OJASHWEE RAMAN

Email: ojashwee.raman2021@vitstudent.ac.in | Phone: 9693592086 | LinkedIn: linkedin.com/in/ojashweeraman

Classification of articles based on title and content.

Code:

```
import pandas as pd
from fuzzywuzzy import process

# Read CSV files
classification_df = pd.read_csv("Dataset_Classification.csv")
categories_df = pd.read_csv("Categories_Description.csv")

# Preprocess data
classification_df['title'] = classification_df['title'].str.lower().str.strip()
categories_df['name'] = categories_df['name'].str.lower().str.strip()

# Function to find category for a title
def find_category(title):
    category_names = categories_df['name']
    matched_name, score, index = process.extractOne(title, category_names)

    # Set a threshold score for matching
    if score >= 80:  # You can adjust the threshold as needed
        return matched_name
    else:
        return "Unknown"

# Add a new 'category' column to the classification dataframe
```

```
classification_df['category'] = classification_df['title'].apply(find_category)


# Print the updated classification dataframe
print(classification_df)


# Save the updated classification dataframe to a new CSV file
classification_df.to_csv("Updated_Dataset_Classification.csv", index=False)
```

Library used:

1. pandas:

- Description: Pandas is a powerful data manipulation and analysis library for Python. It provides data structures like DataFrame and Series, which allow you to easily manipulate and analyze tabular data.
- Purpose in Code: Used for reading and manipulating CSV files, as well as adding and printing data in a tabular format.

2. fuzzywuzzy:

- FuzzyWuzzy is a library in Python for approximate string matching and comparison. It uses algorithms like Levenshtein Distance to calculate the difference between strings, making it useful for tasks involving similarity and matching between strings.
-

In summary, the code utilizes the `pandas` library for reading, preprocessing, and managing data in tabular format, and it employs the `fuzzywuzzy` library for performing approximate string matching to associate classification titles with corresponding categories.

Working of the Code:

- Import necessary libraries: The code starts by importing the `pandas` library for data manipulation and the `process` function from the `fuzzywuzzy` library for fuzzy string matching.
- Read CSV files: It reads two CSV files, `Dataset_Classification.csv` and `Categories_Description.csv`, using the `pd.read_csv` function from the `pandas` library.
- Preprocess data: It converts the text in the 'title' and 'name' columns of the dataframes to lowercase and removes any leading or trailing whitespaces.
- Define the `find_category` function: This function takes a 'title' as input and attempts to find a matching category from the 'Categories_Description.csv' dataframe using

fuzzy string matching. It uses the `process.extractOne` function from `fuzzywuzzy` to find the closest matching category name and calculates a matching score.

- Set a matching threshold: The code sets a threshold score of 80 for a match. If the matching score is equal to or higher than this threshold, the function returns the matched category name; otherwise, it returns "Unknown".
- Add a new 'category' column: The code applies the `find_category` function to each row in the 'title' column of the 'classification_df' dataframe and adds a new 'category' column with the matched category or "Unknown".
- Print the updated classification dataframe: The code prints the updated 'classification_df' dataframe, which now includes the newly added 'category' column.
- Save the updated dataframe: The code saves the updated 'classification_df' dataframe to a new CSV file named "Updated_Dataset_Classification.csv" using the `to_csv` function from the `pandas` library.

Output:

```
                                category
0   diversity, equity and inclusion (dei)
1                 positive financial news
2                                 Unknown
3                   commercial litigation
4                                 Unknown
..                                    ...
95              regulatory or legal issues
96                                Unknown
97                                Unknown
98                                Unknown
99                                Unknown

[100 rows x 8 columns]
```

	id	dt_created	news
0	70804	09-08-2023 08:38	7723975
1	70803	09-08-2023 08:28	7723967
2	70802	09-08-2023 08:27	7723952
3	70801	09-08-2023 08:26	7723954
4	70792	09-08-2023 03:00	7723936
..	...	...	...
95	70586	09-08-2023 02:14	7723599
96	70585	09-08-2023 02:13	7723595
97	70584	09-08-2023 02:13	7723592
98	70583	09-08-2023 02:13	7723587
99	70582	09-08-2023 02:13	7723588

	title
0	firefly aerospace debuts elytra orbital vehicl...
1	virgin galactic announces date of second quart...
2	northrop grumman completes first next generation
3	virgin galactic announces flight window for se...
4	millionero seizes best emerging crypto exchang...
..	...
95	nz issues important documents outlining defenc...
96	algernon neuroscience announces phase 2 dmt st...
97	nsa transitions sharkseer cyber defense tool t...
98	dmt as a treatment for stroke? this company is...
99	nursing home occupancy creeps up to hover at 8...

	snippet
0	Firefly Aerospace, a space transportation comp...
1	Virgin Galactic Holdings, an aerospace and spa...
2	Northrop Grumman Corporation has successfully ...
3	Virgin Galactic has announced the flight windo...
4	Singapore's vibrant cityscape played host to t...
..	...
95	On August 4, New Zealand released three key do...
96	Algernon Pharmaceuticals' subsidiary, Algernon...
97	The Defense Information Systems Agency (DISA) ...
98	Algernon NeuroScience, a subsidiary of Algerno...
99	The demand for skilled nursing care is recover...

	content
0	Formerly known as Firefly's Space Utility Vehi...
1	TUSTIN, Calif.--(BUSINESS WIRE)--Jul 18, 2023-...
2	CLEARFIELD, Utah, Aug. 08, 2023 (GLOBE NEWSWIR...
3	ORANGE COUNTY, Calif.--(BUSINESS WIRE)--Virg...
4	Dubai, UAE, Aug. 08, 2023 (GLOBE NEWSWIRE) -- ...
..	...
95	NZ issues important documents outlining defenc...
96	For more on Algernon's DMT work, see Algernon ...
97	The Defense Information Systems Agency has ass...
98	Algernon NeuroScience (AGN Neuro) , a subsidia...
99	While staffing shortages continue to create ad...

	news__source
0	https://www.prnewswire.com/news-releases/firef...
1	https://www.businesswire.com/news/home/2023071...
2	https://www.globenewswire.com/news-release/202...
3	https://www.businesswire.com/news/home/2023071...
4	https://www.benzinga.com/pressreleases/23/08/g...
..	...
95	https://www.shephardmedia.com/news/defence-not...
96	https://microdose.buzz/news/algernon-neuroscie...
97	https://executivegov.com/2023/08/nsa-transitio...
98	https://www.benzinga.com/markets/cannabis/23/0...
99	https://skillednursingnews.com/2023/08/nursing...

	category
0	diversity, equity and inclusion (dei)
1	positive financial news
2	Unknown
3	commercial litigation
4	Unknown
..	...
95	regulatory or legal issues
96	Unknown
97	Unknown
98	Unknown
99	Unknown

[100 rows x 8 columns]