Reddit User Persona Generator

A Python script that scrapes Reddit user profiles and generates detailed user personas using AI analysis of their posts and comments.

Features

- **Profile Scraping**: Extracts posts and comments from Reddit user profiles using the public JSON API
- **Persona Generation**: Creates detailed user personas with 10 key characteristics
- **Citation System**: Provides citations for each persona characteristic with source links
- **Comprehensive Analysis**: Analyzes basic info, interests, personality traits, communication style, values, tech usage, social behavior, goals, challenges, and lifestyle
- **Text Output**: Saves results to formatted text files

Requirements

- Python 3.7+
- Internet connection
- Required Python packages (see requirements.txt)

Installation

1. Clone or download this repository
2. Install required packages:

```bash
pip install -r requirements.txt
```

Usage

Method 1: Command Line with URL

```bash
python reddit_persona_generator.py "https://www.reddit.com/user/kojied/"
```

Method 2: Interactive Mode

```bash
python reddit_persona_generator.py
```

Then enter the profile URL when prompted.

Method 3: With Optional Parameters

```bash
python reddit_persona_generator.py "https://www.reddit.com/user/Hungry-Move-6603/" --limit 150 --output my_outp
```

Command Line Options

- `profile_url`: Reddit profile URL (required if not running interactively)
- `--limit`: Maximum number of posts/comments to analyze (default: 100)
- `--output`: Output directory for persona files (default: output)

Sample URLs

The script works with standard Reddit profile URLs:

- https://www.reddit.com/user/kojied/
- https://www.reddit.com/user/Hungry-Move-6603/

Output

The script generates a text file named `{username}_persona.txt` containing:

1. **Basic Information**: Age range, gender, location, occupation
2. **Interests and Hobbies**: Extracted from subreddit participation and content
3. **Personality Traits**: Derived from writing patterns and content analysis
4. **Communication Style**: Formal/informal, detailed/concise analysis
5. **Values and Beliefs**: Identified from content themes
6. **Technology Usage**: Tech-savvy level assessment
7. **Social Behavior**: Interaction patterns and community engagement
8. **Goals and Aspirations**: Life goals and ambitions
9. **Challenges and Pain Points**: Identified struggles and concerns
10. **Lifestyle**: Activity level and lifestyle patterns

Each characteristic includes citations with:

- Source post/comment links
- Context snippets
- Supporting evidence

Example Output Structure

```
USER PERSONA: kojied
==================================================

Analysis Date: 2025-07-15T10:30:00
Total Posts Analyzed: 85

BASIC INFORMATION:
------------------------------
  Age Range: 25-40
  Gender: Unknown
  Location: Unknown
  Occupation: Professional/Working

INTERESTS AND HOBBIES:
------------------------------
  • programming
  • technology
  • gaming
  • movies

[... more sections ...]

CITATIONS:
==================================================

Basic Information:
------------------------------
  Characteristic: Age Range
  Value: 25-40
  Source: https://www.reddit.com/r/programming/comments/xyz123/
  Context: "In my career as a software developer..."

[... more citations ...]
```

How It Works

1. **URL Parsing**: Extracts username from Reddit profile URL
2. **Data Scraping**: Uses Reddit's public JSON API to fetch posts and comments
3. **Content Analysis**: Analyzes text patterns, subreddit participation, and content themes
4. **Persona Generation**: Uses rule-based analysis to extract persona characteristics
5. **Citation Generation**: Links each characteristic to supporting evidence
6. **File Output**: Saves formatted persona to text file

Rate Limiting

The script includes built-in rate limiting (2-second delays between requests) to respect Reddit's servers and avoid being blocked.

Privacy and Ethics

- Only accesses publicly available Reddit data
- No authentication required
- Respects Reddit's public API terms
- Generated personas are for research/analysis purposes only

Error Handling

The script handles common issues:

- Invalid URLs
- Private/empty profiles
- Network connectivity issues
- Rate limiting
- Malformed data

Limitations

- Only analyzes publicly available posts and comments
- Cannot access private or deleted content
- Analysis quality depends on available data volume
- Rule-based analysis may miss nuanced characteristics
- No real-time data (analyzes historical posts)

Technical Notes

- Uses Reddit's public JSON endpoints (no API key required)
- Implements respectful scraping practices
- Follows PEP-8 coding standards
- Includes comprehensive error handling and logging
- Modular design for easy extension

Troubleshooting

"No posts found": The profile might be private, empty, or the username incorrect.

"Invalid URL": Ensure the URL follows the format: https://www.reddit.com/user/username/

Rate limiting errors: The script includes delays, but if you encounter issues, try reducing the --limit parameter.

License