DOCUMENTATION FOR REDDIT SCRAPER SCRIPT (reddit_scraper.py)

Purpose

This Python script automatically collects textual data from Reddit for analysis, focusing on the /r/Ljubljana subreddit. It retrieves recent and popular posts, along with their metadata, filters them to include only Slovenian-language content that contains text beyond just titles, and exports the results to a CSV file in a form of a corpus. The resulting dataset can be used for corpus linguistics, sociolinguistic analysis, or natural language processing (NLP) tasks involving informal Slovene text from social media.

Imported Modules

- praw – Provides a Python interface for accessing Reddit's API. It retrieves posts and their attributes (title, author, score, etc.) programmatically.
- pandas – Organizes the collected data into a structured DataFrame and handles CSV export.
- datetime – Generates timestamps for file naming (ensures unique filenames).
- langdetect – Detects the language of each post to include only Slovenian content.
- os – Manages file paths and saves results to the appropriate directory.

Core Function is_slovenian(text)

Function is_slovenian(text) determines whether the provided text is in Slovenian. It uses langdetect to identify language and returns True only if the detected code is "sl". If detection fails, returns False to maintain uninterrupted execution. The input is string (Reddit post content) and the output is boolean value (True if Slovenian, False otherwise).

Reddit Connection and Data Retrieval

The script connects to Reddit using PRAW with valid API credentials (client_id, client_secret, user_agent). Each user must generate personal Reddit API credentials by creating an application on their Reddit account and replace the placeholder values for client_id, client_secret, and user_agent in the script with their own. These credentials are required for Reddit API access. It targets the r/Ljubljana subreddit and retrieves posts from three categories:

- subreddit.hot(limit=500)
- subreddit.new(limit=500)
- subreddit.top("all", limit=500)

All posts are combined, and duplicates are removed based on post IDs, yielding a unique collection before filtering.

Filtering Logic

The script applies several filters to retain only relevant textual Slovenian content:

1. Language filter – Only posts whose combined title and selftext are detected as Slovenian are included.

2. Content filter – Posts without any textual selftext are skipped.
3. Media exclusion – Posts marked as images, videos, or links (post_hint in ["image", "video", "link"]) are discarded.
4. Duplicate removal – Ensures each post appears only once in the dataset.

Data Storage and Output

For each post that passes all filters, the following fields are extracted: title (post title), score (upvote/ downvote score), id (unique Reddit ID), url (direct post link), num_comments (number of comments), created_utc (UTC timestamp of creation) and selftext (main post body text).

The data are stored in a Pandas DataFrame and written to a UTF-8 (with BOM) CSV file on the Desktop. The filename includes a timestamp, e.g.: Ljubljana_slovenian_posts_20251101_123806.csv. The choosen encoding preserves Slovene diacritics (č, š, ž) and ensures compatibility with Excel and corpus tools.

Example CSV Columns

id	title	selftext	score	url	num_comments	created_utc
t3_xabcd	Ljubljanski promet	Danes spet zastoji na Celovški.	52	https://redd.it/xabcd	13	1730462100

Runtime and Performance

The script is executed from the command line: python reddit_scraper.py

Execution time depends on Reddit API response and network conditions (typically 1–3 minutes). Progress messages appear in the terminal, and upon completion, the script reports how many Slovenian posts were saved in the CSV file.

Cleanup and End of Script

After all posts are processed, the script automatically closes any open connections to Reddit and prints a completion message. The resulting CSV file constitutes a clean, text-only corpus of Slovenian Reddit discourse, ready for further computational or linguistic analysis.