[redacted]

✔ Done

[redacted]

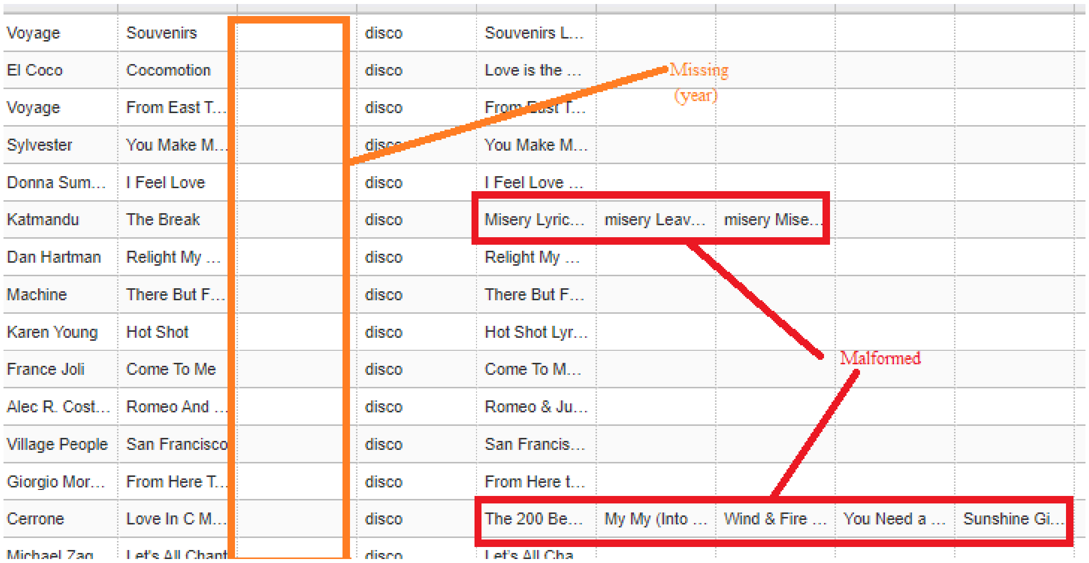
You listened to Taras Matyashovskyy on how to use MLlib to classify music in class. He was classifying music into the two classes "Pop" and "Metal". You can access his code on his GitHub page and read further details from his blog.

However, in his demo, you saw how his system was unable to properly categorize "Hip-Hop" because his model was not trained for it. In this homework, you are going to build a better system. You can use his code as a starting point to help you. But you will have to do changes to it for it to work for this homework.

1. First, download the mendeley music dataset (26MB). Note that it has multiple classes: pop, country, blues, jazz, reggae, rock, hip hop
   - Let's call this dataset **Mendeley dataset**
   - **Note:** In this dataset they have a release_date column. But it is actually not an actual date but just the released year. We will continue this convention.
2. Use a training/test split of 80/20
3. Create and train an MLlib pipeline to classify song lyrics to the above 7 classes. (You can alter Taras Matyashovskyy's code or write your own code from the ground up).
4. Create a web-based user interface where a user can copy-paste lyrics and press a button to activate your system (similar to the demo of Taras Matyashovskyy). But instead of just printing the name of the classified class, your system should show the result in a visual way (eg: pie chart with numbers, bar graph, etc). You are free to select the visual representation.
5. Select a music genre that is not one of the above 7 classes. Create a new data set with at least 100 songs of that genre.
   - If you select one of the following genres, you need to make sure that the 100 songs you collect are not in the following OSF datasets already: Dance, Disco, Electro, Folk, Funk, Metal, Punk, Rap, RnB, Techno
   - **Note:** In your new data set, you only need to have the following columns: artist_name, track_name, release_date, genre, lyrics
   - Remember, the release_date column only should have the year. Not the date.
   - How you collect the data is up to you. You can do it manually if you like. On the other hand, if you want to automate it, you can use the genius API and LyricsGenius or a similar method. I am not going to evaluate how you collected the data. So **do not** upload any code (if any) you used to collect data. Here I just want you to create the data set. How you do it, is totally up to you.
   - If you want to go beyond 100 songs, do so. It will improve your accuracy in the following steps. But only 100 songs are compulsory.
   - Let's call this dataset the **Student dataset** which should have the columns (artist_name, track_name, release_date, genre, lyrics)
6. Merge the relevant columns (artist_name, track_name, release_date, genre, lyrics) of the **Mendeley dataset** and **Student dataset** to obtain the **Merged dataset**.
7. Now, re-train your model; from step 3 to classify all 8 classes. If you selected one of the genres in the OSF data sets (Dance, Disco, Electro, Folk, Funk, Metal, Punk, Rap, RnB, Techno), you are allowed to use your **Merged dataset** as well as the OSF data set to train your model. If you selected a brand new genre (eg: Retro, Soul, Electro), you can stick with your **Merged dataset** to train.
8. Update your visualization to handle all 8 classes.

