Spark Coding exercises

All required datasets are included in the following Google Drive folder
https://drive.google.com/drive/folders/1eTgPaDEjImkgqeY16U9RLOFR22VcQyoT?usp=sharing

Question 01. Consider the JSON file movies.json, whose content represents a movie dataset. Each record of the dataset stores the information of a movie. You can use some JSON viewer application to better view the dataset (e.g., https://jsonformatter.org/json-viewer)
You first need to load the dataset into a Spark DataFrame and then, for each of the following requirements, write a PySpark code segment to fulfill the goal.

1. Show the schema of DataFrame that stores the movies dataset. Show the number of distinct films in the dataset

2. Count the number of movies released during the years 2012 and 2015 (included)

3. Show the **year in which the number** of movies released is highest. Only one highest year is enough.
4. Show the list of movies such that for each film, the number of actors/actresses is at least five, and the number of genres it belongs to is at most two genres.
5. Show the **movies** whose names are longest

6. Show the movies whose name contains the word “fighting” (case-insensitive).

7. Show the list of distinct genres appearing in the dataset

8. List all movies in which the actor Harrison Ford has participated.

9. List all movies in which the actors/actresses whose names include the word “Lewis“ (case-insensitive) have participated.

10. Show top five actors/actresses that have participated in most movies.

Question 02. Consider the CSV file foodmart.csv, whose content represents a transactional dataset. Each record of the dataset is a tuple of values 1 and 0 corresponding to a designated list of items, in which 1 means bought and 0 means not bought.

1. Convert the given dataset to the following format. Note that in each list of items, consecutive items are separated by a single comma.

ID	Items
1	item1 item2 item3
2	item3 item1
…	…

2. Mine the set of frequent patterns and the set of association rules from the above dataset (in new format) with min support of 0.1 and min confidence of 0.9.

Question 03. Consider the CSV file mushrooms.csv, whose content represents a dataset of mushroom species. There are 8124 examples, each of which is presented by 22 attributes and categorized into either “edible” (e) or “poisonous” (p)

1. Prepare the train and test sets following the ratio 80:20
2. Build a decision tree model on the training set
3. Build a random forest model on the training set
4. Evaluate the two models on the same test set
5. Use a pipeline to simultaneously conduct the above experiments.

Question 04. Consider the CSV file iris.csv, whose content represents a dataset of iris plant

species. There are 150 examples, each of which is presented by 4 attributes and categorized into one of the three classes.

1. Cluster the given examples by using k-means clustering with k = 2, 3, and 5.

   ==========

2. Consider the clustering with k = 2 done above. For each cluster, count the number of examples that belong to each of the **three species.**
3. Repeat the counting above for other values of k.

Question 05. Consider the text file users.txt, in which each line contains the information of a user: id, account name, and username (we only use the first two fields). Also, consider the text file followers.txt, in which each line contains a couple of integers, <source> <dest>, indicating there is a connection from the user whose id is <source> to the user whose id is <dest>.

1. Construct a graph from the above information to represent a tiny social network
2. Apply PageRank to the graph to obtain a ranking list of users in terms of followers
3. Find connected components on the given graph