Pontifícia Universidade Católica de Minas Gerais
Pós-graduação em Ciência de dados e Big Data

Projeto Final

Processamento de linguagem natural

Aluna: Daniela Caroline Lucas dos Santos

Prof(a): Bárbara Silveira Fraga

29 de Junho de 2020

Introdução

A língua natural, ou seja, como é falada e escrita não pode ser processada pelo computador, pois como já se sabe esse entende apenas bits que são 0 e 1. Portanto, para que seja possível o processamento, existe um subárea da ciência da computação e Inteligência Artificial conhecida como Processamento de Linguagem Natural (PNL).

O Processamento de Linguagem Natural permite que por meio de um pré processamento do texto, aprendizado de máquina e modelos estatísticos as sentenças possam ser lidas e avaliadas por meio de processamento computacional.

Existem diversas aplicações para o processamento de linguagem natural como, por exemplo, sumarização automática, análise do Discurso, reconhecimento de entidade nomeada (NER), respostas a perguntas, recuperação de informação (IR), extração de informação (IE), recomendação, entre outras. Hoje, devido ao volume de dados produzidos de forma não estruturada essas aplicações têm ficado cada vez mais relevantes para se fazer análise efetivas.

Objetivos

Neste trabalho foi escolhido fazer um tratamento de uma base de filmes, a qual contém a seguinte estrutura:

- IMDB_movie_details

  Nessa tabela é possível encontrar apenas um registro para cada filme registrado na base.

movie_id	ID do filme de acordo com o site IMDB
plot_sumary	Resumo do filme
duration	Duração
genre	Genero (pode conter mais de um)
rating	Média com as notas dadas na review 0 -10
plot_synopses	Sinopse do Filme

- IMDB_reviews

  Nessa tabela é possível encontrar várias reviews direcionadas a um mesmos filme.

review_date	Data em que a review foi feita
movie_id	ID do filme de acordo com o site IMDB
user_id	ID do usuário registrado no site
is_spoiler	Booleano que indica se a review contém spoiler ou não.
review_text	Texto com a review
rating	Nota dada em cada review 0-10
review_sumary	Resumo da review

- IMDB_movies

  Nessa tabela é possível encontrar o registro dos filmes com seu Id e título.

imdb_title_id	ID do filme de acordo com o site IMDB
title	Título do filme
original_title	Título do filme original
year	Ano de lançamento

A partir desses dados foi levantado como objetivo:

- Clusterização - A partir do resumo e da sinopse no filme gerar uma embedding clusterizado usando TSNE e fazer comparações quanto a clusterização.
- Análise de sentimento - Fazer uma sobre os textos de reviews utilizando o algoritmo Vader e fazer a média dos scores destinado a um mesmo filme comparar o score gerado com o 'rating' da tabela IMDB_movie_details.
- Naive Bayes - Usar o algoritmo de classificação para treinar a base e posteriormente definir se aquela review possui spoiler ou não.

Metodologia

Inicialmente, as tabelas foram carregadas e limpas, a fim de retirar todas as linhas que tivessem alguma célula vazia, também foi carregada a tabela movies para ser possível identificar o nome do filme quando necessário.

Para fazer o Embedding também foi carregado o arquivo de embedding treinado da wikipédia e, posteriormente, gerado o word_vectors.

Abaixo serão descritos os métodos e técnicas aplicadas para se fazer a análise dos dados.

Clusterização

O primeiro passo foi clusterizar os dados de acordo com a sinopse e resumo dos filmes, para isso foram lidas as colunas 'plot_synopses' e 'plot_sumary', feito o devido pré-processamento para possibilitar a aplicação do embedding, o qual já era treinado para conseguirmos calcular a distância média entre as palavras de cada sentença da sinopse e do resumo.

Em seguida, foi aplicado o K-means e foram considerados um total de 8 cluster, o número de cluster foi escolhida baseada na possibilidade de clusterização de acordo com o gênero do filme.

Após clusterizado, foi plotado os gráficos usando TSNE para visualizar os clusteres formados. Além disso, foram gerados gráficos de barra para analisar se a clusterização baseadas nas sinopses e nos resumos dos filmes poderia ser usada para dividir os filmes em gênero para melhorar as recomendação de filmes.

