Introduction to data science project

Group A03, QUICKGENRE

Davit Rizhinashvili, Ana Shahpazir

Github link

Business understanding

Background

The entertainment industry, particularly the film sector, has seen exponential growth in content creation and consumption. With a vast array of movies available, efficiently categorizing and recommending films based on genres has become crucial for streaming platforms, distributors, and even production houses.

Business Goals

The primary goal is to make a program that would predict genre from movie title alone, do it fast and accurately.

Secondly, to enhance content discoverability and recommendation accuracy by implementing this model. This will aid in better cataloging of movies and improve user experience in content platforms.

Business Success Criteria

Success will be measured by the accuracy and efficiency of the genre prediction model. An accuracy rate of over 85% in genre prediction will be considered successful. Additionally, the model's ability to integrate seamlessly with existing content management systems will also be a key success factor.

Inventory of Resources

- Dataset: A comprehensive dataset of movie titles and their associated genres. This dataset also contains other metadata for movies, like description, keywords and length, which will not be used for analysis
- Computing Resources: Adequate hardware and software resources for model development and deployment. Particularly, Nvidia CUDA enabled laptop GPU.
- Human Resources: Skilled and passionate students proficient in machine learning and NLP.

Requirements, Assumptions, and Constraints

- The model assumes that movie titles have a significant correlation with their genres. The model also assumes that there will be no bias towards any particular genre.
- There is a constraint in handling titles that are less descriptive or ambiguous.

Risks and Contingencies

- Risk of model overfitting or underperforming with novel or unique titles.
- Risk of model not recognizing one word titles with double meaning.
- A contingency plan involves iterative model refinement and incorporating additional contextual data (like small description) if needed.

Terminology

Key terms include genres, neural networks, NLP, text embeddings, and accuracy.

Costs and Benefits

- Costs: Resource allocation for model development, computational costs, and maintenance. In this case, cost was only time as laptop with Nvidia 3060 was sufficient for the task.
- Benefits: Improved content categorization and user satisfaction, potential increase in platform engagement and retention rates. Faster times of genre lookup just by title.

Data-Mining Goals

- To isolate unique genres.
- To use a neural network model that converts movie titles into latent space vectors and predicts their genres.
- To process and utilize movie title latent data effectively for genre prediction.
- Being able to do multi class classification, meaning predicting multiple genres at once.

Data-Mining Success Criteria

- Achieving an accuracy rate of 85% or above in genre prediction on testing set.
- Demonstrating the model's ability to handle a diverse range of movie titles effectively.

Data understanding

Gathering Data

Outline Data Requirements

- **Primary Data**: Movie titles and their corresponding genres.
- **Format**: Structured data, ideally in a tabular format like CSV.
- **Volume**: A sizable dataset encompassing a diverse range of movies across different periods and cultures to ensure model robustness.

Verify Data Availability

- Sources such as movie databases (e.g., IMDb, TMDB) are identified as potential data providers. Although these would require methods like web scraping to collect data, which for some websites, due to security reasons is prohibited.
- Public datasets available on platforms like Kaggle could also be considered.

Define Selection Criteria

- **Recency**: Movies covering a broad timeline, including recent releases, to ensure model relevance.
- **Diversity**: Titles across various languages and cultures to enhance model's generalization capability.
- **Completeness**: Preference for datasets with comprehensive and maybe multiple genre labels and minimal missing data.

Describing Data

We chose to use movies dataset from kaggle which contains over 40000 entries.

Upon initial inspection, the dataset comprises two primary fields relevant to the project:

- **Original Title**: The movie title, a string variable.

- **Genres**: A JSON array containing genre details, where each genre is represented by an ID and name. example:
  [{'id': 18, 'name': 'Drama'}, {'id': 27, 'name': 'Horror'}]

The dataset's size and diversity (in terms of movie release years, origin countries, and language) are noted, providing an initial understanding of its scope and coverage.

Exploring Data

Initial exploration involves statistically summarizing the dataset and visualizing key aspects:

- **Distribution of Genres**: Analyzing the frequency of each genre to identify any imbalance.
- **Title Analysis**: Observing the length and structure of movie titles, understanding if certain words or patterns are more common in specific genres.
- **Temporal Trends**: Examining if certain genres have become more prevalent over time.

This phase also includes a preliminary assessment of relationships between movie titles and genres, seeking patterns or anomalies that could influence model development.

Verifying Data Quality

- **Completeness**: Checking for missing values in the 'original_title' and 'genres' fields.
- **Consistency**: Ensuring genre labels are consistent and accurately reflect the movie content.
- **Validity**: Verifying that the movie titles are correctly spelled and correspond to actual movies.
- **Uniqueness**: Ensuring there are no duplicate entries of the same movie.

Any issues identified in this phase will be used for cleaning during the data preparation stage.

Project Plan and Task List

1. **Data Collection (8 hours)**
    - Team Member Contribution: Done together.
    - Tasks: Identify and acquire datasets from sources like IMDb, TMDB, or Kaggle.
    - Tools: Web browsers, data downloading tools, APIs if available.
    - Comment: Ensure datasets have a diverse range of movies for robust model training.
2. **Data Preprocessing and Cleaning (16 hours)**
    - Team Member Contribution: 8 hours each from two team members.
    - Tasks: Clean and format data, handle missing values, parse JSON genre data.
    - Tools: Python, Pandas, JSON parsing libraries.
    - Comment: Critical for model accuracy, requires careful handling of genre labels.
3. **Feature Extraction (10 hours)**
    - Team Member Contribution: Done together.
    - Tasks: Convert movie titles into embeddings using NLP models.
    - Tools: Python, Hugging Face Transformers, PyTorch/TensorFlow.
    - Comment: The quality of embeddings is key to successful genre prediction.
4. **Model Development and Training (20 hours)**
    - Team Member Contribution: 10 hours each from two team members.
    - Tasks: Design and train the neural network.
    - Tools: Python, PyTorch/TensorFlow, scikit-learn.
    - Comment: Iterative process; may require adjustments in architecture and parameters.
5. **Evaluation and Optimization (12 hours)**
    - Team Member Done together.
    - Tasks: Test the model, evaluate performance, and fine-tune.
    - Tools: Python, relevant machine learning libraries.
    - Comment: Focus on achieving high accuracy and resolving any overfitting or bias issues.
6. **Documentation and Reporting (6 hours)**
    - Team Member Contribution: 3 hours each from two team members.
    - Tasks: Document the methodology, results, and insights.
    - Tools: Word, Markdown editors, Jupyter Notebooks.

Methods and Tools

- **Methods**: Machine Learning, Natural Language Processing, Data Preprocessing, Model Evaluation.
- **Tools**: Python programming language, Pandas for data manipulation, Hugging Face Transformers for NLP embeddings, PyTorch or TensorFlow for neural network development, and various libraries for statistical analysis.