Mineração de Dados

Prof. Rodrigo Tripodi Calumby - rtcalumby@uefs.br

Desafio Experimental 1 - Classificadores

Tema: Construíndo Preditores do Sucesso Comercial de Filmes

Contexto do Projeto

Grandes estúdios de cinema e produtoras realizam investimentos bastante altos no desenvolvimento de filmes. O sucesso comercial destes produtos dependem de múltiplos fatores. Como membros de uma equipe de Ciência de Dados, vocês foram selecionados para trabalhar no desenvolvimento de modelos de predição de sucesso, considerando múltiplas informações disponíveis sobre filmes já produzidos. Com estas informações será possível estimar o sucesso comercial e popularidade sobre novos filmes. A base de dados a ser utilizada neste projeto contém atributos coletados sobre cerca de 5000 filmes. Estes dados foram obtidos do IMDb que é uma base de dados online sobre filmes, séries de TV, atores, etc. Os dados disponíveis sobre os filmes são apresentados no Anexo I e podem ser baixados na página da disciplina. É importante destacar que algumas informações não puderam ser coletadas para todos os filmes e estão ausentes.

Objetivo Geral

A equipe deverá desenvolver um sistema de predição de indicadores que ajudem as produtoras de filmes a decidir sobre realizar ou não um projeto, considerando seu possível retorno comercial e índices de popularidade. Os indicadores a serem preditos são:

1) Número de curtidas na página do filme do facebook;
2) Número de avaliações por usuários;
3) Número de revisões críticas;
4) Número de votos;
5) Arrecadação;
6) Escore IMDb;

Objetivos Específicos

A equipe deve realizar todo o processo de pré-processamento dos dados para torná-los adequados ao processo de mineração (limpeza, tratamento de dados ausentes, seleção de atributos, etc.). A partir daí, as seguintes atividades devem ser realizadas:

1. Análise descritiva;
2. Escolha dos algoritmos/classificadores a serem avaliados;
3. Definição da metodologia experimental de validação (configurações dos algoritmos, protocolos experimentais, medidas de eficácia, análise de resultados, etc.);
4. Discussão acerca dos resultados obtidos.

Objetivo Extra (opcional)

Como atividade extra, poderá ser realizada a integração de novos dados que possam contribuir para o aumento da eficácia nas predições ou que possam ser preditos a partir dos dados disponíveis. Como sugestão, podem ser coletados mais dados sobre os protagonistas, por exemplo, estatísticas gerais sobre filmes estrelados, quantidade de seguidores no Instagram, número de fotos postadas, número médio de comentários nas fotos, idade, nacionalidade, etc. Novos dados podem também ser derivados dos dados já disponíveis. Haverá pontuação bônus para este objetivo extra de acordo com a qualidade das contribuições apresentadas.

Produtos

A equipe deverá entregar todo os códigos, *scripts*, *worflows*, dados, etc. que forem utilizados/desenvolvidos no projeto, Além disso, a equipe deverá apresentar um relatório, incluindo pelo menos:

- Introdução
- Metodologia
- Resultados e Discussão
- Conclusões
- Referências

Os materiais utilizados/desenvolvidos devem ser entregues por e-mail em arquivo compactado. Este material deve estar organizado em subpastas e com nomes significativos, por exemplo, "../dados/instagram.csv", "../scripts/normalizaçãoAtributoX.sh", "../workflows/experimento1.ows", etc. O relatório deve ser entregue no mesmo e-mail em formato PDF (nome do arquivo: "relatórioEquipeX.pdf"). O relatório deve ser produzido em formato de artigo científico, seguindo o modelo de formatação *IEEE Transactions*. Utilize preferencialmente o *template* LaTeX. O relatório deve ter no máximo 6 páginas (incluindo imagens e referências). O uso de páginas adicionais podem ser solicitadas ao professor com a devida justificativa.

Os produtos devem ser enviados para o endereço rtcalumby(arroba)ecomp(ponto)uefs(ponto)br até as 08h do dia 12/06/18. O assunto do e-mail deve ser: [Mineração 2018.1] Entrega Projeto Prático 1 - Equipe X.

Critérios Gerais de Avaliação

- Atendimento ao prazo de entrega estabelecido.
- Completude do material entregue.
- Atendimento aos objetivos, requisitos e orientações desta proposta.
- Adequação do protocolo experimental aplicado aos dados utilizados.
- Amplitude dos experimentos realizados.
- Rigor científico dos experimentos realizados e resultados apresentados.
- Qualidade da discussão dos resultados e conclusões.
- Qualidade da apresentação visual do relatório.
- Atendimento à formatação indicada.
- Qualidade do texto do relatório:
  - Clareza e objetividade.
  - Atendimento às normas de ortografia e gramática.
  - Adequação das citações.
  - Organização das informações.

Observações

- No relatório, só deve ser feita descrição de métodos não triviais/comuns e que tenham sido utilizados no desenvolvimento do projeto. Não devem ser feita fundamentação teórica daquilo que já foi visto em sala de aula.
- A critério do professor, alunos poderão ser convocados para responder a questionamentos sobre o trabalho desenvolvido.
- Fraudes serão penalizadas com nota 0.0 (zero) na atividade.
- Entregas com atraso de até 24h serão penalizadas em 2 (dois) pontos. Após 24h, entregas não serão mais aceitas, a menos que seja comprovado motivo de força maior que comprometa toda a equipe.

Anexo I - Dicionário de Dados

Variable Name	Description
movie_title	Title of the Movie
duration	Duration in minutes
director_name	Name of the Director of the Movie
director_facebook_likes	Number of likes of the Director on his Facebook Page
actor_1_name	Primary actor starring in the movie
actor_1_facebook_likes	Number of likes of the Actor_1 on his/her Facebook Page
actor_2_name	Other actor starring in the movie
actor_2_facebook_likes	Number of likes of the Actor_2 on his/her Facebook Page
actor_3_name	Other actor starring in the movie
actor_3_facebook_likes	Number of likes of the Actor_3 on his/her Facebook Page
num_user_for_reviews	Number of users who gave a review
num_critic_for_reviews	Number of critical reviews on imdb
num_voted_users	Number of people who voted for the movie
cast_total_facebook_likes	Total number of facebook likes of the entire cast of the movie
movie_facebook_likes	Number of Facebook likes in the movie page
plot_keywords	Keywords describing the movie plot
facenumber_in_poster	Number of the actor who featured in the movie poster
color	Film colorization. 'Black and White' or 'Color'
genres	Film categorization like 'Animation', 'Comedy', 'Romance', 'Horror', 'Sci-Fi', 'Action', 'Family'
title_year	The year in which the movie is released (1916:2016)
language	English, Arabic, Chinese, French, German, Danish, Italian, Japanese etc
country	Country where the movie is produced
content_rating	Content rating of the movie
aspect_ratio	Aspect ratio the movie was made in
movie_imdb_link	IMDB link of the movie
gross	Gross earnings of the movie in Dollars
budget	Budget of the movie in Dollars
imdb_score	IMDB Score of the movie on IMDB