Deliverable:
One zip file containing:

1. Your MLlib code
2. Trained final model
3. Student dataset (**Student_dataset.csv**)
    - If this dataset has columns more or less than (artist_name, track_name, release_date, genre, lyrics), you do not get the points for this.
    - If the content in any column is malformed or missing, you do not get points for this.



Voyage	Souvenirs		disco	Souvenirs L…				
El Coco	Cocomotion		disco	Love is the …				
Voyage	From East T…		disco	From East T…				
Sylvester	You Make M…		disco	You Make M…				
Donna Sum…	I Feel Love		disco	I Feel Love …				
Katmandu	The Break		disco	Misery Lyric…	misery Leav…	misery Mise…		
Dan Hartman	Relight My …		disco	Relight My …				
Machine	There But F…		disco	There But F…				
Karen Young	Hot Shot		disco	Hot Shot Lyr…				
France Joli	Come To Me		disco	Come To M…				
Alec R. Cost…	Romeo And …		disco	Romeo & Ju…				
Village People	San Francisco		disco	San Francis…				
Giorgio Mor…	From Here T…		disco	From Here t…				
Cerrone	Love In C M…		disco	The 200 Be…	My My (Into …	Wind & Fire …	You Need a …	Sunshine Gi…
Michael Zag	Let's All Chant		disco	Let's All Cha				

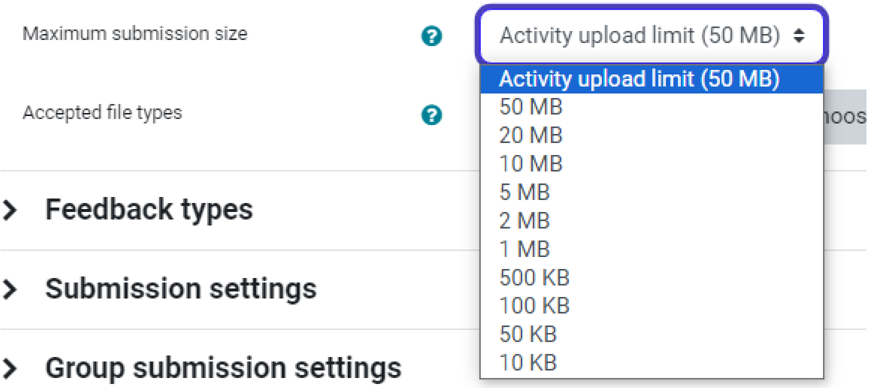
4. Merged dataset (**Merged_dataset.csv**)
5. A single run file (**run.bat**. If and only if you have no other choice, a **run.sh** is acceptable)
    - When I run the above bat file it should launch the visualization webpage
    - It should run the spak commands to load and run your model.
    - I should be able to copy-paste some song text in the browser and press a single button to get the classification run and visualization shown.
    - You can assume that I have set my classpath for spark. So you can call spark at any place by simply executing "spark-shell"
6. A PDF file with a **clickable link** to a Google Drive folder where one or more of the above files are uploaded. You should use this option **if and only if** your zip file exceeds 50MB. You should not put the video file (described below) in the google drive. it should go in the zip file. In addition, it is **highly recommended** to try very hard to put the csv files in the zip file as well.
7. A 1 min video of you demonstrating your system. The video **should not exceed 1 min** and should show the following:
    - You initiating the system.
    - You copy-pasting 4 songs belonging to 4 out of the 7 original classes from the **Mendeley dataset** and show the visualization result. I should be able to see what the actual classes are when you copy the lyrics and then see what the predicted class is. If I don't see these, you do not get the grade for this part.
    - You copy-pasting 1 song belonging to the extra class you added to the **Student dataset** and show the visualization result.
    - You copy-pasting 1 song of a genre not belonging to any of the 8 classes your system has trained for and show the visualization result. If the song belongs to one of the trained classes, you do not get the grade for this part.

