Data Science Assignment

Goal:

Find **contextually similar articles across** two websites moneycontrol.com and economictimes.indiatimes.com

Description:

1. You need to scrape all the links (href or hyperlinks)of news articles from www.moneycontrol.com and economictimes.indiatimes.com/ homepage (only homepage, no need to scrape other sections)

    - We are only interested in news articles so feel free to avoid non-news hyperlinks.
    - save links in a local database (preferably MongoDB, but you can choose any other Database as well). Say you get N hyperlinks from the websites.

2. Scrape each hyperlink you got in the above step and collect text data from the html page of the link (take the first 200 words from each article if article size if greater than 200 words).

3. Now that you have text data for each link, You have to find **contextually similar articles**,
    - For that you have to do some text preprocessing.
    - Do the necessary steps of text preprocessing (tokenize, stopwords etc.) that you think would be required to find contextual similarities between the articles, and write down why you are doing that text processing step alongside the code in the jupyter notebook.
    - Once you have processed text data, try different techniques available for getting contextual similarities between the articles like word frequencies, tf-idf, word2vec, doc2vec etc.
    - For each technique you apply for contextual similarity, generate results as: **for each link from moneycontrol website get the link that is contextually most similar to that link from economictimes website**, examples are given below.
    - For each technique, Manually verify the results that you are getting and point out limitations and advantages of each technique you are using. Summarize your results and findings in 200-250 words.

Submission:

- Code in the Jupyter notebook.
- Write Comments to explain your approach

Example:

Example for Similar news articles across both websites

1. https://www.moneycontrol.com/news/business/shiv-nadar-steps-down-as-chairman-of-hcl-technologies-board-of-directors-5559131.html

   Most similar article:
https://economictimes.indiatimes.com/tech/ites/change-of-guard-shiv-nadar-steps-down-as-hcl-chairman-daughter-roshni-takes-over/articleshow/77012158.cms

2. https://www.moneycontrol.com/news/business/earnings/britannia-q1-profit-jumps-117-to-rs-545-7-crore-revenue-up-27-5559661.html

   Most similar article:
   https://economictimes.indiatimes.com/markets/stocks/earnings/britannia-q1-results-profit-more-than-doubles-to-rs-543-crore-beats-street-estimates/articleshow/77014826.cms