Detailed Dataset Description

Overview

The May 2015 Reddit Comments dataset is a comprehensive collection of Reddit comments posted during the month of May 2015. This is one of Kaggle's most popular text datasets and has been widely used for natural language processing, sentiment analysis, and machine learning projects.

Dataset Format and Size

- **File Format**: SQLite database (database.sqlite)
- **Size**: Approximately 30+ GB (compressed ~20 GB)
- **Number of Records**: Over 54 million comments
- **Time Period**: May 1-31, 2015

DatasetLink: https://www.kaggle.com/datasets/kaggle/reddit-comments-may-2015

Data Structure

The dataset is stored in a SQLite database within a primary table (named "May2015") containing the following fields:

1. **created_utc**: Unix timestamp of when the comment was created
2. **ups**: Number of upvotes the comment received
3. **subreddit_id**: Unique identifier for the subreddit
4. **link_id**: ID of the submission (post) the comment belongs to
5. **name**: Unique identifier for the comment (format: t1_xxxxx)
6. **score_hidden**: Boolean indicating if the score was hidden
7. **author_flair_css_class**: CSS class for author's flair
8. **author_flair_text**: Text content of author's flair
9. **subreddit**: Name of the subreddit where the comment was posted
10. **id**: Comment ID (without the t1_ prefix)
11. **removal_reason**: Reason for removal (if applicable)
12. **gilded**: Number of times the comment was gilded (given Reddit Gold)
13. **downs**: Number of downvotes (deprecated field, usually 0)
14. **archived**: Boolean indicating if the comment is archived
15. **author**: Username of the comment author
16. **score**: Net score (upvotes minus downvotes)
17. **retrieved_on**: Unix timestamp of when the data was retrieved
18. **body**: The actual text content of the comment
19. **distinguished**: Whether the comment was distinguished (mod, admin, etc.)
20. **edited**: Timestamp if edited, or False if never edited
21. **controversiality**: Score indicating how controversial the comment is
22. **parent_id**: ID of the parent comment or submission

Key Characteristics

Content Coverage:

- Spans thousands of different subreddits
- Includes comments from major subreddits like AskReddit, funny, pics, worldnews, gaming, etc.
- Contains diverse topics ranging from casual conversations to serious discussions

Data Quality:

- Raw, unfiltered comments (includes deleted/removed comments marked as "[deleted]" or "[removed]")
- Contains various languages, though predominantly English
- Includes special characters, emojis, markdown formatting, and URLs
- May contain offensive or inappropriate content as it's unmoderated data

Comment Characteristics:

- Wide range of comment lengths (from single words to lengthy essays)
- Nested conversation threads through parent-child relationships
- Contains metadata useful for engagement analysis (scores, gilding, controversiality)

Use Cases

This dataset is commonly used for:

- Natural Language Processing (NLP) research
- Sentiment analysis
- Text classification and categorization
- Word embeddings training (Word2Vec, GloVe, etc.)

Considerations

- **Size**: The dataset is very large and requires significant computational resources
- **Privacy**: Usernames are included, though this is public Reddit data
- **Content**: May contain offensive, controversial, or sensitive material
- **Temporal Snapshot**: Represents Reddit's community and culture from 2015
- **Data Processing**: Typically requires SQL queries or conversion to other formats for analysis

This dataset provides an excellent snapshot of online discourse and social media interaction from a specific time period, making it valuable for both academic research and practical machine learning applications.

Note: For implementation details, normalization schema, and step-by-step data loading process, please also check out the attached **README.md** file.