THE UNIVERSITY OF SYDNEY

School of Computer Science
Dr. Ying Zhou

COMP5349: Cloud Computing **Sem. 1/2021**

Assignment 1: Data Analysis with Spark RDD API

Individual Work: 20% **01.04.2021**

1 Introduction

This assignment tests your ability to implement simple data analytic workload using `Spark` RDD API. The data set you will work on is adapted from Trending Youtube Video Statistics data from Kaggle. There are **two workloads** you should design and implement against the given data set. **You are required to designed and implement both workloads using ONLY basic Spark RDD API**. You should not use Spark SQL or other advanced features.

2 Input Data Set Description

The dataset contains several months' records of daily top trending YouTube video in the following ten countries: Canada,France, Germany, India,Japan, Mexico, Russia, South Korea, United Kingdom and United States of America. There are up to 200 trending videos listed per day.

In the original data set, each country's data is stored in a separate CSV file, with each row representing a trending video record. If a video is listed as trending in multiple days, each trending appearance has its own record. The category names are stored in a few separate JSON files.

The following preprocessing have been done to ensure that you can focus on the main workload design.

- Merge the 10 individual CSV files into a single CSV file;
- Add a column `category` to store the actual category name based on the mapping
- Add a column `country` to store the trending country, each country is represented by two capital letter code.
- Remove rows with invalid video id values
- Remove most columns that are not relevant to the workloads

The results is a CSV file `AllVideos.csv` with 8 columns and no header row. The columns are as follows. The `trending_date` column has the date format: `yy.dd.mm`

```
video_id,trending_date,category,views,likes,dislikes,country
```

3 Analysis Workload Description

3.1 Controversial Trending Videos Identification

Listing a video as trending would help it attract more views. However, not all trending videos are liked by viewers. For some video, listing it as trending would increase its `dislikes` number more than the increase of its `likes` number. This workload aims to identify such videos. Below are a few records of a particular video demonstrating the change of various numbers over time:

video_id	trending_date	views	likes	dislikes	country
QwZT7T-TXT0	2018-01-03	13305605	**835378**	629120	US
QwZT7T-TXT0	2018-01-04	23389090	1082422	1065772	US
QwZT7T-TXT0	...	...	...	...	US
QwZT7T-TXT0	2018-01-09	37539570	1402578	**1674420**	US
QwZT7T-TXT0	2018-01-03	13305605	**835382**	629123	GB
QwZT7T-TXT0	...	...	...	...	GB
QwZT7T-TXT0	2018-01-18	45349447	1572111	**1944971**	GB

The video has multiple trending appearances in US and GB. In both countries, its `views`, `likes` and `dislikes` all increase over time with each trending appearance. As highlighted in the table above, the `dislikes` number grows much faster than the `likes` numbers. In both countries, the video ended with higher number of `dislikes` than `likes` albeit starting with higher `likes` number.

In this workload, you are asked to find out the top 10 videos with fastest growth of `dislikes` number between its first and last trending appearances. Here we measure the growth of `dislikes` number by the difference of `dislikes` increase and `likes` increase between the first and last trending appearances in the same country. For instance, the `dislikes` growth of video `QwZT7T-TXT0` in US is computed as follows:

$(1674420 - 629120) - (1402578 - 835378) = 478100$

The result of this workload should show the video id, dislike growth value and country code. Below is the sample results.

```
'QwZT7T-TXT0', 579119, 'GB'
'QwZT7T-TXT0', 478100, 'US'
'BEePFpC9qG8', 365862, 'DE'
'RmZ3DPJQo2k', 334390, 'KR'
'q8v9MvManKE', 299044, 'IN'
'pOHQdIDds6s', 160365, 'CA'
'ZGEoqPpJQLE', 151913, 'RU'
'84LBjXaeKk4', 134836, 'FR'
'84LBjXaeKk4', 134834, 'DE'
'84LBjXaeKk4', 121240, 'RU'
```

3.2 Category and Trending Correlation

Some videos are trending in multiple countries. We are interested to know if there is any correlation between video category and trending popularity among countries. For instance, we may expect to see a common set of trending music videos in many countries and a distinctive set of trending political videos in each country. In this workload, you are asked to find out the average country number for videos in each category.

The following sample data set contains five videos belonging to category "`Sports`", their trending data are as follows:

video_id	category	trending_date	views	country
1	Sports	18.17.02	700	US
1	Sports	18.18.02	1500	US
2	Sports	18.11.03	3000	US
2	Sports	18.11.03	2000	CA
2	Sports	18.11.03	5000	IN
2	Sports	18.12.03	7000	IN
3	Sports	18.17.04	2000	JP
4	Sports	18.16.04	3000	KR
4	Sports	18.17.04	9000	KR
5	Sports	18.16.04	4000	RU

We can see that video 1,3,4,5, each appears in one country; video 2 appears in three countries; If they are the only videos in `Sports` category, the average country number would be $(1 + 3 + 1 + 1 + 1)/5 = 1.4$ You should print out the final result sorted by the average country number. The sample output of this work load is as follows.

```
('Trailers', 1.0),
('Autos & Vehicles', 1.0190448285965426),
('News & Politics', 1.052844979051223),
('Nonprofits & Activism', 1.057344064386318),
('Education', 1.0628976994615762),
('People & Blogs', 1.0640343760329336),
('Pets & Animals', 1.0707850707850708),
('Howto & Style', 1.0876256925918326),
('Travel & Events', 1.0929411764705883),
('Gaming', 1.0946163477016635),
('Sports', 1.1422245184146431),
('Entertainment', 1.1447534885477444),
('Science & Technology', 1.1626835588828102),
('Film & Animation', 1.1677314564158094),
('Comedy', 1.2144120659156503),
('Movies', 1.25),
('Music', 1.310898044427568),
('Shows', 1.614678899082569)
```

A small number of videos have more than one category name. The category name may change over time. For instance video id "`119YrPUNM28`" has changed its category name from "`News & Politics`" to "`Entertainment`". A video may be given different category names in different countries. For instance, video id "`7klO0p092Y`" is under category "`People & Blogs`" in `CA` and `DE` but under category "`Entertainment`" in `US`. As the number is quite small, you do not need to identify and handle them separately. The sample answer double count them in all categories they appear.

4 Coding and Execution Requirement

Below are requirements on coding and Execution:

- You should implement both workloads in **PySpark** using **Spark RDD API**.
- You should implement both workloads in a single Jupyter notebook. There should be clear indication which cells belong to which workload.
- You should not modify the input data file in any way and your code should read the data file from the same directory as the notebook file.

5 Deliverable and Demo

The source code should be submitted as a single Jupyter notebook file. The due date is week 7 Wednesday 21/04/21 23:59. Please name your notebook file as

`<labCode>-<uniKey>-<firstName>-<lastName>.ipynb`.

There will be a 10 minutes **demo** in week 7/8. **You need to attend the demo to receive mark for this assignment!**

During the demo, the marker will run your notebook on their own environment to check the correctness of the result. You should also have your environment ready to run your code and to answer questions. The marker may ask you to explain the overall computation graph or certain part of the implementation. You may be asked to add some statement in your code to show the structure of an intermediate RDD, or to apply various filters on intermediate RDDs to provide slightly different result.