

# Twitter Keywords Search



Tianyi Tang

<http://www.tty8128.com>

<https://github.com/8128/TwitterKeywordSearch>

# Products

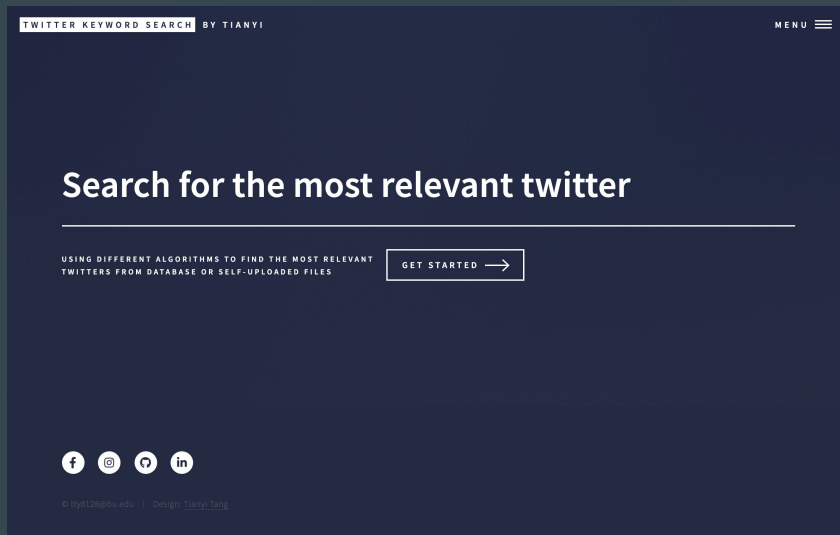
Command line tool

```
-----
|                               |
|           Twitter Keyword Searcher           |
|       Developed by Tianyi Tang               |
|                               |
|       return the top relevant twitter with your keywords       |
|                               |
|-----|
Enter 1 to enter your own CSV file with absolute path,
enter anything else to use default CSV
2
Loading Default CSV File
1041776 twitters loaded

Start to create reversed index map
Reverse Index time: 4744ms

Please Input Your Keywords.
Split them using spaces, and end with enter
```

Spring Boot Website <http://www.tty8128.com>



# Algorithm - Matching

The keywords will be stored into HashSet, each for  $O(1)$  time complexity

HashSet:

- Implements Set Interface.
- Underlying data structure for HashSet is hashtable.
- As it implements the Set Interface, duplicate values are not allowed.
- Objects that you insert in HashSet are not guaranteed to be inserted in same order. Objects are inserted based on their hash code.
- NULL elements are allowed in HashSet.
- HashSet also implements Serializable and Cloneable interfaces.

# Algorithm - Matching

Split the twitter into words, and iterate through every single word, use the contains function to check whether the word match any word inside of the hashset, each word for  $O(1)$  time complexity

An relevance value will be generated using the number of matching.

# Algorithm

## Heap Sort

Initialize a new PriorityQueue(which implements the heap in Java), set the size to the needed size, and override the Comparator to compare them with the relevance value, and make it a min heap

Iterate through all the twitters and compute their relevance values. When the heap size reaches the needed size, and after that the new twitter's relevance value is larger than the min value in the heap, then poll out the min twitter and add the new twitter

## Bucket Sort

Because the twitter's length is smaller than 140, the largest relevance value will not be larger than 140, so buckets of 140 will be initialized.

Iterate through all the twitter and calculate the relevance between the keywords and twitters, then use it to store the twitter to the

bucket[relevance value]

Finally, iterate from the last index of the bucket, add twitters until the list reaches the needed size

# Improvement

- If users want to search several times ?
- Use reverse index

After load all data to list, iterate through all the twitters and store them to a HashMap

HashMap is a Map based collection class that is used for storing Key & value pairs, it is denoted as `HashMap<Key, Value>` or `HashMap<K, V>`. This class makes no guarantees as to the order of the map. It is similar to the Hashtable class except that it is unsynchronized and permits nulls(null values and null key).