Análise de sentimento

A análise de sentimento foi feita utilizando o algoritmo Vader , um algoritmo extremamente forte, capaz de considerar pontuações, letras maiúscula e emojis na sua avaliação, e seu retorno é um dicionário contendo as métricas mostrada abaixo:

Sentiment Metric	Score
Positive	0.674
Neutral	0.326
Negative	0.0
Compound	0.735

Sendo que,

- score >= 0.05 (Positivo)
- score >-0.05 e score <0.05 (Neutro)
- score <= -0.05 (negativo)

O Compound traduz as demais componentes e pode variar entre - 1 e 1.

Para fazer essa avaliação, usei apenas 1/10 da base (tabela IMDB_reviews), por se tratar de uma base muito grande o processamento ficou muito lento. Para esse método foi usada a coluna 'review_text', sem nenhum tipo de pré- processamento.

Posteriormente, a componente 'compound' retornada pela score do algoritmo foi recalculada, isto é, sua escala foi alterada de -1 a 1 para 0 a 10, essa alteração foi necessária para se poder comparar o resultado do algoritmo com o da própria tabela disponibilizado na coluna 'rating'.

Naive Bayes

O algoritmo "Naive Bayes" é um classificador probabilístico, e nesse caso foi utilizado para classificar se uma review possui spoiler ou não.

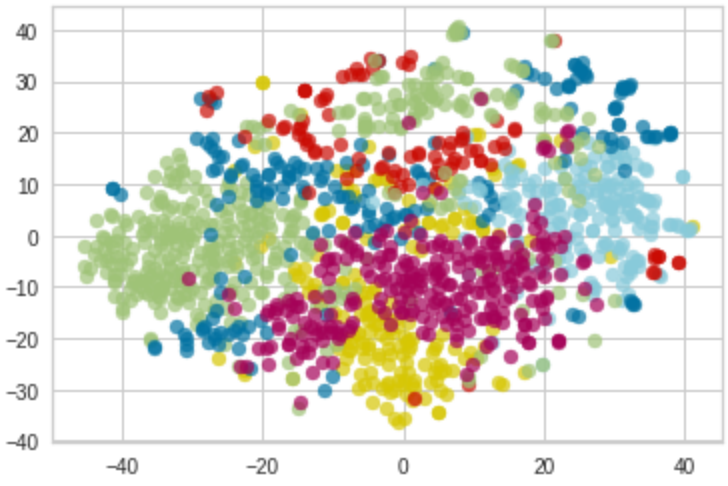
Nesse processo apenas 1/100 da base foi utilizada para viabilizar o processamento, o texto de review foi previamente pré-processado, e posteriormente foi utilizado o Bag of Words para vetorizar as sentenças criadas.

Com o vetor criado a base foi dividida e massa de treino e teste e algoritmo Naive Bayes aplicado, com o objetivo de verificar a eficácia desse algoritmo para essa função.

Análise de Resultados

Clusterização

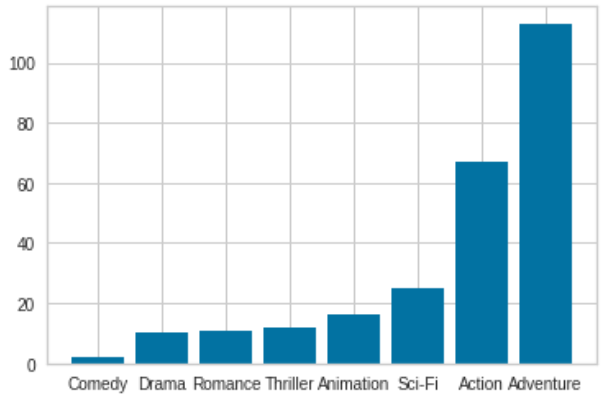
A primeira clusterização foi feita utilizando as sinopses, como pode-se ver os clusters ficaram relativamente bem definidos

