MEMORANDUM

To: Foaad Khosmood, Lecturer, Department of Computer Science, Cal Poly SLO

Foaad@calpoly.edu

From: Chandler Jones

cjones87@calpoly.edu

Date: March 3rd, 2021

RE: **Laboratory 1 – AI Movie Data Generator**

As practice for writing an academic paper, I will begin structuring lab reports in the format we are to read the weeks' papers in, with the eight questions. Proper section headings will be developed for next lab.

1. What is the problem the authors are trying to solve?

The situation is this: People have lots of ideas about movies. They can write a short synopsis of the plot. What they want to know is who should be acting in the movie? Who should direct it? And lastly, what should the title be?

As a top-level lab this scenario is presented as a machine learning problem introduction to CSC 582. Our author hypothesizes that this would be best served as a multi-feature, multi-label classification problem. Besides addressing the solution to this problem, our author has multiple sub-topics with which to familiarize themselves with before tackling the main problem, namely working with pandas dataframes, working with Sklearn, learning the vocabulary of machine learning, and developing a workflow within Google Colab. These presented significant challenges to overcome.

2. What other approaches or solutions existed at the time that this work was done?

Throughout the duration of the lab, our author came across two solutions which approach the difficulties of addressing this problem set.[1][2] These solutions identify two machine learning projects which approach identifying information about movies using a related dataset. Both methods show significant cleaning of the data set. They also each provide examples as for how to split a dataset into a training set and a validation set. . These methods will each be incorporated as a main feature of our solution.

Prateek shows that using a OneVsRestClassifier and Logistic Regression, it is possible to accurately predict genre's using simply the overview of the movie. This is good news as it takes our same input and picks the highest probabilities of a multivariable output, the genres. This is similar to if we were to train our dataset to thus pick a director, another multi-variable output, with a higher variance.

Patel shows that adapting preexisting transformers such as BERT, XLNet, etc. will work for training a multi-classification problem. He shows that speech sentiment can be modified to show whether speech is "toxic, severely toxic, obscene, a threat, or hate". His methods of retraining the preexisting model seem to work.

3. What was wrong with the other approaches or solutions?

Unfortunately, these guides discuss machine learning at a much higher level than I am otherwise familiar with. When discussing "OneVsRestClassifier", "Binary Cross Entropy With Logits", "Hamming Loss", "micro F1 accuracy", or even really "Multi-Label Classification"; I am at a loss.

A preliminary test was run in Google Colab and the author was able to achieve replication of Prateek's methods. Modification was attempted, continuing to use the overview as a feature, and director as an output, however the significantly higher variance in directors caused critical errors. The model would need to be adjusted.

Preliminary google searches into machine learning proved that the meat was hidden behind buzzword articles.[3]

4. What is the authors' approach or solution?

The authors main solution to the problem, which is really that of definition & understanding of machine learning, has been to work through Google's crash course on Machine Learning.[4] This course was found late into the lab and will be minorly beneficial to this lab, however will likely prove significantly useful as the class matures.

The authors hypothesized solution for the present lab is to first use Prateek's method to find the likely genre, and then pick randomly from the top 5% Directors & Actors matching that category. This has the benefits of getting directors & actors suited to make people laugh, however, in the cases where the plot calls for two female leads, this program will not be able to differentiate that fact and may even provide two male leads. Perhaps that could be funnier?

In anticipation for a simple model where a single feature could predict multiple labels. The dataset was modified within the data frame to include only the relevant columns, namely:

id	tagline	title	new_genre	new_keywords	clean_overview	Directors	Actors

With the new columns containing cleaned versions with stop words eliminated, only the top 5 actors remaining, and all crew removed sans the name of the director. Columns were left as single level lists or strings.

5. Why is it better than the other approaches or solutions?

This is a superior approach to this lab as it provides an investment into proper machine learning techniques over a copy paste solution with little understanding.

6. How did they test their solution?

This solution of diving into Google's machine learning course will be tested through progress updates on upcoming labs.

7. How does it perform?

Performance is yet to be determined. Poor up front performance with hopefully make way for significant progress increases as training increases.

Prateek's methods performed very well, he was generally able to guess a single genre of the 1-3 genre's quoted in the dataset. I was rather impressed at this.

My code performs poorly. It is not my best work and with more gumption I could have done much better. I was constantly face to face with the reality that I was beyond my skill level in this lab, and yet I know if I just keep placing one foot in front of the other and plod forward, I would do alright. The paralysis killed me in this lab, and if I had overcome that from the get-go, I may indeed have finished. Time to put in a more concerted effort.

```
/usr/local/lib/python3.7/dist-packages/sklearn/multiclass.py:75:
UserWarning: Label not 1 is present in all training examples.
  str(classes[c]))
```

8. Why is this work important?

This was an important lab, because it brings me face to face with what is possible in machine learning, and to what online personalities claim is rather an easy solution. It sure does seem like an accurate model will make for a relatively accurate solution. With current tools is further seems that accurate models are relatively easy to create by specifying the necessary features leading to the necessary results. This ingrains in my mind the importance of learning these tools, because they will no doubt be able to be put to good use in my career & academics.

References:

1. https://www.analyticsvidhya.com/blog/2019/04/predicting-movie-genres-nlp-multi-label-classification/
2. https://towardsdatascience.com/journey-to-the-center-of-multi-label-classification-384c40229bff
3. https://machinelearningmastery.com/how-to-define-your-machine-learning-problem/
4. https://developers.google.com/machine-learning/crash-course/framing/ml-terminology
5. https://towardsdatascience.com/transformers-for-multilabel-classification-71a1a0daf5e1
6. https://www.kaggle.com/hsrobo/multi-label-classification-evaluation-template
7. https://www.kaggle.com/tmdb/tmdb-movie-metadata
8. https://colab.research.google.com/drive/13nBqYNPXejN1qD3PGwTpfzX-XIjP0Ccm#scrollTo=QJM4XNW8gvey