# HashMap

HashMap<String, HashMap<Integer, Integer>>

First Map: Key - Word / Value - Second HashMap

Second Map: Key - index of twitter in the data list / Value - this word's frequency

When keywords come in, new a HashMap to store twitter and overall frequency,

Get all the keys using keySet() function, make it a list and sort by frequency using Java's Collections.sort and BucketSort

# Website development

<http://www.tty8128.com>

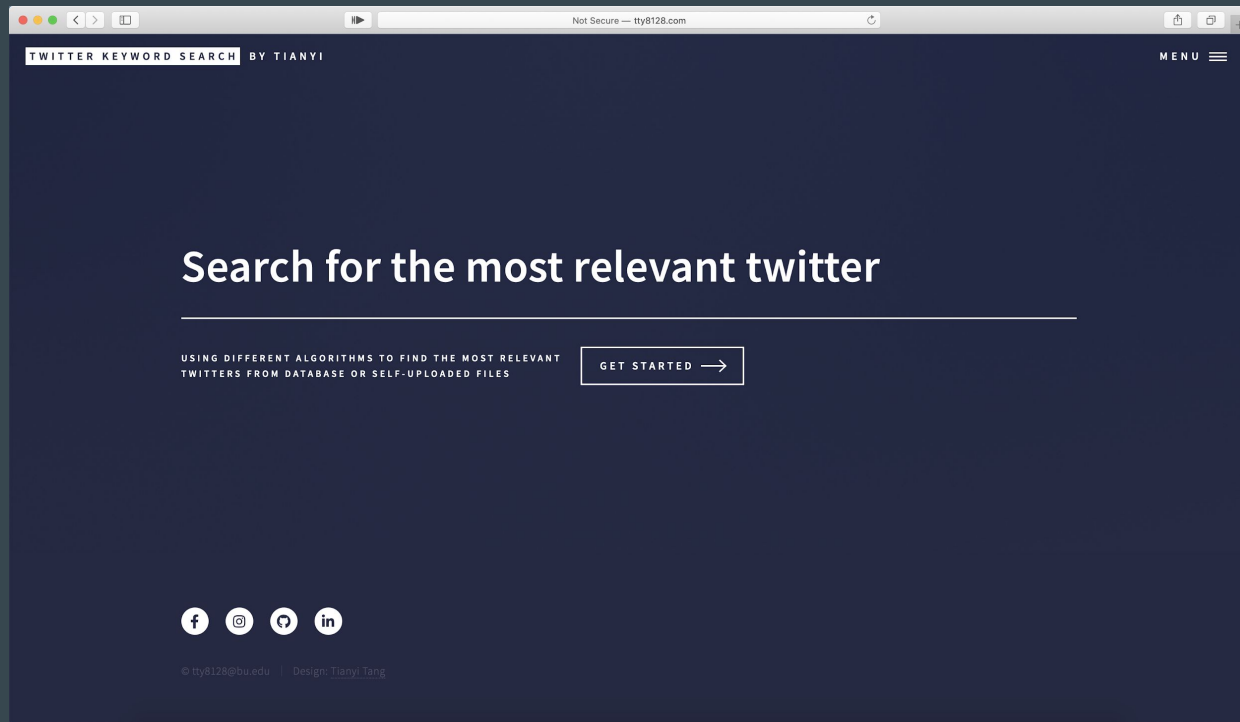
- Using Spring Boot Framework, MyBatis to connect to MySQL
- MySQL (t\_twitter / t\_twitter\_temp)
- HTML5/CSS/thymeleaf

Con:

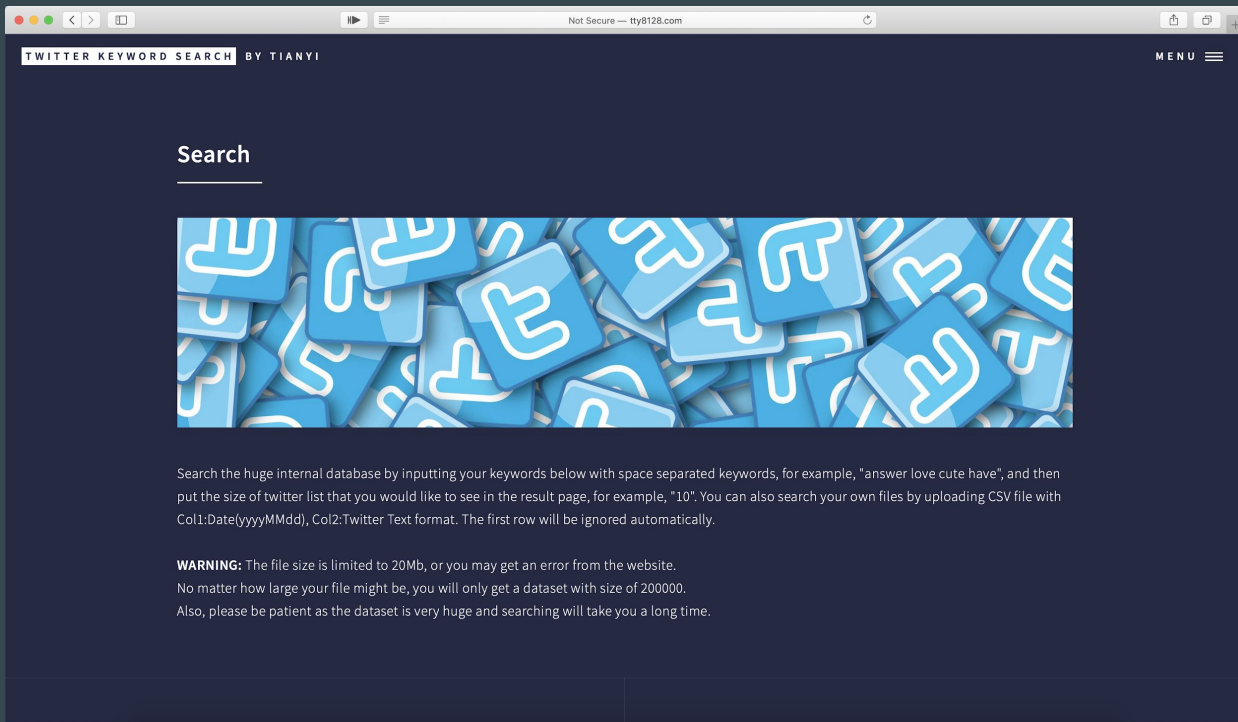
- Cannot use MySQL database on SCC, so it is deployed on Amazon EC2
- Large dataset may cause OOM



