Programming test from Knorex R&P

Position: Machine Learning Operations (MLOps)

*(There are **two questions** in the test, you are required to work on Question 1 and Question 2 only.*

The final one is an optional question to gain more credits, you can skip it.)

Question 1. You are given a file named `raw-bid-win.tar.gz` in the `question1` folder. Please write a program in **Java** or **Python 3** to:

1. Extract the file
2. Parse the files in raw-bid-win folder (which contains lines JSON string) into JSON objects (with the schema described below), then insert them into MongoDB under collection named `individualWins` of database named `analyticsInterview`.

#	Fields	Descriptions
1	auctionId	Self explanatory and found easily in the JSON object, default value: "NA"
2	campaignId	Found under `biddingMainAccount` of the JSON object, default value: "NA"
3	creativeId	Found under `bidResponseCreativeName` of the JSON object, default value: "NA"
4	adgroupId	Found under `biddingSubAccount` of the JSON object, default value: "NA"
5	userAgent	Found under `bidRequestString->userAgent` of the JSON object, default value: "Others"
6	site	Found under `bidRequestString->url` of the JSON object, default value: "Others"
7	geo	Found under `bidRequestString->device->geo->country` of the JSON object, or under `bidRequestString->device->ext->geo_criteria_id` if the former cannot be found, default value: "Others"
8	exchange	Found under `bidRequestString->exchange` of the JSON object, default value: "Others"
9	price	Found under `winPrice` of the JSON object, strip out USD/1M and store as numerical format, default value: 0
10	time	Found under `bidRequestString->timestamp` of the JSON object. Convert the value into Unix Timestamp at the start of the hour, default value: 0

NOTE: If you are unable to parse the line, you may ignore the line. Default value refers to the value you can use if you are unable to parse the key for whatever reason.

Please send back to us: *a zip file containing your java program/python scripts and a dump of your MongoDB containing the data.*

Question 2. Now your teammate, Alice, who is a research scientist, has finished training a deep neural network model using Pytorch. She asks you to help build an API to serve the model so that other teams in Knorex can request via a RESTFUL API to use. To facilitate your work, she gave you two files, including:

- A jupyter notebook to show her model structure as well as how to inference using the pytorch model.
- A state dict (i.e., the weight and bias values of the deep neural network) so that you can load the model and try to inference.

In this task, you will build an API **using python 3** to serve the given model provided in the `question2` folder, named `alice_model.pt`.

NOTE:

- To build such back-end API, we suggest that you can use FastAPI framework, or any framework that you know (for e.g., Flask, Tornado, Django, etc.). However, it would be a plus if you can use FastAPI framework with a demo docker file to illustrate your idea about how to deploy the model.
- The API is required to take 3 numbers as input (which respects to the x tensor), and output 1 number (which respects to the only one value of the output y from the model)

Please send back to us: *a zip file containing all the scripts of your implementation and a README.md file which guides us how to start the API.*

*This question is **optional** only, you can skip this one. However, it should be a big plus if you can solve it.*

Question 3 (optional). Beautiful Soup is a Python library for pulling data out of HTML and XML files. We usually use Beautiful Soup library to extract important content from websites so that we can build datasets utilized for NLP down-stream problems. In this task, you will implement a program **in python 3** using the jupyter notebook named `bs4_and_pandas.ipynb` provided in the `question3` folder to:

1. Extract the main content from this website:
   https://www.investopedia.com/articles/investing/012715/5-richest-people-world.asp
   We aim to get the information of all 10 persons mentioned in the article only, including name, age, residence, co-founder, net worth, ownership stake, and other assets (if the field cannot be found, you can simply put it as None). The remaining text in the article can be ignored.
2. After extracting the required text from the website, please put them into a pandas data frame so that whenever we print that out, it should show the following content:

	Name	Age	Residence	Co-founder	Net worth	Ownership stake	Other assets
0	Elon Musk	50	Texas	Tesla	$223 billion	17% ($150 billion)	Space Exploration Technologies ($40.3 billion private asset), $5.53 billion in cash
1	Jeff Bezos	58	Washington	Amazon	$178 billion	10% ($153 billion)	Blue Origin ($9.15 billion private company stake), The Washington Post ($250 million private asset), and $15.5 billion in cash
...	...	...	...	...	...	...	...

3. Export your pandas data frame into a csv file.

NOTE: If there is any other library rather than Beautiful Soup that you think it might help you to work on this more efficiently, feel free to use that library. In that case, please help input the pip install command or guide us how to install the library using markdown in the notebook.

***Please send back to us:** the `bs4_and_pandas.ipynb` file with your implementation and the csv file.*