Spam Links

Research question

What is the best algorithm for extracting quality seed URLs from social media data?

Motivations

- Identifying seed URLs for web crawling is an extremely labor intensive process.
- Many types of social media research would benefit from the capture of the context of a post, including referenced web resources.

Methodology

1. Extract the top 50 URLs from a tweet dataset, filtering the URLs or tweets by trial algorithm.
   a. AUT is used for this step.
   b. Also, removed all twitter.com URLs.
2. Single human codes each URL as relevant, not relevant, or indeterminate.
3. Counted each URL type for trial algorithm.

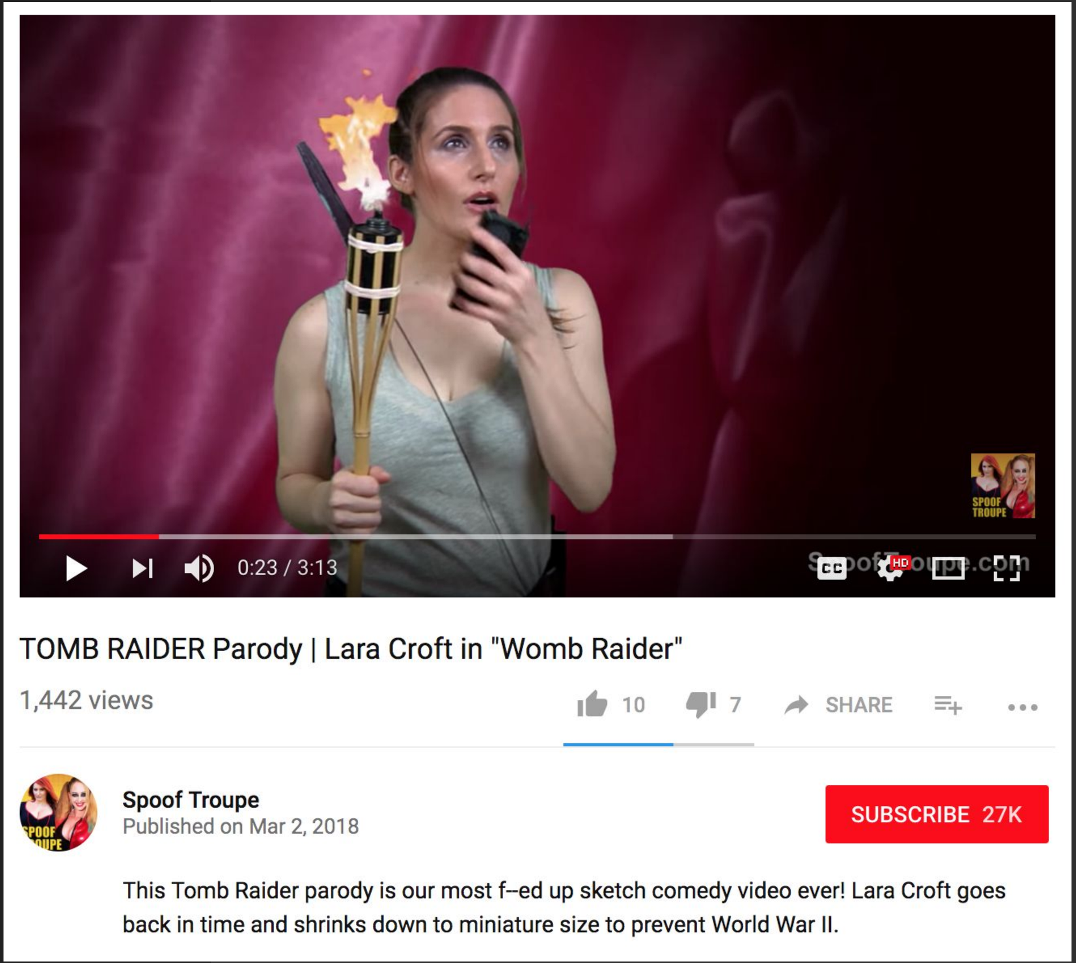
Dataset

10,712,780 tweets related to the Parkland, FL School Shooting (2018/02/14) collected 2018/03/02 - 2018/04/26

Hashtags: #NeverAgain, #neveragainmovement, #neveragainMSD, #boycottNRA, #StudentsStandup, #parkland, #parklandstrong, #DouglasStrong, #MSDstrong, #douglashigh, #douglasshooting, #MSD, #floridahighschoolshooting, #NRA, #BanTheNRA, #MarchForOurLives, #NationalWalkoutDay, #iWillMarch

Usernames: davidhoggs111, DLoesch, AMarch4OurLives

This is spam ...



