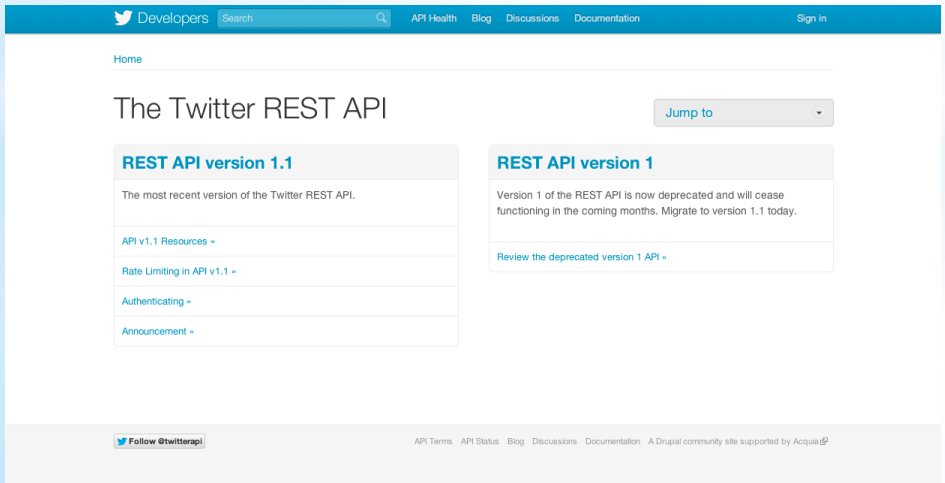
*Technologies Used

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

Technologies Used

Twitter REST API

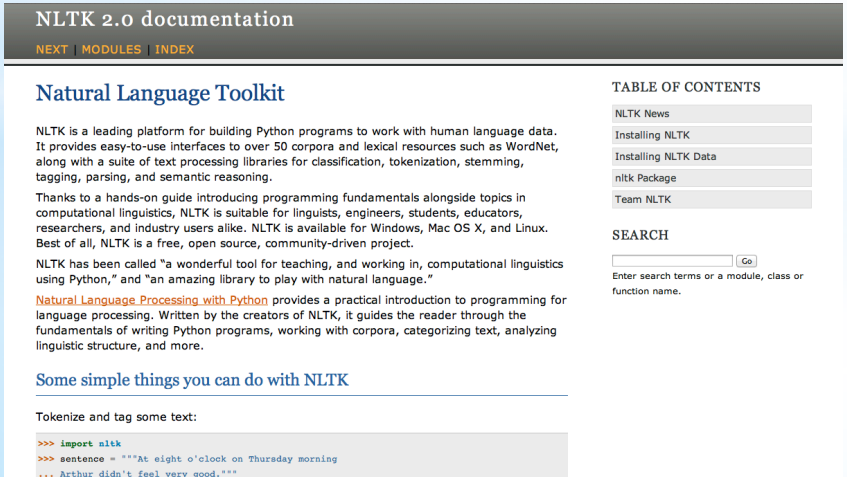
- REST API uses Oauth API to make authorized GET requests
- We power it using Google App Engine
- Can retrieve any user's home timeline, controlling
  - Count
  - Replies allowed
  - IDs returned



*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."""
```

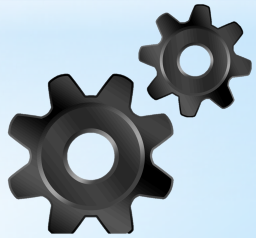
TABLE OF CONTENTS

NLTK News
Installing NLTK
Installing NLTK Data
nltk Package
Team NLTK

SEARCH

Go

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.