Overview

1. Project Purpose and Background
2. Twitter crawling script
3. Web crawling script
4. Reporting Service
The digital archive is a database to collect large amounts of articles about large events and to summarize those events.
Project Purpose

The overall objective of this project was to build a website front end to help the client automate the web archiving process for their big data doctorate research.
A Quick Walk Through
input data
Brings us to yourtwapper

<table>
<thead>
<tr>
<th>Archive ID</th>
<th>Keyword / Hashtag</th>
<th>Description</th>
<th>Tags</th>
<th>Screen Name</th>
<th>Count</th>
<th>Create Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>#WalterScott</td>
<td>Shooting - Walter Scott, SC</td>
<td></td>
<td>ideal_vt</td>
<td>5472</td>
<td>Thu, 09 Apr 2015 13:32:13 -0400</td>
</tr>
<tr>
<td>3</td>
<td>VirginiaTech</td>
<td>VT</td>
<td>#VT</td>
<td>newsEventDetect</td>
<td>2520</td>
<td>Fri, 17 Apr 2015 11:14:22 -0400</td>
</tr>
<tr>
<td>12</td>
<td>hillary clinton</td>
<td>hillary running for president</td>
<td>#hillary16 #clinton16</td>
<td></td>
<td>10979</td>
<td>Wed, 22 Apr 2015 19:41:35 -0400</td>
</tr>
<tr>
<td>11</td>
<td>test</td>
<td>DESCRIPTION</td>
<td></td>
<td></td>
<td>19946</td>
<td>Wed, 22 Apr 2015 18:29:29 -0400</td>
</tr>
</tbody>
</table>
some of the previous twitter searches we have done
focused crawler output
Your request has been processed!

You can view your TwapperKeeper [HERE](#)
You can view your archived website [HERE](#)
Problems Faced

Cloud9/Bluehost
- Blocking of PHP Functions
- Lack of SSH

Calling PHP Scripts non locally
Calling Python Scripts from Flask
Acknowledgements

Mohamed Magdy Gharib Farag
● project sponsor
● mmagdy@vt.edu

PhD. Edward A. Fox
● class professor
● eafox@vt.edu

Sunshin Lee
● extra resource
● sslee777@vt.edu