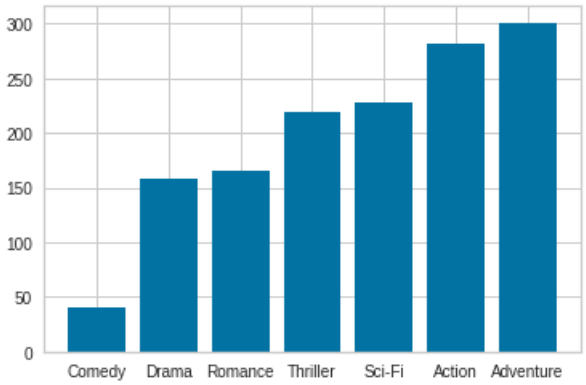


Clusters gerados pelo processamento das sinopses

Entretanto não se pode afirmar que é possível definir o gênero do filme a partir da análise da sua sinopse utilizando esse método, pois ao avaliarmos dois clusters vimos a

distribuição de gênero pode estar bem indefinida, e isso fica confirmado pelos outros gráficos relativos aos demais clusters contidos no 'ProjetoFinalPLN.ipynb'. Logo,é possível que a clusterização tenha se baseado em outros critérios.





Cluster 0 Cluster 4

Outro ponto importante a se levantar, é que um filme pode se enquadrar em mais de um gênero logo era esperado ao realmente que a divisão não fosse bem dividida.

Também foi feita outra abordagem, onde foram extraídos os tops 5 filme de um cluster, simulando uma recomendação.

```
_____________________________
['Star Wars: Episode II - Attack of the Clones']
_____________________________
_____________________________
['The Chronicles of Narnia: Prince Caspian']
_____________________________
_____________________________
['Rogue One']
_____________________________
_____________________________
['Pirates of the Caribbean: On Stranger Tides']
_____________________________
_____________________________
['X-Men']
_____________________________
_____________________________
['300']
_____________________________
_____________________________
['The Scorpion King']
_____________________________
```

O Cluster 0 tanto pelo gráfico quanto pelas recomendações observa-se uma certa coerência de gênero e estilo de filme.

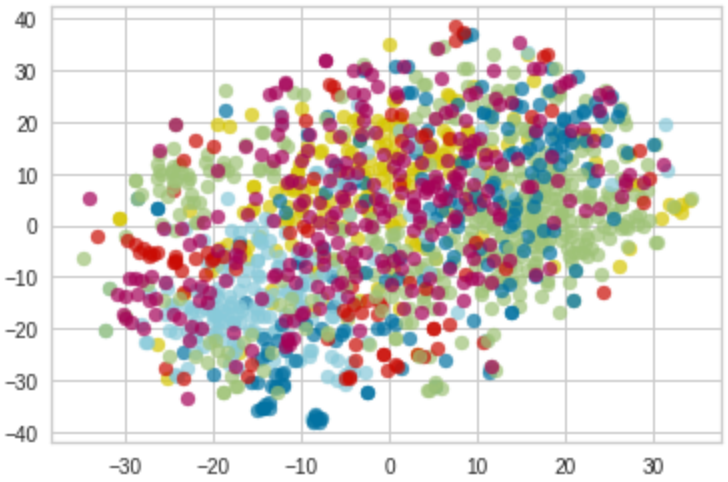
```
_____________________________
['Patriot Games']
_____________________________
_____________________________
['The Usual Suspects']
_____________________________
_____________________________
['True Romance']
_____________________________
_____________________________
[]
_____________________________
_____________________________
['In the Name of the Father']
_____________________________
_____________________________
['Basic Instinct']
_____________________________
_____________________________
['Salinui chueok']
_____________________________
```

O cluster 4 não fica bem definido pelo gráfico e quanto a recomendação, existe uma certa relação pois todos os filmes tratam de crimes e investigação policial, no entanto, os gêneros são diferentes.

Em uma recomendação talvez faça sentido oferecer filmes que trata de assuntos semelhantes mesmo que de gêneros diferentes.

