

Questions and Notes

Victor & Mikkel

October 24, 2020

1 Questions

- `train_data_checkpoint.json`:
Trained on wikipedia and articles for what purpose? To understand "security", "privacy", etc.? When is this used?
- Commercial intentions?:
Is the intention to use this commercially? The API specifically states that this should not be used commercially. If commercial aspects are of interest, we need to have it approved.

2 Notes

- Reddit API:
client ID: `fQuRCESlv_7yOOQ`
client secret: `v6AjzytAKDgqjqyQNczapezXUjk`

3 Overview

- `reddit_API.py`:
Uses the Reddit API to gather top tweets from the "smarthome" subreddit and comments. Now set up to use Victor's account.
Questions:
Why does it not save them? How many tweets can/should we get?
- `scrape.py`:
Generates a user-agent to access the different articles used for generating the *train_data.json* file.
Questions:
Why are there so many functions for scraping? Is it because they differ in format and need different preprocessing to read?
- `pushshift_aggs.py`:
Used to scrape the frequency of query-terms "*privacy*", "*security*", "*trust*"

for the last 3 years on the subreddits. It seems like the *frequency* keyword establishes the sampling rate. This is used to plot how many comments uses these query words over 3 years. *Questions:*
Is this actually the frequency?