The Social Media Macroscope

Yun, J. T., Vance, N. P., Wang, C., Troy, J., Marini, L., Booth, R., Nelson, T., Hetrick, A., & Hodgkins, H.

Summary

In recent years, the explosion of social media platforms and the public collection of social data has brought forth a growing desire and need for research capabilities in the realm of social media and social data analytics. Research on this scale, however, requires a high level of computational and data-science expertise, limiting the researchers who are capable of undertaking social media data-driven research to those with significant computational expertise or those who have access to such experts as part of their research team.

The Social Media Macroscope (SMM) is a science gateway with the goal of removing that limitation and making social media data, analytics, and visualization tools accessible to researchers and students of all levels of expertise. The SMM provides a single point of access to a suite of intuitive web interfaces for performing social media data collection, analysis, and visualization via for open-source and commercial tools. Within the SMM social scientists are able to process and store large datasets and collaborate with other researchers by sharing ideas, data, and methods. This document functions as a brief primer on the initial build of the SMM. As a clarifying note, the SMM is currently in a proof-of-concept stage.

The first tool in the SMM is the Social Media Intelligence & Learning Environment (SMILE) which provides open source functions that collect social media data and analyze it. The tool currently provides access to Twitter and Reddit data and can perform text-preprocessing, sentiment analysis, network analysis and machine learning text classification. Future development of the SMM will add other social media collection and analysis tools and expand the capabilities of SMILE to include more functions and algorithms.

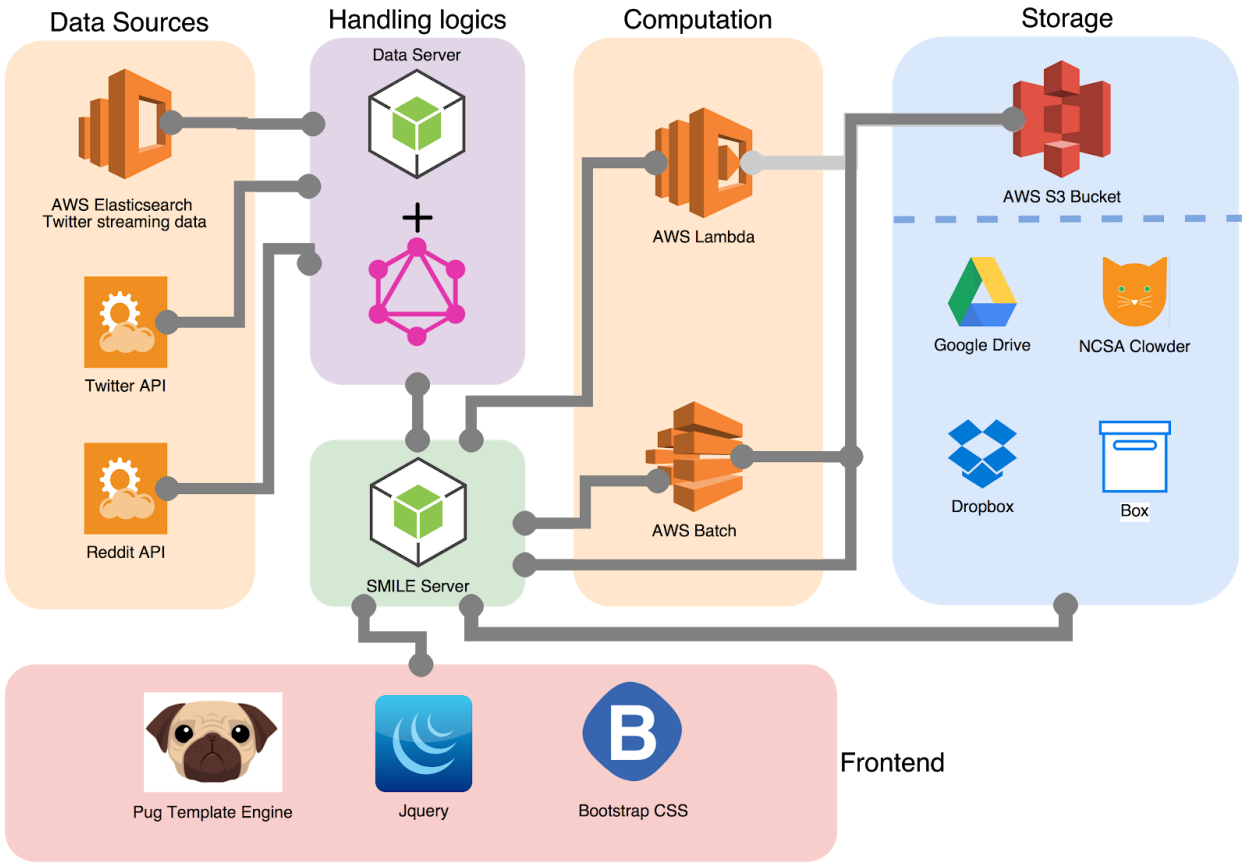
Architecture

The SMM is a web app that uses a virtual machine environment developed by HUBzero at Purdue University. The HUBzero environment allows for the launching of tools in small customized virtual machine containers. This technique allows the tailoring of the environments to fit each tool including their software dependencies and keeps the tool sessions from interfering with one another. The HUBzero environment is hosted on an Amazon Web Services (AWS) EC2 virtual machine.

