Mahtab EzzatiKarami

Answers:

3.2) How could peaks of in_reply_to_user_id be helpful to a company examining a dataset of tweets?

Identify Related High Reply Recipients and Engage with Them: Top results show the users who have the highest number of in_reply_to_user_id field. It means these people get a lot of replies so the interaction from people for their tweets are high. Companies can connect with related high reply recipient users get their business in front of the right audience.

When companies engage with these people who in some cases can be called influencers, they will create the opportunity to gain more followers, more engagement and more traffic to their website. in your field, they already have the audience you want!

p.s. By related, I mean for example Donald Trump is the one whose tweets get the most number of replies, but a sporting clothes company won't necessarily want to advertise their stuff in Trump's account. But they can approach a football player who probably also gets a lot of interaction from people and he will be related to this company's business.

Ad Targeting: Like what I mentioned above, if there is a user related to a company's business, they can target the people replying to their tweets to find the right audience for their products and advertisement. For example, there are a lot of people in sports filed. A company can not target all the people who are interacting with these sports guys. Instead they can target the high reply recipients and target their audience.

4.3) Speculate on the source of the less-common entries. How could entries like this disrupt an optimized workflow (completely automated)?

Generally speaking, less-common entries can refer to bad quality data which can have various sources. Different causes can be:

- Missing Data: Empty fields that should contain data.
- Wrong or inaccurate data: Information that has not been entered correctly or maintained.
- Inappropriate data: Data that's been entered in the wrong field.
- Non-conforming data: Data that hasn't been normalized as per the system of records.
- Duplicate data: A single Account, Contact, Lead, etc. that occupies more than one record in the database.
- Poor data entry: Misspells, typos, transpositions, and variations in spelling, naming or formatting.

Mahtab EzzatiKarami

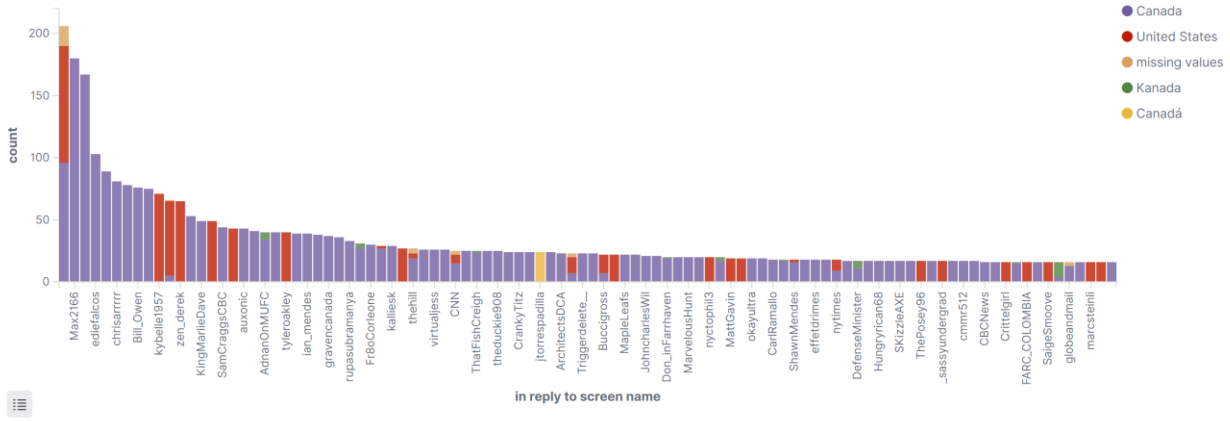
Dirty data wreaks havoc on the entire revenue cycle of an organization. The impacts can range from a transaction level loss to catastrophic effect for an enterprise. The impact of bad data can be:

- Higher consumption of resources
- Higher maintenance costs
- Errors in product/mail deliveries
- Lower customer satisfaction and retention
- Increased churn rate
- Distorting campaign success metrics
- Failure of your marketing automation initiatives
- Dissatisfied sales and distribution channels
- Higher spam counts and un-subscriptions
- Negative publicity on social media
- Misinformed OR under-informed decisions
- Invalid reports
- Lower productivity
- Loss in Revenue

For this specific dataset, as we can see in the 4.1 image, the less common places are Kanada and Canadá. Both of these entries show Canada, but because the dictation is slightly different, they are grouped as a separate place. Here, the number of reply tweets which belong to these two places are not high but if this kind of difference happens in larger scale, it can bias the results. If someone like a data analysist is checking the results, they can handle this difference, but if the process is completely automated it will affect the results.

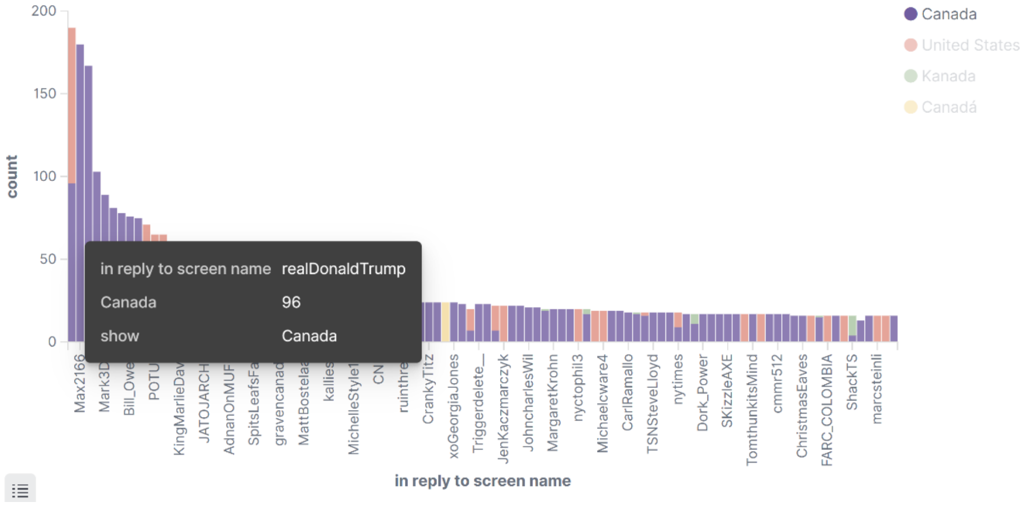
4.4) Are there any missing values for the place.country.keyword field among these users?

As we can see below, there are missing values.



4.5) How many tweets to the most-replied screen name came from "Canada"?

96


count
in reply to screen name realDonaldTrump
Canada 96
show Canada
Max2166
Mark3D
Bill_Owe
POTU
KingMarlieDav
JATOJARCH
AdnanOnMUF
SpitsLeafsFa
gravencanad
MattBostelaa
kallies
MichelleStyle1
CN
ruinthre
CrankyTitz
xoGeorgiaJones
Triggerdelete__
JenKaczmarczyk
JohncharlesWil
MargaretKrohn
nyctophil3
Michaelcware4
CarlRamallo
TSNSteveLloyd
nytimes
Dork_Power
SKizzleAXE
TomthunkitsMind
cmmr512
ChristmasEaves
FARC_COLOMBIA
ShackTS
marcsteinli
in reply to screen name
● Canada
● United States
● Kanada
● Canadá