STAT 231: Assignment Dataset Information

For your assignments you will analyze a dataset of tweets downloaded from Twitter (`twitter.com`). Note that neither a Twitter account nor access to Twitter is expected or required. If you are unfamiliar with Twitter please see the section titled Twitter: A Brief Introduction on Page 4 of this document. You may also find it helpful to read about the service online.

Data Source

We have prepared a 'primary' dataset from which you will download an individual sample. The primary dataset was created using the rtweet package in R. (While you do not need to use the rtweet package in STAT 231, you may find it interesting to read more about it here: `cran.r-project.org/web/packages/rtweet/index.html`.) The package accesses the Twitter application programming interface (API) to download information relating to tweets.

We downloaded tweets from a list of 30 Twitter accounts on Wednesday, September 1, 2021. Of these, 20 are 'personal' accounts (representing an individual) and ten are 'organization' accounts. Tweets that were retweets of other accounts, and tweets that were replies to Twitter users, were excluded. Tweets that were published on September 1, 2021 were also excluded. The resulting dataset contains a total of 19,141 tweets. A summary of the 30 accounts we sampled are summarized on Page 2.

Downloading Your Sample

In your assignments you will analyze a sample of tweets sampled from the primary dataset. To construct your sample, you will need to choose five accounts from the list of 30 in the primary dataset. Three of these must be personal accounts and two must be organizational accounts. You will use the same sample for all five assignments, so choose carefully! The Midterm and Final Reports will analyze a different set of tweets. Further details will be announced later in the term.

When choosing the accounts you will analyze, you may wish to look at the list of variates that are provided and think about which accounts may be interesting to compare. For example, would you expect certain accounts to get more/fewer retweets, or write tweets with longer words? Perhaps you'd expect some accounts to tweet at different times of day, or on different days of the week. There is no 'good' or 'bad' set of accounts to choose, but if you think about some of the questions we might want to ask (and answer) based on the available variates, you may find the analyses we do throughout the term more interesting!

Please do not deliberately choose a set of accounts to match those chosen by a classmate. This will not make the assignments any easier (even if you choose the same accounts you will still get a different sample of the data), and you will miss the opportunity to study the data that most interest you. It may, however, be helpful to discuss with classmates when deciding which accounts to select - a friend may think of something you hadn't considered, and vice-versa! As you'll know from what you learned in STAT 230, there are $\binom{20}{3} \times \binom{10}{2}$ possible sets of accounts that can be chosen. We will therefore be very surprised if any two students make exactly the same selection.

To download your sample go to `shiny.math.uwaterloo.ca/sas/stat231/tweetdownloader/` and follow the instructions. The password is "gimmethedata" (not including the "quotation" marks, all lower case). Note that this website includes some summary information about the primary dataset: the number of followers each account had at the time the data were downloaded, the number of tweets that appear in the primary dataset for each account, and the date of the first tweet that appears in the primary dataset for each account.

The file you download will include a random sample of tweets from each of the five accounts you have chosen. As soon as you download your data please check it immediately for any obvious errors. If you have any trouble downloading your sample please ask for help on Piazza as soon as possible!

Once you are happy with your sample dataset, upload the file to the dropbox on LEARN.

The Accounts

We constructed a list of 20 personal accounts and 10 organization accounts. The accounts are largely based in the Eastern time zone to minimize geographical dissimilarities. The tables below briefly summarize each account (note that elected officials may change roles over time). You are encouraged to do additional research into these accounts to learn more about who they represent!

Personal Accounts (Choose 3)

Username	Name	Description
@alessiacara	Alessia Cara	Singer-songwriter.
@amanda_parris	Amanda Parris	Broadcaster and writer.
@b0rk	Julia Evans	Computer scientist and author.
@BardishKW	Bardish Chagger	Member of Parliament for Waterloo.
@benhuot	Benoît Huot	Paralympic swimmer.
@christibelcourt	Christi Belcourt	Visual artist and author.
@CPHO_Canada	Theresa Tam	Chief Public Health Officer of Canada.
@DesmondCole	Desmond Cole	Journalist, activist, author, and broadcaster.
@farwell_WR	Mike Farwell	Waterloo region local radio talk show host.
@JohnTory	John Tory	Mayor of Toronto.
@jonnysun	Jonny Sun	Author and illustrator.
@kevinolearytv	Kevin O'Leary	Businessman and television personality.
@MargaretAtwood	Margaret Atwood	Poet and novelist.
@NaheedD	Naheed Dosani	Physician and health justice advocate.
@NahlahAyed	Nahlah Ayed	Journalist and broadcaster.
@SimuLiu	Simu Liu	Actor and writer.
@solmamakwa	Sol Mamakwa	Member of Ontario Provincial Parliament.
@tagaq	Tanya Tagaq	Singer.
@thehazelmae	Hazel Mae	Sports broadcaster.
@theJagmeetSingh	Jagmeet Singh	Leader of the New Democratic Party.

Organization Accounts (Choose 2)

Username	Description
@AnishNation	Anishinabek First Nations political advocate group.
@GKWCC	Greater Kitchener Waterloo Chamber of Commerce.
@OHLRangers	Kitchener Rangers hockey team.
@ONThealth	Ontario Ministry of Health.
@OntScienceCtr	Ontario Science Centre museum.
@ROMtoronto	Royal Ontario Museum, Toronto.
@ROWPublicHealth	Region of Waterloo Public Health.
@TheTorontoZoo	Toronto Zoo.
@TorontoDefiant	Toronto Defiant, Overwatch esports team.
@TorontoStar	Toronto Star, newspaper.

