Homework 4 Part B

10-605/805: Machine Learning with Large Datasets

Due Thursday, November 3rd at 11:59PM Eastern Time

Instructions: There are two parts to this homework, which will have **different deadlines**.

- Part A is due on October 27th and is worth 20% of the grade.
- Part B (i.e., this document) is due on November 3rd and is worth the remaining 80% of your grade.

Submit your solutions via Gradescope, following the template below. Note that this assignment does not contain a theoretical written part and the programming part is more open-ended than previous homework. Because of this, we will not be autograding your submission. Instead, you will submit a report consisting of responses to questions asked in the notebook. Submit this part to the Homework 4: Part B submission slot. We may refer to your code submission to verify your work or in case we have any doubts. You must submit your code to Gradescope under the HW4 Programming submission slot to receive credit for the assignment.

Submitting via Gradescope: When submitting on Gradescope, you must assign pages to each question correctly (it prompts you to do this after submitting your work). This significantly streamlines the grading process for the course staff. Failure to do this may result in a score of 0 for any questions that you didn't correctly assign pages to. It is also your responsibility to make sure that your scan/submission is legible so that we can grade it.

1 Part A: Data Conversion and Preparation

Please complete Part A following it's write-up before attempting this part.

2 Part B: Modeling

In this part of the homework, you will perform exploratory data analysis (EDA) and data cleaning, and then train models with the original features. You will then perform feature engineering similar to what we did in Homework 1 (TF-IDF and Bag-of-Words), and then train models with these new features.

2.1 Setting up EMR and Spark

With our data ready in S3, it's now time to configure and create an EMR (Elastic MapReduce) cluster and run Spark, starting from the notebook `hw4.ipynb`. Include at least Hadoop, JupyterHub, and Spark in your cluster software configuration. Set `maximizeResourceAllocation` to true (see here) in software settings. Select your EC2 key pair.

You'll have to do ssh port forwarding (recommended) or attach additional security groups/inbound rules to the master node in order to access the JupyterHub web interface. This step could be tricky and we cover how to do this in detail in lecture. Then, log in to jupyter (learn about login credentials here), upload your notebook, and start working!

Again, cost management is key. Because we might be running a cluster of machines, this could easily blow up your budget. We recommend using at most 1 Driver and 1 Core of type `m5.xlarge` while developing and debugging on a subset of MSD. You could scale this up to multiple Core workers when doing the final run. You may also utilize AWS spot instances to save cost during development.

Note that, although EMR is made up of EC2 instances, unlike EC2, you cannot stop an EMR cluster—you can only terminate it. See here for why this is the case. You should plan your strategy accordingly. **Do not forget to download your code** before you terminate a cluster when you are done.

2.2 Preprocessing

To run the Jupyter notebook, `hw4.ipynb`, for preprocessing, you would first need to copy this file into your cluster by using the following command on your local terminal:

```
scp -i <your_keypair> <file> <destination>
```

After that ssh into your cluster by using the following command:

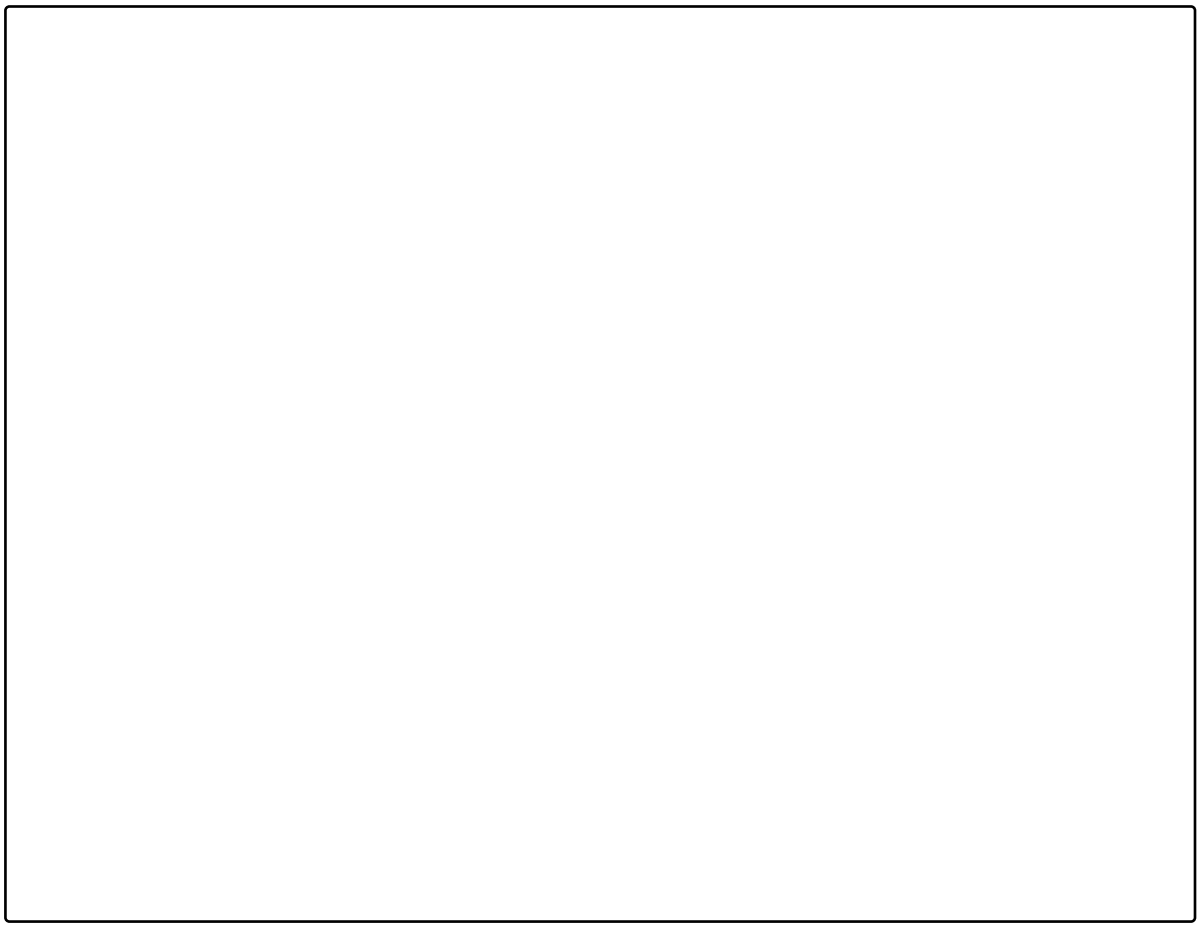
```
ssh -i <your_keypair> -N -L localhost:<local_port>:<cluster>:9443 hadoop@<cluster>
```