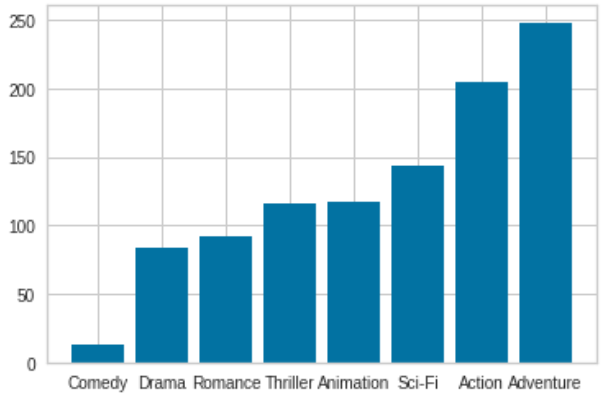
Infelizmente o título de um dos filmes não pode ser recuperado por não ter registro na iMdb_movies.

A clusterização feita utilizando o resumo dos filmes ficou bem espaçada e pouco definida.

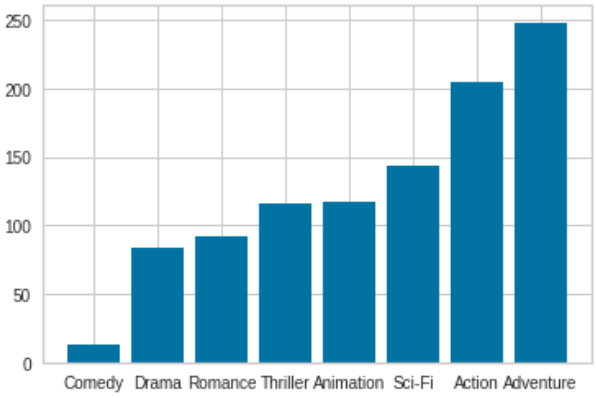


Clusters gerados pelo processamento dos resumos

Entretanto, apesar da dispersão os gráficos de distribuição de gênero por cluster ficaram semelhantes aos da sinopse.



Clusters 3



Clusters 4

Em seguida, foram geradas as recomendações para os clusters, levando em consideração os clusters 3 e 4 avaliadas acima tivemos os seguintes resultados:

```
_____________________________
['Saving Private Ryan']
_____________________________
_____________________________
['Enemy at the Gates']
_____________________________
_____________________________
['Das Boot']
_____________________________
_____________________________
['Rogue One']
_____________________________
_____________________________
['The Hunt for Red October']
_____________________________
_____________________________
['Captain America: The First Avenger']
_____________________________
_____________________________
['Iron Man 2']
_____________________________
```

Ao gerar o top 5 de recomendações de filmes, o cluster 3 foi extremamente coerente, oferecendo filmes da marvel e outros relacionados a guerra que condizem também com os roteiros dos filmes da Marvel.

```
_____________________________
['The Descent']
_____________________________
_____________________________
['The Parent Trap']
_____________________________
_____________________________
['Metropolis']
_____________________________
_____________________________
['The Piano']
_____________________________
_____________________________
['Blue Valentine']
_____________________________
_____________________________
['Sunrise: A Song of Two Humans']
_____________________________
_____________________________
['Buffy the Vampire Slayer']
```

O cluster 4, no entanto, falhou ao oferecer um filme de terror e um filme infantil na lista misturado com filmes de drama.

Outras análises semelhantes a essa podem ser feitas observando o arquivo 'ProjetoFinalPLN.ipynb'.

Análise de sentimento

Usando o algoritmo Vader foram geradas os scores para as reviews dos filmes, entretanto, na própria tabela de review é possível encontrar uma coluna relativa a avaliação do usuário com um valor de 0 a 10. Visando avaliar se o retorno no algoritmo estava coerente fiz uma mudança de escala no resultado retornado pelo Vader, o qual disponibiliza um score que varia de -1 a 1, para isso usei a seguinte fórmula:

novo_score = 10(velho_score +1)/2

Com o valor (novo_score) gerado, fiz uma comparação com alguns filmes da base, que pode ser visto abaixo:

```
_____________________________
['The Shawshank Redemption']
score tabela = 10

score calculado = 10.0
_____________________________
['The Godfather']
```

```
score tabela = 10

score calculado = 10.0
_____________________________
['The Dark Knight']
score tabela = 8

score calculado = 9.9
_____________________________
['Pulp Fiction']
score tabela = 9

score calculado = 9.1
_____________________________
['The Matrix']
score tabela = 4

score calculado = 9.9
_____________________________
['Sen to Chihiro no kamikakushi']
score tabela = 10

score calculado = 7.0
_____________________________
['Rear Window']
score tabela = 6

score calculado = 6.0
_____________________________
['Nuovo Cinema Paradiso']
score tabela = 7

score calculado = 10.0
_____________________________
['Oldeuboi']
score tabela = 1

score calculado = 0.3
```

