COMM061
Natural Language Processing

Group Declaration

Danylo Kovalenko - 6413526
Prakash Jha - 6659329
Jamie Dance - 6661320

UNIVERSITY OF SURREY

Project Topic

Our chosen topic is Genre Classification. We will be classifying Netflix shows and films into a variety of genres based on their description. As each title can fall into different categories, this is a multi-label classification problem. The number of shows on Netflix has nearly tripled since 2010[1] and as a result of the Covid-19 lockdown, the number of users has skyrocketed. Using NLP can help to automate and simplify the process of genre assignment, and ultimately be used to improve the recommendation system, as users can have preferences for a combination of genres.

Dataset

We will be analysing the "Netflix Movies and TV Shows" dataset, taken from Kaggle. The dataset consists of 7,787 instances - some of the titles available on Netflix in 2019 - and 12 features. The number of records is sufficient to implement different models and it contains a vast proportion of the total Netflix library. We believe this dataset is relevant to the current situation whilst also being a joint interest of each member of the group.

The dataset can be found at the following link:
https://www.kaggle.com/shivamb/netflix-shows

Plan

The aim of the project is to find a model that most accurately predicts the probability of a title being related to a number of genres based on its description. It is likely that some genres in the dataset do not have enough titles associated with them so we will limit the scope by grouping together or excluding such genres from our models.

We will conduct some exploratory data analysis and clean the dataset as a group. This will ensure that our further experiments are based on a common input. We might also build a simple Linear Discriminant Analysis (LDA) model together in order to set a baseline for the evaluation of individual tasks. Then, we will work individually to experiment with various pre-processing techniques and build a number of models before coming back together to compare our findings, choose the best solution and deploy it. Each of us will prioritise one chosen algorithm but will attempt other algorithms depending on time constraints.

Development Environment

We will be using Google Colab as a development environment. Google Colab is a cloud version of Jupyter notebooks. It will allow us to see any code written by other group members in real-time. We have also created a GitHub repository, which will help us track any major changes between iterations. This should lead to an organised and streamlined process.

We will make use of a number of Python libraries, namely:
NumPy, Matplotlib, NLTK, Scikit-learn, Spacy and Gensim, PyTorch, TensorFlow. We will also make use of other common Python libraries, such as collections and re.

Individual Tasks

As discussed previously, we will go through the initial stages, such as data exploration and some basic pre-processing, together. Then, each of us will conduct individual experimentation. Each group member will try additional pre-processing techniques, different term frequency options and algorithms. Danil will focus on different Naive Bayes algorithms, such as Gaussian and Multinomial. Prakash will focus on Decision Trees and Jamie will experiment with Deep Learning methods. We may attempt additional algorithms, depending on the time constraints. We will also try tuning hyperparameters individually. Throughout the project, we will be holding regular weekly meetings to exchange our findings, as well as use a number of communication channels to continually track our progress.

[1] Shivam Bansal, "Netflix Movies and TV Shows", https://www.kaggle.com/shivamb/netflix-shows