Naming Conventions:

- The zip file should be named as <your Index number>.zip. For example, if your index number is 123456X, it should be named as 123456X.zip
- MP4, MKV, AVI, or any other video format that is compatible with VLC media player 3.0.12 is acceptable.
- The video also **must** be named with your index number. For example, if your index number is 123456X, and you are using MP4, it should be named as 123456X.mp4
- The **run.bat**, the **video file**, **Student_dataset.csv**, and **Merged_dataset.csv** should be placed at the root of the zip file. (i.e., not within any folder). You can use folders for other files in your submission. But these four files should be kept at the root.

MLlib and Visualization Homework File size

A number of students have complained about the file size. But due to the university no longer having money to maintain storage, they do not give us the ability to let you upload files greater than 50MB.



Maximum submission size: Activity upload limit (50 MB)

Accepted file types

So as a solution, I am going to allow you to put your model file (and or other files EXCLUDING the video file) on a google drive folder and put the shared link in a pdf in the uploaded zip file. But that is the ONLY change. Any violation will earn a zero for that component. That includes but not limited to:

1. The video file MUST be in the zip uploaded to the moodle. NOT in the google drive folder. (refer to the "Redirect penalty" below)
2. The google drive folder must be shared as "anyone with the link can view". We will NOT wait around "requesting access". We click the link and if it doesn't load, we give a zero for the missing parts and move on to the next student.
3. We have asked for a clickable link on the pdf. If your link is in any other form, (eg: in a txt file where we need to manually copy-paste), it will be ignored and the relevant parts will be considered missing and given zeros.

FAQs

Q: Can I use other tools/technologies (eg: pySpark)
A: The TA (or me depending on who is grading) will set up the MLlib environment. Then if there is a doubt or an issue in the student submission we will run the code and try to give the student some marks No other systems will be installed. Because if we do that N number of students will use N number of technologies and frameworks. We do not pay our TAs enough to suffer that much. So ... ultimately the point is the code submission is there as a backup plan or an insurance policy to give the student "some" points in case the default parts fail for some reason. You may use other technologies but in doing so, you should understand that you will not have that backup grade. But as the explanation goes, it is a backup. If you do a good job in the default submission, maybe you won't even need it. So if you are 100% confident of the other parts of the submission, you can take the risk of using something else.

Penalties

1. The zip file name is anything but your index number: **-20 Points**
2. The video file name is anything but your index number: **-20 Points**
3. Anything other than a single zip file is uploaded: A total grade of **zero (0)** is given without consideration.
4. Time Penalties for the video (in minutes):
   - 0:30 - 1:00: no penalty
   - 1:00 - 2:00: Time Penalty: **-5 Points**
   - 2:00 - 3:00: Time Penalty: **-10 Points**
   - 3:00 - 4:00: Time Penalty: **-20 Points**
   - 4:00 - 5:00: Time Penalty: **-35 Points**
   - 5:00 - 6:00: Time Penalty: **-50 Points**
   - 6:00+: A total grade of **zero (0)** is given without watching
5. Late submission penalties (in hours):
   - 00:00: no penalty
   - 0:00 - 12:00: Late Submission Penalty:**-5 Points**
   - 12:00 - 24:00: Late Submission Penalty: **-10 Points**
   - 24:00 - 48:00: Late Submission Penalty: **-20 Points**
   - 48:00 - 72:00: Late Submission Penalty: **-40 Points**
   - 72:00+: A total grade of **zero (0)** is given without watching

6. Redirect penalty (ie, if the video file is not in the zip file and instead a link is given):**-15 Points**
7. If I note that you have stolen the data rows in **Student_dataset** from some other data set instead of curating it yourself: **-50 Points**

[redacted]

Submission status	[redacted]
Grading status	[redacted]
Time remaining	[redacted]
Last modified	[redacted]
File submissions	[redacted] [redacted]
Submission comments	▸ Comments (0)

Contact us

🌐 ☎ ✉

Follow us

f in

Contact site support

[redacted]

Data retention summary

Get the mobile app

Get the mobile app

This theme was developed by

conecti.me