The AWS environment is extensively utilized by SMILE to enable easy scaling of compute resources and componentization of its analytical methods. As is pictured in the architecture diagrams below SMILE uses the EC2 machine only for requesting its functions. All of the computation and storage is accomplished by other AWS services.

The main computation is done in AWS Lambda which is a serverless code service. Lambda houses the Twitter and Historical Reddit data collection scripts and versions of all 4 analysis methods. AWS Batch which is an on demand cluster environment houses versions of all of these scripts to be used for longer jobs as Lambda has a run limit of 5 minutes. The main Reddit search script is executed only in Batch as it has a long run time due to API rate limits.

All of the storage of social data and results is done in AWS S3 which is a simple storage environment. Each Lambda or Batch script places its results in S3 and return a link to the S3 file to the main SMM EC2 instance.



SMILE Functions

Search Twitter

SMILE offers two methods of searching Twitter for data via Twitter's API. Both methods are facilitated by using the GraphQL query language to interact with Twitter's REST API.

- *Tweets* connects to Twitter's Search API endpoint to collect and return up to 180,000 posts from the last 7 days that match the keywords you provide.
- *Twitter Users* connects to Twitter's User Search endpoint to collect and return up to 1,000 accounts that match the keywords you provide.

You can find a list operators(AND, OR, NOT, etc) you can use in the search here, URL encoding will be done automatically.

Historical Twitter

SMILE also offers the option to search our historical backlog of Twitter data. Our historical database includes tweets collected from Twitter's 1% streaming API since June 2017. There may be some small gaps in the available tweets due to collection errors.

Search Reddit
SMILE offers three methods to search for Reddit data from the Reddit API. All three methods are facilitated by using the GraphQL query language.

- *Search Reddit Posts* connects to the Reddit API and returns the 1000 most recent posts containing the keyword given.
- *Posts in Subreddit* connects to the Reddit API to return the 100 most recent posts in a specific Subreddit.
- *Comments in Subreddit* connects to the Reddit API to return the 100 most recent comments in a Subreddit.

These processes can take a significant amount of time to complete for large numbers of posts due to limitations on API calls.

Historical Reddit
SMILE two methods to search for historical Reddit data. Both methods are facilitated by using the GraphQL query language to connect to Pushift.io's Reddit API.

- *Historical Reddit Posts* connects to the Pushshift.io endpoint for Reddit Posts to collect and return up to 10,000 Reddit posts who's titles match the keywords you provide.
- *Historical Reddit Comments* connects to the Pushshift.io endpoint for Reddit Comments to collect and return Reddit comments that match the keywords you provide. This can also be used once you completed a search for Reddit posts to return all of the comments to those posts.

Both of these processes take a significant amount of time to complete for large numbers of posts due to limitations on API calls.

Text Preprocessing
SMILE currently provides text preprocessing powered by the Natural Language Toolkit (NLTK) Python Library. This processing breaks post text down into phrases and words, removes stopwords, stems words, lemmatizes words, tags parts-of-speech and provides visual analysis of the most commonly used words and connections between word usage.

Sentiment Analysis
SMILE currently provides sentiment analysis powered by the Valence Aware Dictionary and Sentiment Reasoner (VADER) algorithm. This algorithm is designed specifically for use on Twitter data and provides sentiment analysis for each tweet and the corpus of collected tweets as a whole.

Text Classification
SMILE currently provides supervised machine learning text classification powered by the Scikit-Learn Python Library. This classification takes three separate tasks to create and test you prediction model.

- *Step 1* splits your collected data into a training and testing set based on a ratio you provide and delivers the training set to you in CSV format. You will need to download that training set and note the training set identifying code that you are given. Then, label

each post in the training set if the categories you would like to train your model to recognize.

- *Step 2* requires you to upload your labeled training set and give your identifying code to match it to the saved testing set. This training set is used to create a machine learned model to identify posts that match the categories you trained.
- *Step 3* Tests your model by using it to attempt to identify which posts in the training set match categories your model is trained to identify. You can then evaluate the successfulness of your model based on this test.

Training a machine learning model can take multiple attempts through this process before you get a successful model.

Network Analysis

SMILE currently provides a tool for network analysis powered by the NetworkX Python Library to help you analyze how content moved through the network of posters in your dataset. This tool identifies mention, reply and retweet interactions between users and can help you discover the influence of specific accounts on the network as a whole.

Conclusion

As a community grows around the SMM, we envision additional tools, algorithms, and functionality to be added by members of the larger community. Technical support can be requested via the SMM, and any project-level questions can be sent to Joseph Yun (Leader, Social Media Analytics, Social Media Lab, Technology Services, University of Illinois at Urbana-Champaign) at jtyun@illinois.edu.