On the Information Content Value of Social Media: A System for Real-Time Data Collection and Analysis

Advisor: Manolis Kaparakis **Syed Mansoor Alam & Ross Petchler '12** Quantitative Analysis Center, Wesleyan University

Introduction

Motivation: To assess the practical value of social media content and data potential. To build a system to consume and process this content, and develop related methods in programming, analysis and presentation.

About The Project: This presentation presents the ongoing development of the *Ducky* system. Currently focused on use of the Twitter API, we collect and analyze thousands of targeted tweets each day. We have begun web development focused on data presentation and accessibility. Our data frequently supports various academic studies.

Topics of Interest

- How do we trace points of influence on social media networks, and how does that influence translate to real life?
- Methods in tracking and forecasting user behavior
- How to best leverage visualization in information display

Methods

Basic Workflow: Twitter API Example

We are currently streaming the 113th Congress…

- Data collection runs continuously and passes data through through scripts for additional parsing, analysis, and preparation for presentation. (i.e. sentiment analysis, geocoding)
- Data is stored in tables, and exported based on researcher specification (i.e. which variables and format).
- Data is then ready for further analysis by researcher
- After running the initial Stream script, subsequent processes are largely automated.
- The project webpage updates every hour, populating graphs and visualizations with new data.

System Structure

Basic Structure and Workflow:

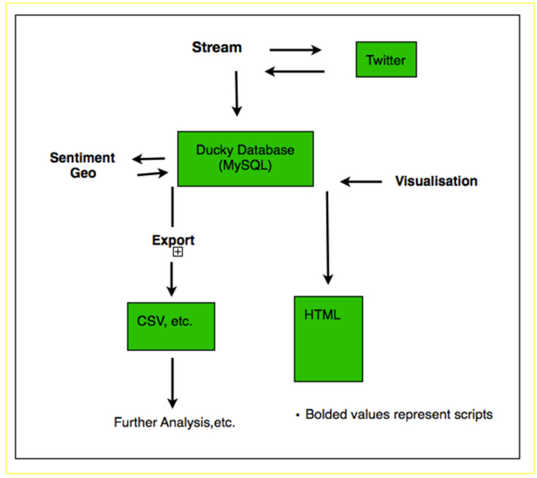


Figure I: Basic workflow and structure of the system, using the Twitter API as an example.

Sample Features and Considerations:

- Database Type: Relational vs. Key-Based, Table Structures
- Primary Language: Python vs. PHP
- Error handling – email alerts, exponential backoff, fail-safe scripts (see Figure II)
- Visualization

Evaluating and Presenting Content

Limitations

- Do measures in a virtual environment translate accurately to the real world?
- How do the limits of social media (i.e. 140 character Tweets) influence content value?
- Complex media data is difficult to analyze computationally (i.e. Vine, Instagram, YouTube)

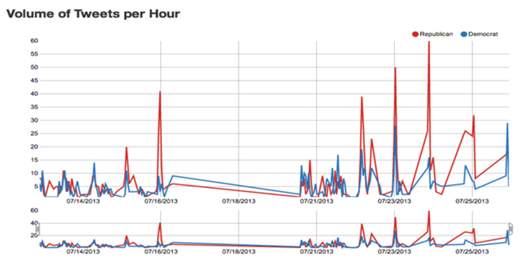


Figure II: The gap in the middle is caused by rate-limiting stream abortion. We have developed scripts that allow us to retroactively retrieve this missing data.

Republican Congress Members

Hashtag	Frequency
Obamacare	88
tcot	31
4jobs	29
NSA	27
StopGovtAbuse	26
treyonplane	20
BuyAmerican	18
immigration	16
NY19	16
FairnessForAll	15

Figure III: Most frequent hashtags presented in table form, dynamically updated.

Cont.

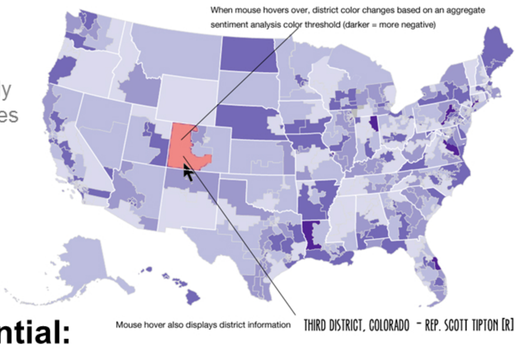


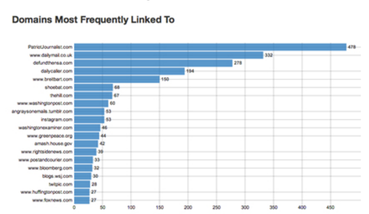
Figure IV: A sample dynamically updated choropleth with features allowing users to interact with the graph to obtain more info.

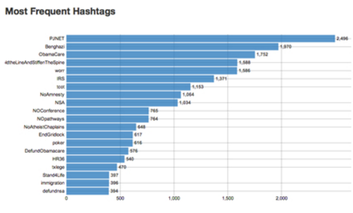
Data Uses and Potential:

- To showcase consumer behavior and public opinion, provide context to various events.
- Analyze how information travels across users.
- Provide data for campaigns to identify key issues in different constituency demographics, possibly negating the need of costly and ineffective canvassing
- Disaster Response

Beyond Social Media [examples]:

- Retrieve and process data from computational lab simulations (i.e. physics)
- Web mining of text and images
- Financial monitoring and forecasting
- Parallel Data Streams (i.e. finance and carbon emissions)





Future Directions

Future Directions:

- Web-app development for simple client-side initiation and retrieval of data
- Greater focus on utilizing web-visualization techniques for effective presentation (i.e. web sockets)
- Incorporation of machine learning techniques

Discussion:

- The evolving state of data sources and availability requires us to adapt our systems and designs. Otherwise, technology will inevitably render our operations obsolete.
- How can we best collect and process heterogeneous data (i.e. blogs) ?
- What distinguishes data from content?

ADDITIONAL REFERENCE:
qacprojects.wesleyan.edu

Acknowledgements: I would like to thank Professor Kaparakis for this opportunity (and unthank him for beating me in squash), Ross Petchler '12 for his knowledge & support and Eric Stephen for the company.