A eficiência do vader é impressionante, os valores calculadas ficaram bem próximos dos valores da tabela, provando o quão forte é o algoritmo usado.

Naive Bayes

Com o objetivo de avaliar a eficiência do algoritmo Naive Bayes para classificar reviews com spoiler ou não, foi impressa a matriz de confusão para os dados de treinamento e teste, e também uma matriz contendo os: ‘verdadeiros positivos’, ‘falso positivo’, ‘falso negativo’ e ‘verdadeiro negativo’

.

Ao avaliar os dados de treinamento é possível notar que a precisão é alta e o recall também, indicando que os elementos selecionando são relevantes e que a quantidade de elementos relevantes selecionados também é alta.

Analisando a matriz verdade vemos que apenas 339 reviews foram classificadas incorretamente como falso positiva, e pensando em um situação prática seria marcar alguma review que não contém spoiler como ‘true’ do que como ‘false’, dessa forma impedimos de qualquer maneira que o usuário da plataforma seja exposto a algo que não queira ler.

```
------------------Estatísticas de treinamento------------------
Acurácia: 0.9212543554006969
Acurácia de previsão: 0.9212543554006969
              precision    recall  f1-score   support

        True       1.00      0.89      0.94      3181
       False       0.77      1.00      0.87      1124

    accuracy                           0.92      4305
   macro avg       0.88      0.95      0.91      4305
weighted avg       0.94      0.92      0.92      4305

                  Is_Spoiler(prev)=False  Prev. Is_Spoiler(prev)=True
Is_Spoiler=False                    2842                         339
Is_Spoiler=True                        0                        1124
```

Ao analisar as estatísticas da massa de teste o resultado é um pouco pior, como esperado. A acurácia diminui para quase 50%. Nota-se que a precisão e o recall para ‘True’ é melhor do que para ‘False’, indicando que o algoritmo quem mais facilidade em identificar uma mensagem que contem spoiler do que uma que não tem.

Entretanto, ao analisar a matriz verdade o número de falso positivos é alto assim como falso negativo, indicando que o algoritmo tende a sempre classificar como positivo e não reconhece quando não há spoiler.

```
------------------Estatísticas de teste------------------
Acurácia de previsão: 0.537979094076655
              precision    recall  f1-score   support

        True       0.71      0.64      0.67      1061
       False       0.20      0.25      0.22       374

    accuracy                           0.54      1435
   macro avg       0.45      0.45      0.45      1435
weighted avg       0.57      0.54      0.55      1435

                  Is_Spoiler(prev)=False  Prev. Is_Spoiler(prev)=True
Is_Spoiler=False                     678                         383
Is_Spoiler=True                      280                          94
```

Conclusão

A clusterização possibilitar diversas análises, portanto, para o resultado mais assertivo acredito que seriam necessárias outras abordagens, a fim de apurar os resultados.

O resultado da análise de sentimento foi bastantes satisfatório, além do algoritmo Vader ser simples de ser usado, garante resultados realmente surpreendentes. Entretanto, gostaria de desenvolver mais e expandir os horizontes, pois acredito que seria possível outros caminho e visualizações.

A classificação usando Naive Bayes se mostrou pouco eficiente, porém, acredito que se aumentasse a base de treinamento os resultado poderiam ser melhores.

Gostaria de ter feito mais, entretanto, tive muita dificuldade para processar grandes volumes de dados usando o google Colab. No geral, foi gratificante fazer esse trabalho, aplicar os conhecimento na prática e ver coerência dos resultados contribuiu para meu conhecimento.