Successfully downloaded: 16.8B

Viable samples: 13.1B

Download attempted: 40.0B

CommonCrawl samples: 88.0B

Near-duplicate: 0.5B, NSFW image: 2.8B, NSFW text: 0.8B, Total: 3.7B

Dead links and other download errors: 23.2B

Download not attempted: 48.0B