Python Intern Final Assessment Task

Task Window: 28 hours

Deadline: 11:59 PM, 18 October, 2024

Suppose you are given a django model:

class News(models.Model):

url = models.URLField(max\_length=2048, unique=True)

title = models.CharField(max\_length=255)

meta\_description = models.TextField()

news\_type = models.CharField(max\_length=20)

news\_subcategory = models.CharField(max\_length=20)

media\_type = models.CharField(max\_length=255)

image\_urls = models.TextField()

published\_date = models.DateTimeField(null=True, blank=True)

updated\_date = models.DateTimeField(null=True, blank=True)

keywords = models.TextField(blank=True, null=True)

source = models.CharField(max\_length=255)

last\_scraped = models.DateTimeField()

international = models.BooleanField()

old = models.BooleanField(null=True, blank=True)

sentiment = models.CharField(max\_length=15)

views = models.IntegerField(default=0)

news\_score = models.FloatField(default=0.0)

rating = models.FloatField(default=0.0)

engagement = models.IntegerField(default=0)

author = models.CharField(max\_length=255)

content = models.TextField()

class Meta:

db\_table = 'news'

def \_\_str\_\_(self):

return self.title

**Model Description:**

url : The URL of a news article like- <https://bangla.bdnews24.com/tech/f7b7b6bab500>

title : The title of the article

meta\_description : The meta\_description of the article

news\_type : The news\_type of the article. The categories:

1. Sports

2. Economics

3.Politics

4. National

5. International

6. Entertainment

7. Tech

8. Opinion

9. Lifestyle

10. Science

11. Health

12. Crime

13. Education

14. Job Search

news\_subcategory : The news\_subcategory of the article like for Sports category sub categories can be cricket, football, golf, tenis, etc.

media\_type : The media\_type of the article-TV Media/Newspaper/ Online/Magazine

image\_url : The image\_url of the article

published\_date : The published\_date of the article

updated\_date : The updated\_date of the article(if any)

keywords : The keywords of the article (generated by a LLM if there are no keywords found in the webpage)

source : The source of the article

last\_scraped : The time of scraping the article

international : True if the news has an international perspective in terms of Bangladesh (generated by a LLM)

old : True if the article is older than 3 days else False

sentiment : The sentiment of the article - positive, neutral or negative (generated by a LLM)

views : This data will be taken as 0 by default

news\_score : The importance score of the article in the context of Bangladesh (generated by a LLM)

rating : This data will be taken as 0 by default

engagement : This data will be taken as 0 by default

author : The author of the article

content : The full content of the article

To populate the database you need to scrape some data from a popular News Portal like-

1. Prothom Alo
2. Daily Star
3. Samakal
4. Dhaka Tribune
5. The Business Standard
6. Kaler kontho

You will be given a url to scrape in the email sent to you. The url will be a certain category page of the news portal for example- <https://www.jugantor.com/sports>

1. Your task will be to scrape all the urls from that given url and filter out the urls of that certain category. For example- for the url <https://www.jugantor.com/sports> you will extract all the urls that are related to sports and avoid the others.
2. Then after extracting the urls, you will extract all possible data for each url according to the requirements of the model except the ones to be generated by LLM and the ones with default value 0 in a json file.
3. Finally, use an open-source LLM to generate a score of the importance of the news and to analyze the sentiment and international perspective of the news using Groq/Gemini/Nvidia/Huggingface (Free API for inference). Integrate the data into the json.
4. Upload your code and the produced json file in a github public repository and send it replying to the email sent to you.