# Index Page



# Index Page



# Search Page

**WARNING:** The file size is limited to 20Mb, or you may get an error from the website.  
No matter how large your file might be, you will only get a dataset with size of 200000.  
Also, please be patient as the dataset is very huge and searching will take you a long time.

### Search Internal Dataset

KEYWORDS

Input your key words and separate them using spaces...

Size

INTERNAL SEARCH

CLEAR

### Search External Dataset

Choose File no file selected





KEYWORDS

Input your key words and separate them using spaces...

Size

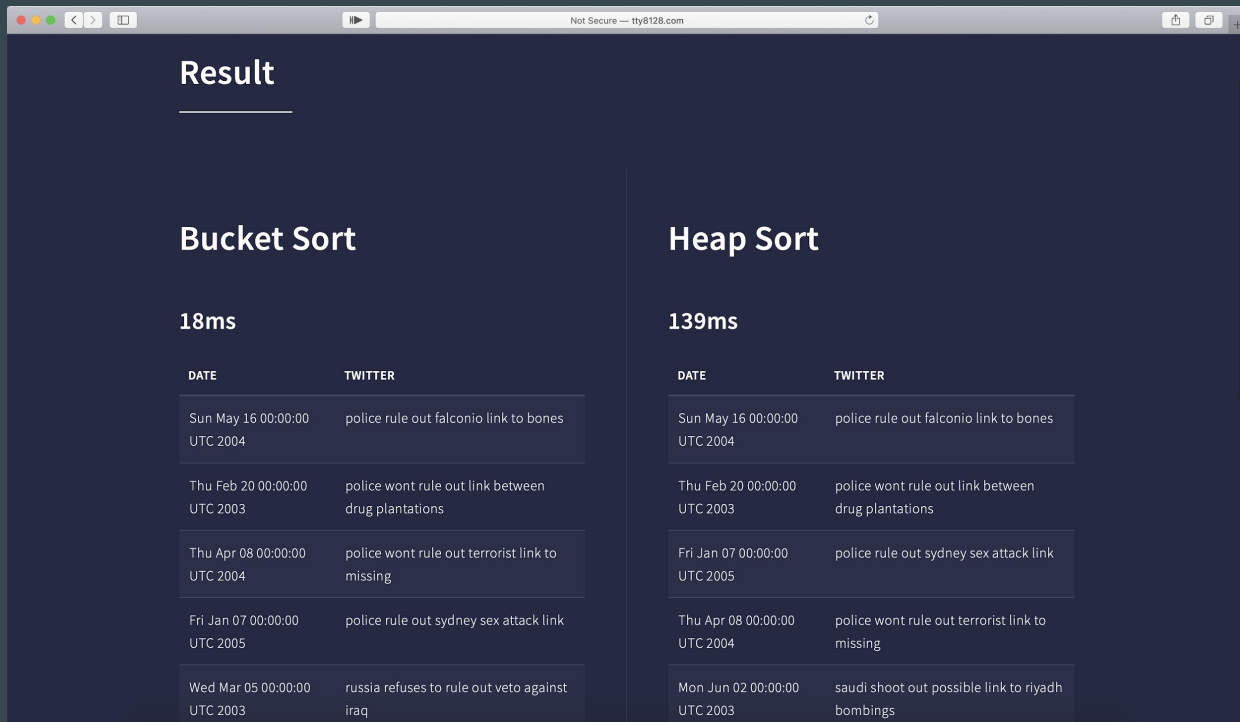
EXTERNAL SEARCH

CLEAR



© ty8128@bu.edu | Design: [Tianyi Tang](#)