Change the S3 bucket name to be the one you created when you set up your EC2 instance while loading data in Part A.

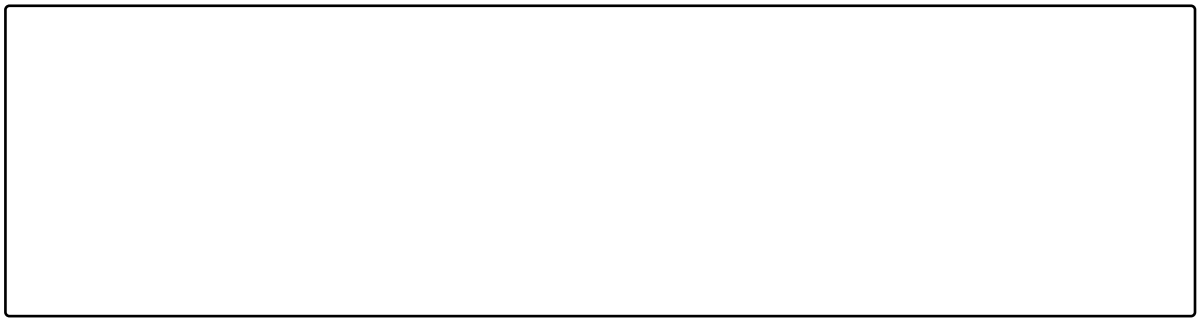
2.3 Exploratory Data Analysis

(a) *[2 points]* Explain why the two features seem problematic (after performing .summary() operation).

(b) *[3 points]* Histograms (remember to label them)



(c) *[3 points]* Explain what is strange about `year`'s distribution and what might cause this. Describe how you could filter `year` to make its histogram look more balanced.



(d) *[2 points]* New histogram for `year`.

(e) *[5 points]* Provide plots for the three pairs. Describe your findings.

(f) *[3 points]* Think about what simple technique you could use to visualize large datasets while retaining a similar data distribution. Briefly describe what you did.

2.4 Data Cleaning

(a) *[2 points]* Your justification for dropping the two features.

(b) *[5 points]* Compare the two numbers and explain the advantages and potential problem of doing this step. What other techniques could you use to potentially do better?

(c) *[2 points]* State the two features.

(d) *[6 points]* Explain your proposed solution and discuss its pros and cons.

(e) *[2 points]* Report the percentage:

2.5 Baseline

(a) *[3 points]* Explain why treating this as a classification problem might be a sensible choice.

(b) *[2 points]* Report what percentage of songs are assigned the "popular" label.

(c) *[2 points]* Explain why we shift the year.

(d) *[5 points]* Explain what scaling means and why we want to perform scaling before the learning step.

(e) *[5 points]* Explain the difference between these two metrics and when AUC might be more useful than accuracy.

(f) *[8 points]* Calculate the train and test AUC of both models and report them.

Models	Train AUC	Test AUC
Logistic Regression		
Random Forest		

2.6 Featurization: Bag-of-Words and TF-IDF

(a) *[3 points]* Explain what the `vocabSize` hyperparameter means in the context of Bag-of-Words.

(b) *[3 points]* Other than featurizing texts, what other feature engineering would you do on the dataset? Briefly describe one.

(c) *[3 points]* Explain where this number "31" comes from.

2.7 Modeling with New Features

(a) *[8 points]* Evaluate train and test AUC for each model and report them.

Models	Train AUC	Test AUC
Logistic Regression		
Random Forest		

(b) *[8 points]* Include the plot and your explanations.

2.8 Do Your Best

(a) *[2 points]* Your final AUC:.

(b) *[4 points]* Your model and hyperparameters.

(c) *[4 points]* Describe your approach.

2.9 Reflection

[5 points] What challenges did you face in HW4? How did you overcome these challenges? What did you learn from HW4 ?"

3 Collaboration Questions

1. (a) Did you receive any help whatsoever from anyone in solving this assignment?

   (b) If you answered 'yes', give full details (e.g. "Jane Doe explained to me what is asked in Question 3.4")

2. (a) Did you give any help whatsoever to anyone in solving this assignment?

   (b) If you answered 'yes', give full details (e.g. "I pointed Joe Smith to section 2.3 since he didn't know how to proceed with Question 2")

3. (a) Did you find or come across code that implements any part of this assignment?

   (b) If you answered 'yes', give full details (book & page, URL & location within the page, etc.).