Team name: HealthEdBot
Names & USC ID
Cici Chang (8706354957)
Daoyang Li (7168949829)
Yifan Yang(8386626967)
https://github.com/DSCI560

README for Lab4 Part2 Python Script

This document provides a detailed overview of a Python script designed for data processing, natural language processing (NLP), and clustering analysis. The script utilizes several popular Python libraries to accomplish tasks such as data cleaning, keyword extraction, document vectorization, and clustering of textual data.

Imported Libraries

- pandas: Data manipulation and analysis.
- numpy: Support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays.
- re: Regular expression operations for string searching and manipulation.
- urllib.parse: Parsing URLs to break them down into their components.
- spacy: Advanced Natural Language Processing in Python.
- nltk: Natural Language Toolkit, a suite of libraries and programs for symbolic and statistical natural language processing.
- gensim: Modeling of textual data, specifically for topic modeling and document similarity.
- sklearn.cluster: Machine learning library for Python, providing KMeans clustering.
- matplotlib.pyplot: Plotting library for creating static, interactive, and animated visualizations in Python.
- sklearn.decomposition: Principal Component Analysis (PCA) for dimensionality reduction.
- schedule: Job scheduling for Python.
- time: Time access and conversions, used here for scheduling.

Workflow Description

The script begins by loading a dataset of posts from a CSV file, followed by data cleaning to remove duplicates. It then defines a series of functions for preprocessing the data, including extracting the domain from URLs, removing stopwords, and extracting keywords and topics from the titles using Spacy and NLTK. Next, it vectorizes the titles using Doc2Vec from the Gensim library for further analysis.

The processed data is then clustered using the KMeans algorithm from scikit-learn, based on the vectorized titles. A Principal Component Analysis (PCA) is applied to the vectors for dimensionality reduction, enabling the visualization of clusters in a 2D plot using Matplotlib. Lastly, the script includes a

scheduling component that allows for the periodic execution of a predefined job, in this case, updating the database.

User Interaction

The script allows for user interaction through a command-line interface. Users can input keywords to search within the clustered data. Based on the input, the script displays the closest cluster and its messages, optionally visualizing the clusters each time a search is performed.

Team name: HealthEdBot
Names & USC ID
Cici Chang (8706354957)
Daoyang Li (7168949829)
Yifan Yang(8386626967)
https://github.com/DSCI560