European University of Lefke

Faculty of Engineering

Graduation Project 1

PREDICTING MOVIE GENRE USING PLOT SUMMARY

KELVIN GWARI

154420

ABSTRACT

The project uses the supervised text classification tools and machine learning methods to predict movie genres based on their movie descriptions. The movies and their corresponding data which is the plot summaries, the titles and genres will be taken from TMDB with over 10 000 movie data. As a movie can belong to several genres, this project will use the multi label classification. The main aim is to process natural languages, therefore only the movie descriptions will be used.

Supervisor

Asst. Prof. Dr. Vesile Evrim

May 2019

1. Introduction

1.1. Problem Definition

In a world where everything is being digitalized and the Internet of Things (IoT) is taking inroads to our day to day activities, it would be nice to have the entertainment industry to follow suit as well. With the invention of the internet, everything is now found online from even grocery shopping. Therefore since movies are all over the internet and they are attracting a large portion of the population, it is with great importance that the information found about a movie which is mostly the plot summaries can be used to predict the genre of a movie. This will be achieved by text classification and machine learning techniques. The movie information will be taken from The Movie Database (TMDB) through a corpus [1] and will also be stored into our own genre database for ease of access and to speed up the process. The data will be placed into a file with each file holding one movie occurrence. The data will be processed and will provide a decision boundary that will decide whether the genre of a movie is correct or incorrect. The text processing will be done in a supervised learning way where a group of words will be put in sets and the output produced and grouped according to the sets. The plot summaries are mostly made by users in the Internet Movie Database (IMDB). Sometimes the plot summaries won't be exactly matching with the genres of the movie. Therefore the system will give a better guidance for better plot summaries which will be matching with the genres.

1.2. Literature survey

Online movies are often registered with their genres and plot summaries. These movie genres are often inconsistent with the real story line of the movie. This is because the classification of these genres is done manually and involves the collection of users suggestions sent to Internet Movie Database. In the past only a handful of people have done this automation process of the movie genre prediction using plot summaries.

Ka-Wing Ho investigated different ways to classify movie genres by synopsis [8]. Since a movie can belong to multiple genres, [8] used the SVM method to do the multi label classification. One group of movies that belong to one genre will be the positive sample and the rest will be the negative sample and then train the classifier with the two disjoint sets. Multi-label K-nearest neighbor was also used. In this method, the knowledge of movies that belong to the same genre should share common keywords was used. If one were to consider a movie

synopsis y as a point in the hyperspace, movies that have similar genres combination of y will be close to it.

Mo Velayati [9] published a paper where he also focused on the movie genre classification. A Naïve Bayes classifier with multinomial model was used to achieve this task. In order to extract features the plot summary was converted into a “bag of words” where individual words were the features. Words were assigned term frequencies the number of times the term occurred in the entire corpus and also document frequency that showed how many documents have the word in their plot text. They then train the system, the purpose of the training was basically to calculate the threshold value and use it to decide whether a genre belonged with the movie or not.

Alex Blackstock and Matt Spitz investigated Classifying Movie Scripts by Genre with a MEMM Using NLP-Based Features [10]. Two evaluation metrics to analyze the performance of two separate classifiers, Naïve Bayes Classifier and a Maximum Entropy Markov Model Classifier were devised. During the investigation they faced challenge of inconsistent format of movie scripts and multiway classification presented by the fact that each movie script they had was labelled with several genres. To solve this challenge I will use the multi-label classifier [3].

1.3. Goals

The main goal is to identify movie genres using machine learning tools. This will give users a more precise and more accurate movie genre only from the plot summary of a movie. In this project I will be trying to process natural language from the plot summaries. The processing will be achieved by eliminating some stop words such as prepositions. Movies can have twists in them which will make them to belong to different genres at once. To solve this problem i will be using supervised learning. Each set will have instances associated with a set of labels and the task will be to predict a set of unseen occurrences through analyzing training set with known label sets.[3](*deep dive into multi-label classification*). Results show that there are methods of tackling this multi label problem, using algorithm adaptation methods or problem transformation methods. During the course of the project I will learn how to use python language and various machine learning techniques.

2. Resources

2.1. Required Hardware

Computer

A computer will be an ideal hardware to host the software required by the project such as the IDEs and the user interfaces to the database management systems and the system at large.

The suitable computer requirement is at least a **2GB RAM** machine.

2.2. Required Software

Python language

It is a language that is mostly used for its simple syntax and dynamic nature. This language is an open source language and is supported by a lot of resources and high quality documentation. [4]. The language is very readable by non-developers and with its syntax, developers will not worry much about the programing language but rather they will concentrate more on the problem to be solved [5]. I chose this programming language because it has libraries that are best suited with machine learning and the data science branch as a whole.

