Abstract:

The impact of social media has despicable impact in our social and personal life. Particularly, microblogging services like twitter with its tiny 140 character limited post carry’s a huge important and meaningful information. Millions of user post intersecting topic on varies subject/communities/topics.

Perphaps, searching on a topic gives results from various tweets from different channels/posts/users/accounts. Going through all those tweets, replays, hashtags and comments may be very burdensome and time consuming. Also, not all tweets are quality one. The action performed after going through of all tweets such as post, replay, follow etc for different tweet accounts may be again burdensome. In some situations, user wants to share the same message/post/replay to all those community posts. Its only possible by individually sending them.

This paper focuses on detecting the community account most relevant hashtags and identify the most suitable administrators to follow using tweet scraping technique.

This paper purpose a tweet scraping technique to filter a group of tweets from a search result. The tweets are sorted in the order of importance using the repetitive score model technique. It also support in the extracting of key references such as hashtags (literals starting with ‘#’), replays (literals starting with @) and hyperlinks. These key references are sorted in the most occurrence wise.

Thus, this technique eases the process of finding the same community group under different account or users. System is to be developed such that a single tweet can be posted to all the community hashtags. A group of hashtags with common agenda is said to be community message.

Introduction :

The rise of social media has huge impact on communication and sharing the information. Tweeter has posts 5 million tweets everyday. Tweeters accounts are run by individual or community. The tweets are posted based on the some incidents, events or on the interesting topics. Such type of post are concerned for some group of user. If the user interested in a particular item, event or topic ,he/she can often provide a few relevant keywords to a tweet search function. A huge tweet list from various users and account or community account pops up. These tweets are filtered according to tweets search algorithm. The tweeter is based on various parameter that hidden from the user. If a user wants to find appropriate community post form all those post, then it is very difficult to go through all the posts. The message from different community for the same topic will be there. This could cause the user to repeatedly read the same message. And also of the user wants to replay a message to users for such a post , then he has to replay to all such tweets one after the other, which is burdensome. If he is interested to follow such a user or group from such community, then he has separately follow them one after the other.

Literature survey:

Nowadays, quite a lot of researchers are working on extracting information about types of events, entities or relationships from textual data. A form of information extraction is text mining, an information retrieval task. Scrapers are basically adopted to transform unstructured data and save them in structured databases. In screen scraping, a special form of scraping, a program extracts information from the display output of another program. So that, the output which is scraped is created for the end user for additional functionalities. Similarly, there are web scrapers which extracts the information from the website and process to give meaningful information.

There are different tweeter scrapers available in the market. But they differ in functionality. In this paper, we focus on twitter scrapers that extract textual information from twitter resource and identify the hashtags and Administrators of the group and sort them.

Twitter background:

Background:

Twitter is a microblogging service that was founded in early 2006 to enable people to share short textual messages “tweets” with others in the system. Because the system was originally designed for tweets to be shared via SMS, the maximum length of a tweet is 140 characters. Though the service evolved to include more uses besides SMS, such as web and desktop clients, this limitation persisted, and so was re-narrated as a feature. Twitter’s Creative Director Biz Stone argues, “creativity comes from constraint” [1].

## Twitter conventions

Twitter members are restricted to communicating in only 140 characters. As members appreciated the innovation and its affordances, a progression of idea that enabled clients to add structure and unique presentation of tweets. For instance, users created approaches to reference different users, merged on special characters to convey the meaning, and converged on topics to be talk upon. Twitter participants began using the @user syntax to refer to specific users (e.g., @twiiterScrap) to address one another.

Topics or event are identified through the special character hashtag(#) and a topic/event name following it.The practice of using keywords to label tweets most likely parallels the use of “tags” to freely categorize web content. Tagging gained visibility with social bookmarking [2].

Early Twitter clients additionally started giving connects to outside substance by incorporating the URL in their tweets. Since URLs are ordinarily long, they take up an excessive number of characters. Subsequently, individuals began utilizing "URL shorteners" (e.g., http://bit.ly) to create remarkable, condensed URLs that divert to the coveted site.

Twitter developer API:

Twitter has given REST API's which can be utilized by third party developers to access and read Twitter information. They have likewise given a Streaming API which can be utilized to get to Twitter Data in real time. Most of the product written to get to Twitter information give a library which works as a wrapper around Twitters Search. The Search API can dig for tweets posted previously, though Streaming will be speedy and catches tweets as they are really posted.

With Twitter's Search API you can only sent 180 Requests every 15 minutes. With a maximum number of 100 tweets per Request this means you can mine for 4 x 180 x 100 = 72.000 tweets per hour. By using Twitter API you are not limited by this number. Search API is that you can access Tweets written in the past 7 days. This is availability for a free tier Twitter API. Eventually you can upgrade to professional tier for more resources accessibility[3].

Proposed methodology:

Twitter scrap:

Algorithm :

Input: S is the search key to be searched in tweeter

Output: ST sorted tweets

Function TweeterScrap(S)

tweets:= get tweet results form Tweeter API

tweetScore:=0

For Each tweetText in tweets do

tweetScore:= repetitive-score-model(tweetText)

ST:=sort the tweets according to score(tweetScore)

Return ST

Algorithm :

Input: T is the list of tweets

Output: HL sorted list for hash list. AL sorted list for Assignee list. LL sorted list for hyperlink list.

Function sortReferences(T)

tweets:= get tweet results from Tweeter API

For Each tweetText in T do

HTL:=Extract hastags from TweetText

ASL:=extract Assignee from TweetText

HPL:=Extract hyperlink from TweetText

HL:=sort the list (HTL)

AL:=sort the list (ASL)

LL:=sort the list (HPL)

Return HL,AL,LL

Architectural diagram:

Twitter Developer API

Twitter Scrap

Algorithm

UI Actions

Follow, Post, Comment

Filtered Recommended twitters

Search

UI

User login

Twitter scrap Flow Chart:

Twitter data

Selection of Tweets

Tokenization & Extraction

Repetitive score model(tweets)

Filtering tweets

Twitter scrap algorithm:

Conclusions: In this paper we have presented a twitter scrap technique that eases the users to find the relevant and prioritized tweets. It also enables users to see the popular hashtags to post on. It suggest the users to follow the most popular administrator on the given domain. This paper provides the functionality to post single tweet to many domains(hashtags).

As for the future work, we can consider language wise tweeter scrap and limiting it to regional ends.

References:

[1] Zinko, C. What is Biz Stone doing? *San Francisco Chronicle*, K-1, April 5, 2009.

[2] Golder, S. and B.A. Huberman. Usage Patterns of Collaborative Tagging Systems. *Journal of Information Science* 32, 2 (2006), 198-208.

1. [4]. Kobayashi, M. & Takeda, K. Information retrieval on the web. ACM Comput. Surv. 2000; 32: 144-173.

http://dx.doi.org/10.1145/358923.358934