Notes: Our primary aim when compiling this list was to offer a varied set of personalities and organizations for you to analyze, while also ensuring the statistical methods of the course could be adequately applied to the data. **The inclusion of an account on this list does not indicate they are endorsed by the instructors or the university in any way.** We have not read all of the tweets in the primary dataset and make no guarantees as to their content. In particular, we cannot guarantee that the primary dataset does not contain tweets that may offend some individuals. If you are concerned about being exposed to material that you may find offensive (or in any other way distressing) we recommend you take care in your choice of accounts. Note, however, that no photos or videos will be downloaded, only the text of the tweets by that user. If you have any concerns please contact Professor Wallace, who will be happy to assist you in selecting a sample.

Variates

Your downloaded dataset will be a comma separated value file (.csv). Each row corresponds to an individual tweet. Each column corresponds to a variate. The variates are:

- username: the username (or 'handle') of the account that sent the tweet (in other words, the name of the 'user').
- profile: a brief description the user provides in the profile of their account.
- followers: the number of accounts following the user at the time the data were downloaded.
- following: the number of accounts the user is following at the time the data were downloaded.
- published: the date and time the tweet was published on Twitter. Note that the time is adjusted to the Eastern time zone but this may not have been when the tweet was posted in the user's local time (for example, if someone tweeted while travelling).
- day.of.week: the day of the week the tweet was published (Monday-Sunday).
- time.of.day: the time of day the tweet was published on Twitter, expressed as seconds after midnight.
- tweet.gap: the number of seconds since the user last tweeted.
- first.tweet: a binary indicator of whether this was the first tweet the user sent that day, taking the value 1 if it is the first tweet of the day, and 0 otherwise.
- text: the text contained in the tweet. Note that special characters (such as emoji) will not be properly rendered.
- length: the length of the tweet text in number of characters, as reported by the Twitter API.
- words: the number of 'words' in the tweet (calculated by counting character string separated by a space). Text such as web links, hashtags, and mentions of other accounts count as words.
- long.words: the number of words 10 characters or longer in the tweet, *excluding* urls, hashtags, and mentions of other accounts. Note that the length of a word includes punctuation.
- likes: the number of users who have liked the tweet.
- retweets: the number of users who have retweeted a tweet.
- source: the software or platform used to send the tweet, such as a mobile phone app, web browser, or specialist software.
- media: the number of media items (images or videos) used in the tweet. Note that a tweet can contain at most four images or one video, and cannot contain both images and videos.
- urls: the number of urls (website links) mentioned in the tweet.
- hashtags: the number of hashtages used in the tweet.
- mentions: the number of other twitter users mentioned in the tweet.
- media.binary, urls.binary, hashtags.binary, mentions.binary: binary indicators of whether or not the tweet features at least one media item, url, hashtag, or mention.
- quote: a binary indicator of whether the tweet was 'quoting' another tweet.
- quote.verified: if the tweet is quoting another tweet, this indicates whether the quoted user is a 'verified' account or not.
- quote.likes: if the tweet is quoting another tweet, this indicates the number of users who have liked the tweet being quoted.
- quote.retweets: if the tweet is quoting another tweet, this indicates the number of users who have retweeted the tweet being quoted.

Twitter: A Brief Introduction

Twitter is a social networking service which allows users to post short messages (known as tweets). Tweets are limited to 280 characters in length. Users may also add photos or videos to their tweets. Photos or videos may be posted with or without accompanying text. Tweets will often feature hashtags, which are special words or phrases that are used to help users search for tweets on a certain topic. Tweets may also mention other accounts. Mentioning another account will automatically send a notification to that user, alerting them to the fact they have been mentioned in the tweet.

Users may follow other accounts. Following an account means that tweets in that account will automatically be shown to you in your twitter feed. The number of followers an account has is the number of users who have chosen to follow that account. This may be thought of as an indicator of how popular that account is. The number of accounts a user follows is also reported (and denoted following).

Twitter accounts may be verified. This means that Twitter has confirmed the account truly belongs to the individual it claims to represent. Examples of accounts that will typically be verified include those used by celebrities, politicians, and companies.

Users may interact with tweets in a number of ways. These include:

- Like: this adds the tweet to a list of liked tweets for that user. The action of liking a tweet may be thought of as an endorsement of that tweets content. The term favorite is also sometimes used to describe this action.
- Retweet: this shares the tweet with all the users that follow the account who retweeted it. A retweet may be thought of as a stronger endorsement of the tweet than a like.
- Quote: a user may quote a tweet which allows them to add their own comment about the original tweet, while also sharing the original tweet with that users followers. Quote retweets are often not endorsements of the original tweet, and may be used to express disagreement.

While liking or retweeting a tweet is not always an indication that a user endorses the tweet, you may assume this is true for the purposes of analyzing the Twitter data in this course. Many users consider it desirable to post tweets that receive a large number of likes, retweets, and quotes.

If you are ever uncertain of terminology specific to the Twitter service, or have any questions about how people use Twitter, then please ask for help on Piazza!