# Result Page



Result																									
<h2>Bucket Sort</h2> <p>18ms</p> <table><thead><tr><th>DATE</th><th>TWITTER</th></tr></thead><tbody><tr><td>Sun May 16 00:00:00 UTC 2004</td><td>police rule out falconio link to bones</td></tr><tr><td>Thu Feb 20 00:00:00 UTC 2003</td><td>police wont rule out link between drug plantations</td></tr><tr><td>Thu Apr 08 00:00:00 UTC 2004</td><td>police wont rule out terrorist link to missing</td></tr><tr><td>Fri Jan 07 00:00:00 UTC 2005</td><td>police rule out sydney sex attack link</td></tr><tr><td>Wed Mar 05 00:00:00 UTC 2003</td><td>russia refuses to rule out veto against iraq</td></tr></tbody></table>	DATE	TWITTER	Sun May 16 00:00:00 UTC 2004	police rule out falconio link to bones	Thu Feb 20 00:00:00 UTC 2003	police wont rule out link between drug plantations	Thu Apr 08 00:00:00 UTC 2004	police wont rule out terrorist link to missing	Fri Jan 07 00:00:00 UTC 2005	police rule out sydney sex attack link	Wed Mar 05 00:00:00 UTC 2003	russia refuses to rule out veto against iraq	<h2>Heap Sort</h2> <p>139ms</p> <table><thead><tr><th>DATE</th><th>TWITTER</th></tr></thead><tbody><tr><td>Sun May 16 00:00:00 UTC 2004</td><td>police rule out falconio link to bones</td></tr><tr><td>Thu Feb 20 00:00:00 UTC 2003</td><td>police wont rule out link between drug plantations</td></tr><tr><td>Fri Jan 07 00:00:00 UTC 2005</td><td>police rule out sydney sex attack link</td></tr><tr><td>Thu Apr 08 00:00:00 UTC 2004</td><td>police wont rule out terrorist link to missing</td></tr><tr><td>Mon Jun 02 00:00:00 UTC 2003</td><td>saudi shoot out possible link to riyadh bombings</td></tr></tbody></table>	DATE	TWITTER	Sun May 16 00:00:00 UTC 2004	police rule out falconio link to bones	Thu Feb 20 00:00:00 UTC 2003	police wont rule out link between drug plantations	Fri Jan 07 00:00:00 UTC 2005	police rule out sydney sex attack link	Thu Apr 08 00:00:00 UTC 2004	police wont rule out terrorist link to missing	Mon Jun 02 00:00:00 UTC 2003	saudi shoot out possible link to riyadh bombings
DATE	TWITTER																								
Sun May 16 00:00:00 UTC 2004	police rule out falconio link to bones																								
Thu Feb 20 00:00:00 UTC 2003	police wont rule out link between drug plantations																								
Thu Apr 08 00:00:00 UTC 2004	police wont rule out terrorist link to missing																								
Fri Jan 07 00:00:00 UTC 2005	police rule out sydney sex attack link																								
Wed Mar 05 00:00:00 UTC 2003	russia refuses to rule out veto against iraq																								
DATE	TWITTER																								
Sun May 16 00:00:00 UTC 2004	police rule out falconio link to bones																								
Thu Feb 20 00:00:00 UTC 2003	police wont rule out link between drug plantations																								
Fri Jan 07 00:00:00 UTC 2005	police rule out sydney sex attack link																								
Thu Apr 08 00:00:00 UTC 2004	police wont rule out terrorist link to missing																								
Mon Jun 02 00:00:00 UTC 2003	saudi shoot out possible link to riyadh bombings																								

# Contact Page

TWITTER KEYWORD SEARCH BY TIANYI

MENU

## Advice?

If you have any advice for my project, please inform me!

NAME

EMAIL

MESSAGE

SEND MESSAGE

CLEAR

Email

tty8128@bu.edu

Phone

(+1) 857-498-0354

Address

40 Malvern St  
Boston, MA, 02134  
United States of America

**Thank You**