Classificação de Imagens por Similaridade

Ingrid Rosselis Sant Ana da Cunha, Pedro Olmo Stancioli Vaz De Melo
{irscunha,olmo}@dcc.ufmg.br

Motivação

- Redes sociais:
  - Tag automático de pessoas em fotos;
  - Organização automática de imagens em álbuns;
  - Extração de insights sociais em imagens públicas.
- Acessibilidade para pessoas com deficiência visual.
- Categorização de imagens.
- Clusterização de imagens similares.
- Análise de imagens médicas.
- Análise e filtro de conteúdo.

Objetivos

- Obter imagens de pôsteres de filmes, séries e animações, bem como suas principais informações.
- Tratar dados.
- Fazer análise exploratória dos dados obtidos.
- Estudar algoritmos e técnicas para classificar e clusterizar imagens.

Dados

- Web scraping em Python.
- Bases separadas por fonte (sites IMDB e MyAnimeList).
- 9 categorias selecionadas: comédia, ação, aventura, terror, romance, sci-fi, drama, mistério, fantasia.
- Extraídos 1000 instâncias por categoria.
- Exclusão de duplicatas como processo pós extração e divido por base.
- IMDB: aproximadamente 5000 instâncias.
- MyAnimeList: aproximadamente 4000 instâncias.

Trabalhos Futuros

- Aplicar algoritmos sobre os dados obtidos para encontrar relações interessantes e descobrir padrões em pôsteres.
- Tipos de algoritmos que podem ser utilizados:
  - Clusterização;
  - Classificação;
  - Extração de features.

Conclusão

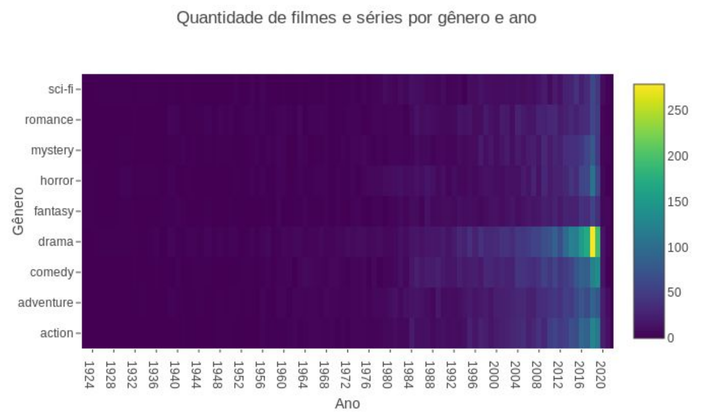
- Redes neurais convolucionais são mais recomendadas para classificação de imagens.
- Classificação de imagens por rótulos (gêneros, por exemplo) é um processo amplamente difundido atualmente.
- Técnicas para extrações de cores dominantes podem gerar correlações interessantes entre palhetas de cores e outro atributos das imagens, como gênero e ano.

Visualizações iterativas pode ser encontradas em:
https://irscunha.github.io/poc1/

Ou pelo QR code:

Análises Exploratória da Base IMDB

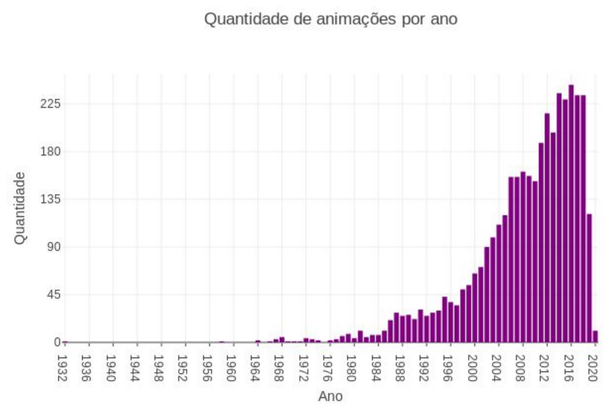
- Pico em 2018:
  - 495 filmes e séries.
- Produção parece mais esparsa, comparando com MyAnimeList:
  - Custo geralmente mais alto que o de uma animação.



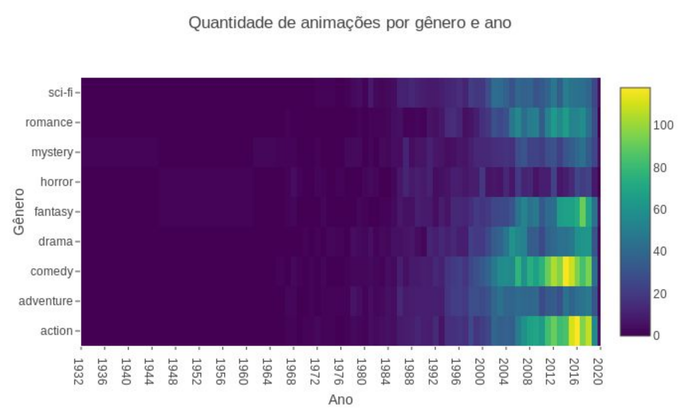
- Dos dados coletados, 2018 teve 495 instâncias e, destas, 279 continham o gênero drama.

- Pode-se notar pelo Wordcloud que o número de filmes e séries que possuem continuação é menor em comparação com as animações.

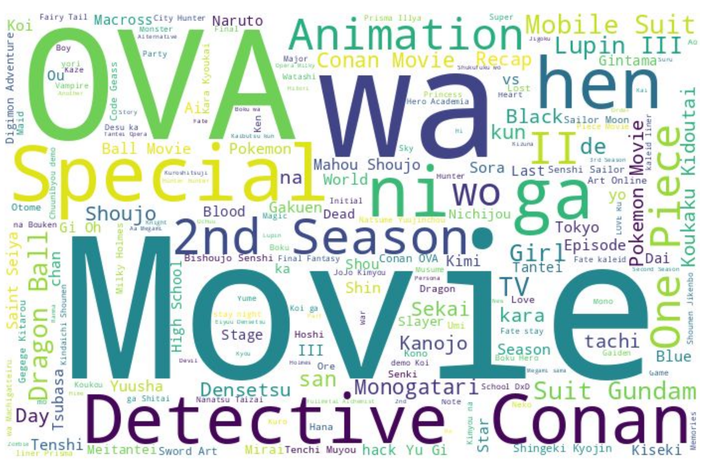
Análises Exploratória da Base MyAnimeList



- Produção de animações (em forma de filmes ou séries) é muito mais uniforme que a vista em filmes, como mostrado no primeiro gráfico.



- Gêneros de ação, aventura e comédia são os destaques em número de animações feitas



- Muitas partículas da língua japonesa, como “wa”, “ni” e “ga”.
- Divisão explícita entre OVA (“original video animation”, episódios especiais em DVDs), filmes, episódios especiais e segunda temporada.

Avalie-me!

40 ANOS 1976-2016
UFMG | ICEx
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO
UFMG