Data Engineer Assessment 1 :
The assignment is to create 2 pipelines, one with automated scrapping based on ticker in instructions, second one based on static input .

*** Create pipeline with Airflow**

Pipelines:

- **Pipeline 1:**
  - Data Sources: (search for only two keywords: HDFC, Tata Motors, fetch 5 latest articles for each ticker)
  - Data Source1: https://yourstory.com
  - Data Source2: https://finshots.in/
  - Schedule: 7pm every working day
  - **Steps:**
    - Fetch data (text data of article) from Data Source1, Data Source2
    - Do basic cleaning and processing (prepping/deduplication on title/text data for that ticker) on the data
    - Generate sentiment score for the company (assume a mock/dummy API which can be called for it with input as news text and response as float between 0 to 1)
    - Persist final score in some DB, Data Lake or anything of your choice, and anything else you may consider necessary (with justification)

- **Pipeline 2: schedule: 8pm every working day:**
  - Condition: skip if pipeline 1 has failed/not completed on same day run
  - Data Source : https://grouplens.org/datasets/movielens/, ml-100k.
  - Metadata and other details are given there. http://files.grouplens.org/datasets/movielens/ml-100k-README.txt
  - Create 4 tasks,
    - Find the mean age of users in each occupation
    - Find the names of top 20 highest rated movies. (at least 35 times rated by Users)
    - Find the top genres rated by users of each occupation in every age-groups. age-groups can be defined as 20-25, 25-35, 35-45, 45 and older
    - Given one movie, find top 10 similar movies. The similarity calculation can change according to the algorithm.

Described below is one way of finding similar movies. You can define your own algorithm.

Finding the most similar movies based on user ratings.

users	**movie**	**rating**
U1	M1	R1
U2	M1	R2
U1	M2	R3

Hint: Here, we have to find out if user U1 rated 2 movies M1 and M2, then, how. much similar are they in terms of their ratings.

If we do that for all the users and all the movies, it will give us list of similar movies.

- Constraints: The movies have similarity threshold of 95% and co-occurrence threshold of 50.
- Similarity threshold - Similarity of ratings
- Co-occurrence Threshold - least number of times two movies are rated together by same user.

For example:
Top 10 similar movies for Usual Suspects, The (1995)

Close Shave, A (1995)	score: 0.9819256006071412	strength: 56
L.A. Confidential (1997)	score: 0.9816869323101214	strength: 113
Sling Blade (1996)	score: 0.980468570034675	strength: 94
Rear Window (1954)	score: 0.980441832864182	strength: 115
Shawshank Redemption, The (1994)	score: 0.9792067644351858	strength: 177
Manchurian Candidate, The (1962)	score: 0.9789963985081663	strength: 75
Wrong Trousers, The (1993)	score: 0.9787901543866219	strength: 68
Good Will Hunting (1997)	score: 0.9781245483949754	strength: 65
Apt Pupil (1998)	score: 0.9762169825124449	strength: 54
Godfather, The (1972)	score: 0.9754550490486855	strength: 176

you can do processing using spark/pandas/sql.

- **Constraints:**
  - Pipeline2 should only run when Pipeline1 has successfully ran (all steps) for same day.
  - If at any stage pipeline crashes, we should get alerts (you can mock that API too)

- **Note:**
  - Feel free to assume if at any point you are stuck and write back the justification of assumption.
  - Create a ci/cd from GitHub to pick up latest code
  - Also set up alerts on failure of tasks
  - Please share the code with github link or zip with working Dockerfile to run and install/setup.
  - Please share evidence of working version with a short video recording or screenshots, as applicable.
  - Please make sure to provide a bash script to up the system (docker application and other configurations env variables (if any)). The process should not require any manual inputs