TOMB RAIDER Parody | Lara Croft in "Womb Raider"

1,442 views

Spoof Troupe
Published on Mar 2, 2018

This Tomb Raider parody is our most f--ed up sketch comedy video ever! Lara Croft goes back in time and shrinks down to miniature size to prevent World War II.

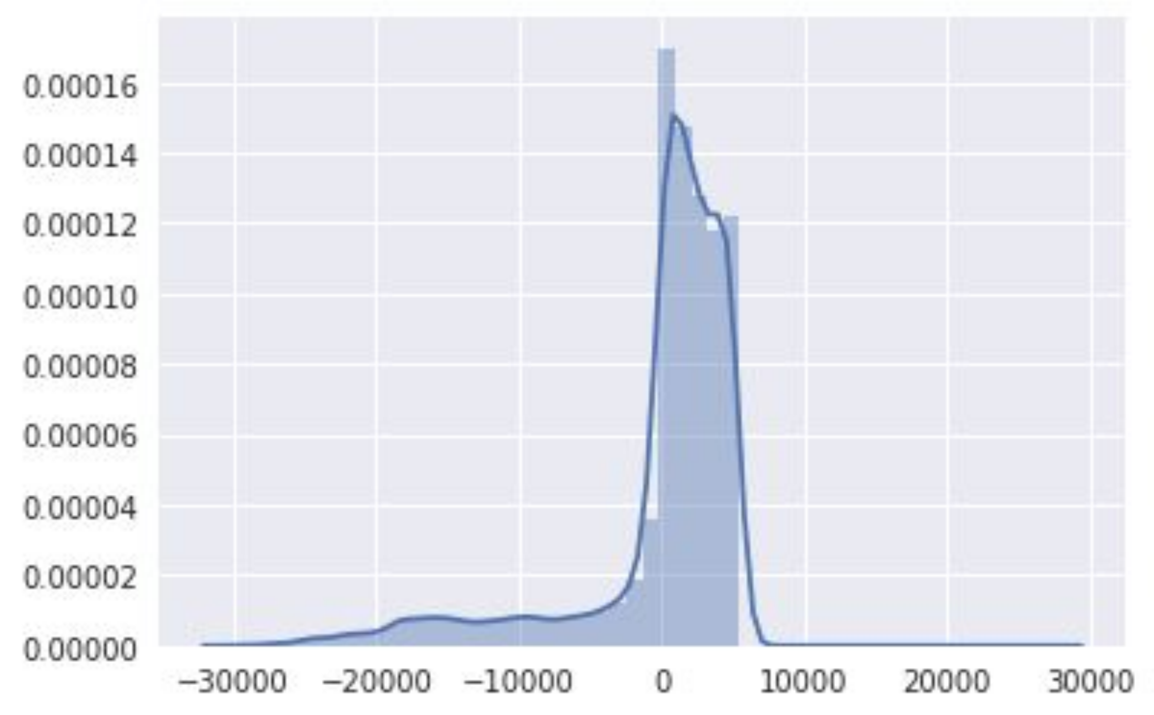
Null hypothesis

Top URLs by count of occurrences in the dataset produces the highest quality URLs.

Trial algorithms [1]

