Classificação de Gêneros de Filmes Baseado na Sinopse

INFORMAÇÕES DO GRUPO

Daniele Hidalgo Boscolo	7986625
Eduardo Sigrist Ciciliato	7986642
Guilherme Vicentin Nardari	7986521
Hiero Martinelli	7986646
Willian Fagner Policiano	8066190

INFORMAÇÕES DA DISCIPLINA

Nome Disciplina: SCC530-Inteligência Artificial
Professora: Solange Oliveira Rezende
Estagiária PAE: Camila Vaccari Sundermann

IDENTIFICAÇÃO DO PROBLEMA

Coleção de Dados

A coleção de dados utilizada no experimento foi extraída da base de dados do **IMDb** (Internet Movies Database), que é uma base aberta que conta com uma boa parte dos filmes já lançados registrados.

Este experimento será baseado em **classificação** de **gêneros** de filmes utilizando a **sinopse** do mesmo, portanto, foi extraído uma coleção de registros contendo Sinopse, Gênero e ID do filme, no período do ano de **1998 até 2013** obtendo assim **12500** registros (tuplas) de dados que possuem os atributos gênero e sinopse para serem divididos entre conjunto de treinamento e conjunto de teste.

Problema

Foi realizado a classificação de filmes de modo automático, para facilitar classificar as novas entradas de filmes na IMDb nos gêneros mais clássicos como por exemplo ação, aventura, horror, comédia, etc.

Com uma classificação automática através da sinopse dos filmes, é necessário apenas inserir os gêneros mais complexos nos novos filmes, diminuindo a carga de trabalho ao editar a base de dados.

PRÉ-PROCESSAMENTO E EXTRAÇÃO DE PADRÕES

Pré-Processamento

Após a obtenção dos dados, foi utilizado um script em **Python** para converter essas informações para um banco de dados **Mysql**.

Como o problema baseia-se em mineração de texto, foi preciso primeiro fazer a seleção das tuplas a serem utilizadas e depois usar um script em **Shell** para convertê-las em arquivos de texto, esses arquivos foram utilizados pelo **TPT** para remover as **stopwords**, normalizar, calcular o valor-atributo e criar um arquivo **ARFF** com a **bag-of-words**.

Extração de Padrões

A extração de padrões foi realizada pelo Weka, à partir da **bag-of-words** gerada pelo software TPT.

Os padrões encontrados foram descobertos utilizando-se o algoritmo **Naive Bayes Multinomial**, pois foi o que teve melhor porcentagem de acerto.

PÓS-PROCESSAMENTO

Avaliação

Através da ferramenta ***Experimenter*** do Weka foi comparado com 4 iterações os algoritmos **Naive Bayes Multinomial** e **Naive Bayes.**

O resultado desse experimento mostra que o primeiro utiliza menos **CPU**, demora menos tempo para processar e retorna **10% de acerto** a mais do que o segundo para este conjunto de dados.

Resultado e Conclusões

Utilizando a sinopse de 12.500 filmes a porcentagem de acerto foi de 49.8%

Gênero com menor conflito: Documentário.
Gêneros com maiores conflitos: Ação com Aventura, Ação com Sci-Fi, Comédia com Romance e Drama com Romance.

Assim, foi possível ver como os gêneros interagem entre si (exemplo: romance pode tanto ser uma comédia romântica, como um drama) e ao mesmo tempo como é complicado a classificação de texto que obteve uma taxa de acerto relativamente baixa.