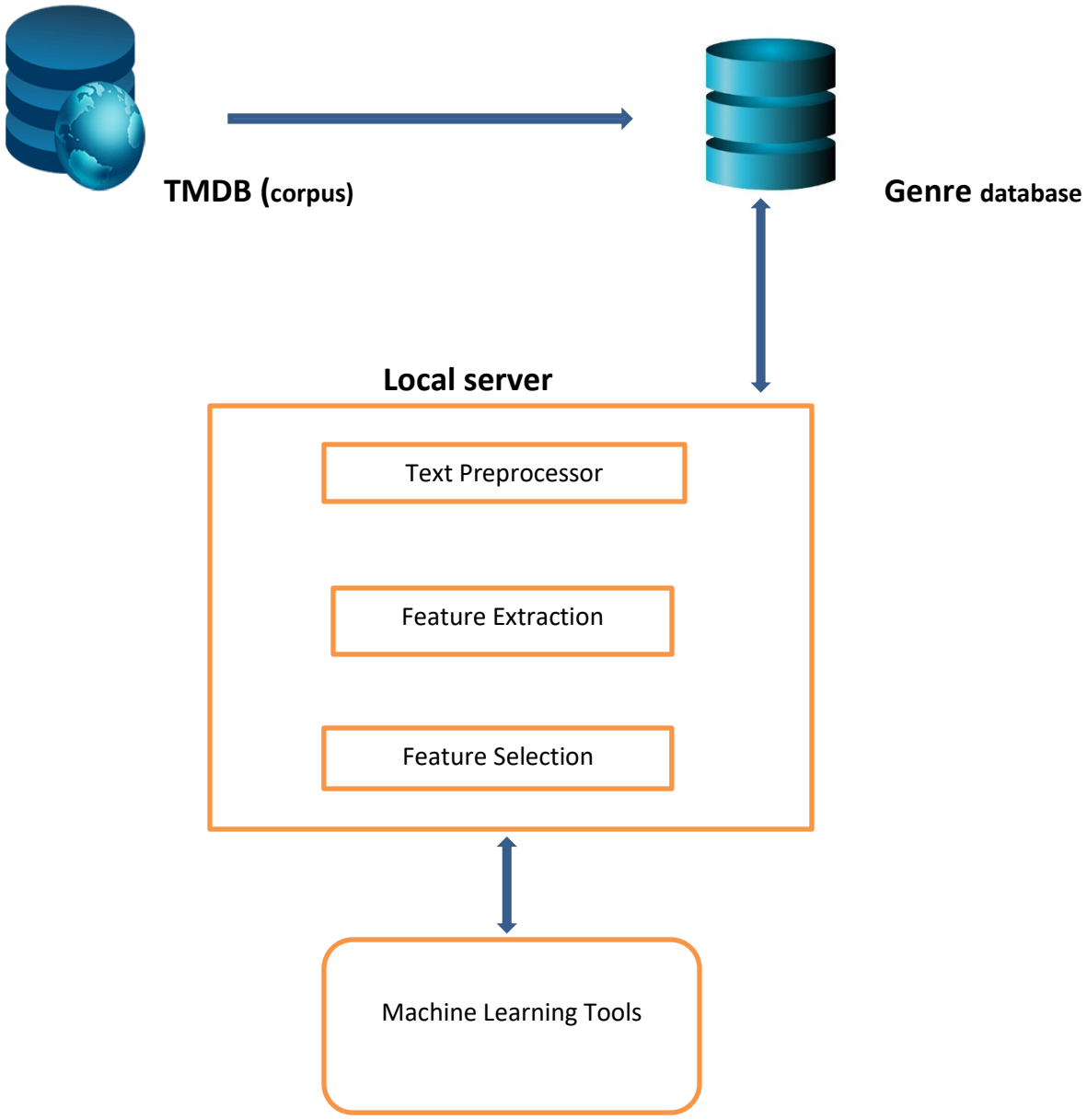
MySQL

This is a database management system used to store data to be used in the long run. MySQL is well known for its data security and its reliability in performance and so on. [6]

PyCharm

PyCharm is a python IDE that is used for machine learning and AI development. [7]

3. Framework


TMDB (corpus)
Genre database
Local server
Text Preprocessor
Feature Extraction
Feature Selection
Machine Learning Tools

Module

3.1. Corpus

The data for the corpus will be collected from [1] and will be comprising of the movie titles, plot summaries, genres, year etc. i will be using the data from this database especially the plot summaries, titles, year and also genres. The genres will be used to compare our results from our system with the actual results that the movie is said to belong with.

This module will be implemented by Kelvin Gwari. The module will take approximately 3 weeks to complete.

3.2. Genre Database Module

After getting the data from the corpus i will store this data in our own database which i will name genre database. In this module, the data will be accessed on a local machine which will make the natural language processing faster. The data will be in relevant tables.

The module will be implemented by Kelvin Gwari who will design the database and populate it with the data from the corpus. The implementation will take 2 weeks.

3.3. Text Processors module

This module is responsible for processing the text from the plot summaries. The processing involves the elimination of stop words from the text to achieve a cleaner text for further processing stages.