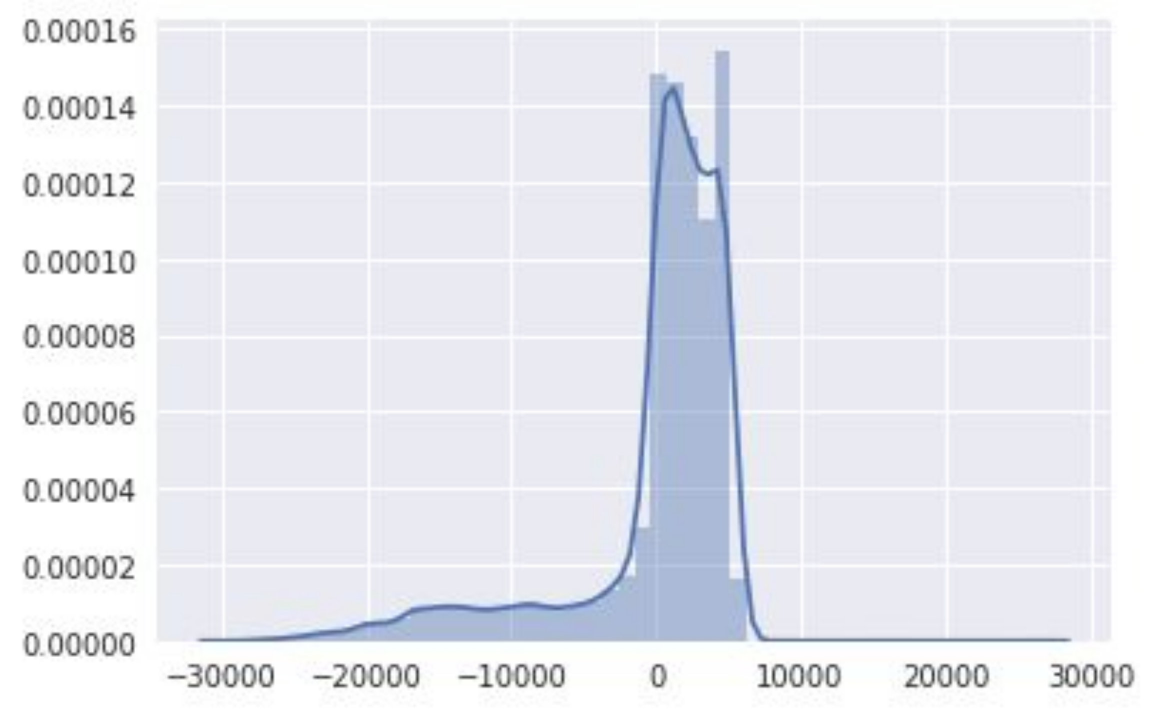
- Remove users with a default profile image.
- Remove users with no description in profile.
- Only include users in middle 50% of follower count (> 75 and < 751).
- Only include users in middle 50% of friends count (> 154 and < 836).

Change in URL Position with Follower Filter (full list)



1,047,051 → 32,209 URLs

Change in URL Position with Following Filter (full list)



1,047,051 → 30,114 URLs

Trial algorithms [2]

- Original tweets only (not retweets, quotes, or replies).
- Only include users in middle 50% of age of account (> 3/16/2011 and < 1/31/2016).

Findings

- There was not a significant number of spam links in top 50.
    - For null hypothesis, only 3 spam links.
- For original tweets only trial algorithm, no spam links.
- For all other trial algorithms, no significant effect on spam links.

Limitations

- Single dataset
- Single coder
- Only code top 50
- Limited algorithms tested
- URL normalization not handled
- URL unshortening not handled

😱 Hacking Lead to More Questions… 😱

- At what point do we see URL churn with this criteria?

Example Scala code [1]

```
import io.archivesunleashed.spark.matchbox._
import io.archivesunleashed.spark.matchbox.TweetUtils._
import io.archivesunleashed.spark.rdd.RecordRDD._
import org.json4s._
implicit lazy val formats = org.json4s.DefaultFormats

// Load tweets from Never Again dataset
val allTweets =
RecordLoader.loadTweets("/mnt/data/neveragain/*.json", sc)

// Count number of tweets in the dataset = roughly 12,000,000
allTweets.count()
```

Example Scala code [2]

```
// Creating a new set of tweets filtered by followers_count
val tweetsByFollowerCount = allTweets.filter(tweet => (tweet \ "user"
\ "followers_count").extract[Int] > 75 && (tweet \ "user" \
"followers_count").extract[Int] < 751)

// Creating array of the expanded_urls from tweets, with filtering to
remove twitter.com URLs
val followerCountUrls = tweetsByFollowerCount.map(tweet => try {
(tweet \ "entities" \ "urls" \ "expanded_url")  } catch { case e:
Exception => null}).filter(u => u.isInstanceOf[JString] ).map(u =>
u.extract[String]).filter(u => ! u.startsWith("https://twitter.com"))
```

Example Scala code [3]

```
// Counting frequency of URLs and sorting in descending order
val topFollowerCountUrls = followerCountUrls.countItems().sortBy(_._2
, false)

// Save output as CSV file
topFollowerCountUrls.coalesce(1,true).map(a => a._1 + "," +
a._2).saveAsTextFile("followerCountUrls")
```

The Team

- Brian Griffin, Old Dominion University
- Jayanthy Chengan, University of Toronto
- Justin Littman, GWU Libraries
- Russell White, Library & Archives Canada
- Shawn Walker, Under the Blazing Hot Sun