Elaboration I - Planning

Kylie Reynolds

FOAR705, Semester 2, 2019

Contents

1 Introduction

Part of the research that I am currently thinking of undertaking for my MRES Thesis is focused around active bystander interventions in the context of violence against women (VAW) activism online. As part of the research, I am thinking of incorporating an investigation into Twitter responses to male and female activists for the prevention of VAW, the negative responses to this support for violence prevention, and online positive bystander action in support of the allies.

2 Problem

To complete the MRES Thesis project there are several tasks which have been identified and which will need to be completed to stay on track within the project time frame. These tasks were identified in the scoping phase of this analysis (see Appendix A). I have narrowed down the scope of this list and have identified my biggest time constraints which may impede my progress, and where solutions for addressing these issues must be found. These are as follows:

- the collection of data from Twitter;
- linking connections between posts to keep track of relationships;
- the analysis of data;
- output and storage of data; and
- the accumulation of references for tweets.

3 Goals

To achieve the goal of effectively managing the Twitter data that is to be collected, analysed and reported on to complete my thesis in a timely manner, I will need to identify technological solutions for the tasks identified above.

An initial investigation will be completed to identify already tried and tested existing tools that may be adopted - which will also save time - and a list will be outlined in the section below. These tools, over the coming weeks, will be tested to see if they are appropriate for adoption in the next phase of this project. If no tools can be found, development of the necessary tools may be required,and an evaluation and reassessment of the timeline and possibilities of this project will need to be considered.

4 Identification of tools for testing

- Find out about Twitter Application Program Interface (API)
- Identify potential tool that will:
  - collect Twitter content data
  - retains the links and relationships between threads
  - output data in csv format
- find out how to cite and reference Tweets
- identify an automated referencing tool to use

4.1 Tools identified for assessment

- Twitter Premium search and API application
  https://developer.twitter.com/en/docs/tweets/search/ overview/premium
- Netlytic
  https://netlytic.org/home/
- Octoparse
  https://www.octoparse.com/
- Reaper
- How to cite a tweet in APA
  - http://www.easybib.com/guides/citation-guides/apa-format/tweet/
  - https://apastyle.apa.org/learn/faqs/cite-twitter

5 Testing stage

5.1 Twitter API

To apply for the Twitter API and Token codes to access data from the site, I needed to sign up for a Developers account. I created a new Twitter account for the project. To sign up for the account I had to fill in an application form on the Developer site, which needed to be approved by Twitter. The Developer application was approved straight away.

I discovered that I needed to first set up an App on the site as part of the verification process, and then I would be able to access the API and Token codes to be able to access data.

There were different sign-ups which allowed different types of services. I wanted free access so I chose the Premium package as this allowed access of data up to a 30-day end point. The straight search function allowed a 7-day end point for

Twitter content.

To create an App I needed to enter a website URL. I had to create a GitHub page which have me a URL address, which I subsequently used to create the App. Once the App was created I was able to access the API and Token codes, and the Twitter Premium search sandbox.

5.2 Data collection and connecting relationships

5.2.1 Twitter Premium search

I am able to do an advanced search for tweets which searches up to 30 days. However, I was unable to find information which states that I am able to download replies to a tweet which were connected to an original tweet and I am not sure that is possible.

There are many resources on the site which tell me how to search, however, I am having trouble understanding the vast amount of information there is. I could not identify where to search within the develop site. There are additional resources on the site which outlines the search parameters that I will need to look more deeply into to discover how best to keep the relationship links between tweets and replies, and how best to accomplish the collection of data to the specifications that I would like.

The information on the site states that I am able to complete 50 requests per month for no cost. I need to find some training or advice on how to use this site.

5.2.2 Octoparse

Octoparse is a web based data scraping tool which can be used to scrape data from Twitter. The program seems to use a looping system to gather the data from each tweet searched and outputs the data in a csv spreadsheet format. The website https://www.octoparse.com/ gives details about the different pricing arrangements available. For no fee I will be able to complete 10 crawls and will be able to download up to 10,000 records.

There are tutorial videos on the Octoparse website which detail the process of scraping and how to do it specifically with Twitter. This looked like an easy tutorial to follow.

I was able to download the software onto my laptop which was easy to do. I attempted to search for relevant data on the program by following the tutorial, however, I found it difficult to set up the loops to gather the specific data that I am after. I do think this tool will be good if I can get some guidance from someone who understands the programming process better than I do, and after I

have more data carpentry skills. Before I can decide whether or not this software would be suitable I would need to learn more. At this stage I was unable to download data.

