# Overview

# Collecting removed Reddit comments

The three-stage process to collect Reddit comments that were removed by moderators. An illustration of this approach is shown in figure below.



* Used **comment\_streaming\_scripts** for **stage 1**.
* Used **collection\_and\_reporting\_script** for **stage 2 and 3**.

Reference paper: <https://arxiv.org/pdf/1904.03596.pdf>

1. **comment\_streaming\_scripts**: These scripts are currently running in the background on the effect server and continuously streaming live **comments** (as and when they are being posted on Reddit).
2. **submission\_streaming\_scripts**: These scripts are currently running in the background on the effect server and continuously streaming live **posts**/submissions (as and when they are being posted on Reddit).
3. **collection\_and\_reporting\_script.ipynb**: This notebook collects comments through ids using the Reddit API wrapper (praw). It also can be used to report the statistics on the collected data.
4. **discussion\_thread\_script.ipynb**: This notebook generated a complete discussion thread starting from the removed comment to the root (parent post). This is a bottom-up approach, i.e. for every removed comment the script finds the immediate parent then >> grandparent then >> great-grandparent and so on till the root (parent post).
5. **removed\_com\_toxicity\_script.ipynb**: This notebook finds the toxicity of each removed comment using Google’s Perspective API.