The module will take 1 month to complete.

3.4. Feature Extraction module

This module is for creating a new set of features that captures most of the useful information that includes the keywords by combination of the existing features using various machine learning algorithms.

The module will take 2 months to complete and Kelvin Gwari will be responsible or the implementation.

3.5. Feature Selection module

This module is responsible for dimensionality reduction, especially when dealing with a lot of features (words) there is need of dimensionality reduction. This will be achieved by selecting a subset of the existing features without a transformation.

This module will take 1 month to implement.

3.6. Machine Learning Tools

After the processing, extraction and the selection modules are done the features will be placed into a vector and the Naïve Bayes and other machine learning tools will be use to classify the data.

This module will take 2 weeks to implement. Developer will be Kelvin Gwari.

3.7. The user interface module

This is where one will input data, this data will be the plot summaries of a movie. This will be processed using the different modules mentioned above in sections 3.3, 3.4 3.5 and 3.6. The genres that will be the output will be produced on the same interface.

3.8. Human Resources module

Developer: Kelvin Gwari will be responsible for the implementation of all the modules during the course of the project.

4. Risk Analysis

Risk Summary	**Overloading of data**
Risk category	**System risk**
Probability	**medium**
Impact (0-5)	**3**
Description	**A system may crash when a lot of data have to be processed with a machine that has limited computational power.**
Risk Mitigation	**Set a maximum quota for the data to be processed.**

Risk Summary	**Inaccessible movie data**
Risk category	**System risk**
Probability	**low**
Impact (0-5)	**5**
Description	**Taking longer periods to access data or failing to acquire data from an external database.**
Risk Mitigation	**Build our own database in our local machine.**

Risk Summary	**Additional of genres in the movie industry**
Risk category	**Product risk**
Probability	**low**
Impact (0-5)	**4**
Description	**When a new genre will be added in the movie industry and our system wasn't trained to recognize it.**
Risk Mitigation	**An update of the genre list that will cover both the old and the new movies**

5. Conclusion

5.1. Benefits

Predicting movie genres based on plot summaries is a project that will see many movie lovers in the world stop wasting time looking for the genres while watching the movies but rather they will use the system to decide which genre they want using the plot summaries provided. This will also enable the people who set genres to set accurate and precise movie genres for the viewers. Since I will be using python, a language I am starting to learn to write such big programs, it will help me intensify my knowledge on the language. My knowledge of algorithms

will be taken to a whole new level, this is because during the course of the project, I will be using algorithms of different types and testing their time complexity on different levels of development. The natural language processing field requires one to have an extensive knowledge of machine learning techniques which I will also acquire during the course of the project. Many organizations in the world are now using MySQL as their preferred database management system. With this knowledge, I decided to prepare myself for the future by just using MySQL database on this project.

5.2. Ethics

With the diversity of movies currently present in the world of entertainment. It is of great importance to present the ethics and morals of how the system will work to protect and respect cultures and different religions. Any malice use of the system to violet the safety and health of the public, the developer i.e. Kelvin Gwari, will not be held responsible for the action. The system will and must be used with human culture and conservation of the human rights in mind. It will not be implemented to get movies from the adult entertainment industry ever. The sets of the data used in the supervised training will not include any genre that is harmful or inappropriate.

5.3. Future Works

The movie genre prediction based on plot summary system can be expanded in the future by adding the movie success predictor. The success of the movie can be predicted using information from the Internet Movie Database (IMDB). This data include the casting crew, the director, main actors in the movie etc. We can also add the rating system which will rate movies based on the user comments. This will take the comments from users of the same IMDB and based on what they will be commenting whether negative or positive will then be used with natural language processing and provide the ratings of the movie.

6. References

[1]: IMDB data. ftp://ftp.fu-berlin.de/pub/misc/movies/database/frozendata/

[2]: The Movie Database (TMDB). Retrieved from (TMDB API) https://www.themoviedb.org/documentation/api /

[3]: Dive into multi-label classification..! Kartik Nooney (June 7, 2018) https://towardsdatascience.com/journey-to-the-center-of-multi-label-classification-384c40229bff/

[4]: Why Is Python So Good for AI, Machine Learning and Deep Learning? Jakub Protasiewicz (Aug 31, 2018) https://www.netguru.com/blog/why-is-python-so-good-for-ai-machine-learning-and-deep-learning/

[5]: Python Programming Language official page. https://www.python.org/

[6]: MySQL Database Management System. https://dev.mysql.com/

[7]: PyCharm Ide. https://www.jetbrains.com/

[8]: Movies genres classification by synopsis. Ko-Wing Ho, 2011.

[9]: Movie genre classification. Mo Velayati, 2017. https://mvelayati.com/2017/07/19/movie-genre-classification/

[10]: Classifying Movie Scripts by Genre with a MEMM Using NLP-Based Features. Alex Blackstock Matt Spitz, 2008.