5.2.3 Netlytic

Netlytic is a text content social networks analyzer which is cloud-based. See the website here: https://netlytic.org/. To use this software an API key and Token codes are needed. This also required me to sign up and to create a login. This software was easy to use and I was able to successfully download a set of data in csv format for analysis. The data collected did not keep the thread of Tweets and replies together, however, I adjusted the way that I was searching for the data to be able to capture what I needed, with some manipulation of the spreadsheet by adding columns and variable data to spreadsheet.

With this software I will be able to collect a data set of up to 1000 records and set up 5 searches for free.

I believe this may be a tool that I can use to capture twitter data, however, I do not think that there is Open Source access to the program commands needed to recreate or adapt the software to my needs exists. I will need to look into that a little more. There are a few journal articles on this particular software which I will need to download and read to find out more about it's suitability.

5.3 Output and storage of data

The output of collected data from Twitter will be in csv format. Which will be cleaned using OpenRefine and raw data will be stored on my laptop in GitHub and backed up on Onedrive.

5.4 Referencing

Endnote will be used as a reference tool for the project.

A Appendix

A.1 Decomposition

Once the literature review, guiding questions have been defined, and a methodology has been decided on, I will have a better understanding of the type of data I need to be collecting. The following tasks need to be conducted in order to successfully fulfil the goals stated above. I will need to:

1. Create some guiding principles for data collection.
2. Decide on a way to store the data that is to be collected.
3. Identify the exact information to be collected and stored.
4. Set up somewhere to store data and set up files (spreadsheet/database).
5. Decide on field names for spreadsheet and data entry purposes, and define and document exact information to be captured.
6. Set up validation rules and controlled vocabulary.
7. Create a text format metadata file.
8. Conduct a search and identify examples of Twitter content.
9. Find examples of Twitter content for analysis.
10. Entering data of original tweet.
    (a) Data entry of original Prevention of violence against women tweet content.
    (b) Collecting relevant comments of both backlash commenters and positive active bystander commenters in defence/support linked to original tweet for qualitative analysis.
    (c) Tag keywords to aid in analysis.
    (d) Collecting data about tweets for quantitative component.
    (e) Write content analysis notes for qualitative component.
11. Enter details of citation/reference details of Twitter content to be used in the study.
12. Output data in a tidy data format for analysis.
13. Store the RAW data somewhere safe and backup the data.

A.2 Pattern Recognition

In the list of tasks above there are some clear task groupings which have emerged. They include:

- Tasks 1-7 relate to the planning and initial stages of data collection and storage.
- Tasks 8 -9 relate to searching for and identifying content for the study.
- Task 10 (a-e) relates to the data entry, researcher note taking and identifying relationships.
- Task 11 capturing information to be able to properly reference and track the Twitter content used in the study.
- Tasks 12 and 13 relate to the output of the collected and stored data for analysis.

A.3 Algorithm Design

Developing the step by step instructions for solving this and similar problems:

1. In the planning stages of data collection (tasks 1-3 and 7) must be developed manually by the researcher and the details of the information in the format of a text document will be stored electronically in a repository for easy access and version control.
2. An easy to use searchable database software program could be developed with a template where fields could be entered to set up the data collection page to the specifications of the researcher developed planning information (tasks 1-3 – docs could be uploaded to database). Task 6, set up validation rules and controlled vocabulary can be included in this process.
3. To action tasks 8-9, which are related to searching for and identifying content on Twitter, I would need to identify if there is software that enables an efficient method of trawling for Twitter content using search terms identified in the data collection and planning stages of this study. If not, I will need to discover if it is possible to create such software easily.
4. Once the stage above is complete I will need to figure out if it is possible to feed data from the original tweet posts and their comments from the original posts directly into the database, or to import data into the database software discussed above.
5. Reference data must also be transferred into the database.
6. There must be capabilities for tagging content or setting up parent/child relationships between Twitter comments and their original tweets.

7. Qualitative analysis of tweets collected must be conducted by researcher, however, there must be a field for the research to make notes on the Tweet record.

8. To manage the output, the database could be set up so that a spreadsheet in CSV format could be downloaded (formatted using tidy data best practice principles).

9. For referencing details, the software could be set up so that reference details auto populate in a field on the database record of the tweet, and a tick box could be included on the record to mark it as content that will need to be referenced/cited in the Thesis report. There could be capabilities for a downloadable reference list from this program.