DSCI 560 - Data Science Professional Practicum Lab 4 Part 2

Team name: Cleanup Crew
Team members: Jonathan Kusuma (ID:4897950942), James Susanto (ID:3466104463)

Introduction

In the continuation of our project, we have enhanced the functionality of our Reddit data scraping pipeline. Lab 4 Part 2 introduces an automated and periodic data collection, processing, and clustering mechanism. Using the PRAW alongside Doc2Vec and KMeans clustering, the application now organizes fetched data into meaningful groups and visualizes the clustering results. This provides an insightful view of the dominant topics within the 'tech' subreddit.

Requirements

- Python 3.x
- PRAW (Python Reddit API Wrapper)
- MySQL Server
- Tesseract-OCR
- Gensim for Doc2Vec
- Scikit-learn for KMeans clustering
- Matplotlib for visualization
- Additional Python libraries: mysql-connector-python, Pillow, requests, nltk, pytesseract, numpy, argparse, json

Installation

1. Ensure Python 3.x and MySQL Server are installed.
2. Install Tesseract-OCR:

“sudo apt-get install tesseract-ocr tesseract-ocr-eng”

1. Install the required Python libraries:

“pip install praw mysql-connector-python Pillow requests nltk pytesseract gensim scikit-learn matplotlib numpy argparse”

Running the Application

1. Make sure the MySQL Server is active.
2. Input your Reddit API credentials into the scraper.py and clustering.py scripts.
3. Execute the data scraping and processing script:

“python3 scraper.py <interval_minutes>"

1. Run the clustering and visualization script:

“python3 clustering.py”

Notes

- The scripts are set to periodically fetch and process 5000 posts from Reddit's 'tech' subreddit.
- scraper.py handles data scraping, initial processing, and storing in the database.
- clustering.py retrieves stored posts, applies clustering, and provides a visualization of the data clusters.
- To ensure proper operation, scripts must be run with the specified interval in minutes as a command-line argument.