Team: A8

|  |  |  |  |
| --- | --- | --- | --- |
| Subreddit | Reason | Number of subscribed users | Amount of data collected |
| r/adidas | To understand the customers preferences | 25.9k | 328,022kb |
| r/beyonce | Determine active user’s interest in adidas/sports | 72.3k | 206kb |
| r/PharrellWilliams | Determine active user’s interest in adidas/sports | 798 |  |
| r/BadBunnyPR | Determine active user’s interest in adidas/sports | 2.6k |  |
| r/BlackPink | Determine active user’s interest in adidas/sports | 105k | 66k |
| r/naeun | Determine active user’s interest in adidas/sports | 681 |  |
| r/YaraShahidi | Determine active user’s interest in adidas/sports | 550 |  |
| r/ZoeSaldana | Determine active user’s interest in adidas/sports | 8.1k | 82kb |
| r/AdrieneMishler | Determine active user’s interest in adidas/sports | 219 |  |
| r/karliekloss | Determine active user’s interest in adidas/sports | 10.1k | 102kb |
| r/NinjasHyper | Determine active user’s interest in adidas/sports | 4.3k |  |
| r/kerwinfrost | Determine active user’s interest in adidas/sports | 118 |  |

Data description:

As we are using multiple subreddits we are extracting large amounts of data both streaming and historical. Some of these subreddits are not active and some posts in the data were in different languages other than english, hence chosen to ignore the irrelevant data or rely more on historical data for text mining purposes. Also we encountered many pictures and videos based posts which also were ignored for the time being. We also look forward to studying the data before and post the covid pandemic breakout given the special situation in the year of 2020. Overall the data extracted for few celebrities might not be sufficient, but we are focusing on celebrities with sufficient and non-noisy data.This is still a work in process.