Documentation: 3.2 Step - web scraping using BeautifulSoup

Overview

This document explains a Python script that searches the web for information about companies, collects relevant text, and analyzes it to determine its meaning and sentiment. The goal is to understand whether the content mentions anything risky, positive, or neutral about each company.

What Does the Script Do?

1. **Search for Company Information:** Uses Bing search to find web pages mentioning specific company names.
2. **Extract Text from Web Pages:** Collects meaningful text from those pages.
3. **Analyze the Text:** Identifies keywords, checks sentiment (positive, negative, or neutral), and extracts names of people, organizations, or places.
4. **Score the Information:** Assigns a score based on keywords and sentiment.
5. **Save the Results:** Outputs the findings in a CSV file for easy analysis.

Required Tools and Libraries

Make sure you have these installed:

- **pandas:** For working with data tables.
- **requests:** To fetch web pages.
- **BeautifulSoup (bs4):** To extract text from web pages.
- **nltk:** For sentiment analysis.
- **spacy:** For natural language analysis.

Run this command to install them if needed:

pip install pandas requests beautifulsoup4 nltk spacy

And download the sentiment lexicon:

import nltk

nltk.download('vader_lexicon')

How It Works

1. Company Data Input

- The script reads company names from a CSV file (`Offshore Leaks-entities.csv`).
- It processes the first 20 company names for efficiency.

2. Web Search

- Searches Bing for each company name.
- Filters out social media and irrelevant pages.

3. Extracting Text

- Visits valid web pages.
- Extracts meaningful text and ignores unwanted content like ads or menus.

4. Keyword Matching

- Looks for important words in the text:
    - **High-Risk Words (+30 points):** crime, fraud, money laundering
    - **Medium-Risk Words (+5 points):** penalty, lawsuit, investigation
    - **Negative Words (-1 point):** stock

5. Sentiment Analysis

- Determines if the text feels **positive**, **negative**, or **neutral**.
- Adjusts the score based on sentiment.

6. Scoring System

- Combines keyword matches and sentiment results.
- Higher scores indicate higher relevance or risk.

7. Save Results

- Results are saved in a CSV file (`Step_3.2.2_company_analysis_with_scores.csv`) with these columns:
    - **Company Name**
    - **URL**
    - **Extracted Text (snippet)**
    - **Score**
    - **Matched Keywords**
    - **Entities (e.g., names of people, places)**

- **Sentiment** (Positive, Negative, Neutral)

Important Functions Explained

- `random_headers()` – Rotates user-agent headers to avoid detection.
- `is_valid_url(url)` – Skips unwanted domains like social media.
- `bing_search_scrape(company)` – Searches Bing for a company.
- `extract_text_from_url(url, company_name)` – Extracts clean text from web pages.
- `analyze_text_with_nlp(text)` – Finds key names and sentiment in the text.
- `calculate_score_with_reason(text, snippet, company_name)` – Assigns a score based on keywords and sentiment.
- `process_company(company_name)` – Runs all the above steps for one company.

Multithreading for Speed

- The script processes multiple companies at the same time using multithreading.
- It pauses briefly between batches to prevent overloading search engines.

How to Run the Script

1. Ensure all dependencies are installed.
2. Place the CSV file (`Offshore Leaks-entities.csv`) in the same folder.
3. Run the script:

python script_name.py

4. Check the output CSV file for results.

Example Output

Company	URL	Score	Sentiment	Matched Keywords	Entities

ABC Corp	example.com/abc	35	Negative	fraud, scam	John Doe, USA
XYZ Ltd	example.com/xyz	5	Neutral	penalty	Europe

Troubleshooting

- **Error fetching data:** Check your internet connection.
- **Blocked IP:** Reduce the number of threads or use a VPN.
- **Empty results:** Increase the number of company names.

Future Improvements

- Use proxies to avoid blocking.
- Add more advanced keyword detection.
- Support for additional search engines.

Author: Natalja Krjuckova
Date: 24/12/2024

End of Documentation