Weibo Top History

# Purpose

The purpose of the this web application is to record and analyze the Weibo Top topics and gather it as long as possible to become the source of big data analysis.

This data has huge potential. Later it can be used to categorize the topic like “positive”, “negative”, “entertainment”, “politics”, etc. which can help to analyze Chinese people’s online behaviors. Combined with AI and data science technology, it can have many useful applications.

# Requirements

So the requirements will be made to stages.

**1st stage** – create a crawler to which can fetch data from Weibo Top page (<https://s.weibo.com/top/summary?cate=realtimehot>) and put the top 50 topics into database. Every 30 mins (or shorter time 10 mins) triggered by web application check the top topics and then update the timing into database.

Database Design

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| id | topic | content | category | highest\_rank | hotpoints | first\_time | last\_time | duration |
|  |  |  |  |  |  |  |  |  |

* [id, VARCHAR(16)] is primary key, the format is YYYYMMDDHHMMSSII;
* [topic, VARCHAR(150)] is the title of the topic, this key should be indexed to improve performance;
* [content, VARCHAR(300)] is the short content of the topic;
* [category, VARCHAR(100)] for 3rd stage usage;
* [highestrank, INT UNSIGNED] is the highest rank the topic ever reached;
* [hotpoints, INT UNSIGNED] is the highest hot points the topic ever had;
* [firsttime, DATETIME] when the topic entered the Weibo top page;
* [lasttime, DATETIME] the last time the topic stays in the Weibo top page;
* [duration, INT UNSIGNED] is duration of the topic staying in Weibo Top in seconds;

Total size of one row is 16+150+300+100+4+4+8+8+4=594 bytes.

**2nd stage** – build a website which can show the current **Weibo top 50**, **historical longest duration Weibo top 50**(in the database it will store more data for future data analysis) and **hot points ranking in history**.

**3rd stage** – utilize data science technology to give some insights.

# System Components

* Web crawler tool to craw the Weibo Top page;
* MariaDB database;
* Web application to show the data;
* …