Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Antônio Côrtes Rodrigues

MONOGRAFIA DE PROJETO ORIENTADO EM COMPUTAÇÃO II

Em Alta no YouTube: Comparação de vídeos populares entre países

Belo Horizonte
2019 / 2º semestre

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Ciência da Computação
Curso de Bacharelado em <CURSO>

Em Alta: Comparação de vídeos populares entre países

por

Antônio Côrtes Rodrigues

Monografia de Projeto Orientado em Computação II

Apresentado como requisito da disciplina de Projeto Orientado em Computação do Curso de Bacharelado em Ciência da Computação da UFMG

Prof. Dr. Jussara Marques de Almeida
Orientadora

Belo Horizonte
2019 / 2º semestre

RESUMO

Este trabalho tem como intuito detectar similaridade entre países no que se refere ao conteúdo de vídeos assistidos na plataforma do YouTube. Os vídeos de interesse são da categoria Em Alta do YouTube, que se referem aos vídeos mais populares de um determinado dia. Estes dados foram tirados de um banco de dados disponibilizado na plataforma Kaggle. É previsto que países próximos tenham culturas similares, e que suas preferências de conteúdo estejam alinhadas. Nove países foram analisados neste estudo: Canadá, Estados Unidos, México, Grã-Bretanha, França, Alemanha, Índia, Rússia e Coreia do Sul. Cada vídeo no YouTube possui diversas informações atreladas a ele, como número de visualizações, números de usuários que gostaram do vídeo, número de usuários que não gostaram do vídeo, etc. Neste trabalho, analisamos as distribuições gerais de números de visualizações e likes para países, e agrupamos os países por estas distribuições, traçando perfis de comportamento geral dos vídeos. Em seguida, separamos vídeos por categorias de interesse, e analisamos como os países se assemelham para vídeos destas categorais.

Palavras-chave:.similaridade, clusterização, vídeos, YouTube, banco de dados

ABSTRACT

This work's objective is to detect similarities between countries by analyzing videos watched on YouTube on each country. The videos used came from the Trending resource from YouTube, that detects the most popular videos in a given day. The data utilized in this work was obtained from Kaggle. It is expected that countries in the same region watch similar content, considering they usually have similar culture. Nine countries were analyzed in this work: Canada, United States, Mexico, Great-Britain, France, Germany, India, Russia, South Korea. Each video on YouTube has a number of data associated with it, such as number of views, number of people who liked the video, number of people who disliked the video, etc. In this work, we analyze the distributions of views and likes for each country, and we group them up using these distributions, identifying general behavioral profiles for videos. We then proceed with an analysis of certain categories of interest in which we identify countries with similar behaviors inside each of these categories.

Keywords: similarity, clustering, videos, YouTube, databases

LISTA DE FIGURAS

Figura 1 Representações das matrizes de similaridade entre os países……..…...3

LISTA DE SIGLAS

t-SNE	t-distributed Stochastic Neighbor Embedding

LISTA DE TABELAS

Tabela 1 Definição dos intervalos criados para números de visualizações ……..…….6

Tabela 2 Definição dos intervalos criados para números de likes……6

Tabela 3 Valor da métrica Silhouette para um dado número de clusters dos vetores representando a distribuição de vídeos em todo o intervalo de dados...……7

Tabela 4 Categorias dos vídeos disponíveis no YouTube com seus identificadores. Marcado em vermelho as categorias escolhidas para análise.…….12

Tabela 5 Valor da métrica Silhouette para um dado número de clusters dos vetores representando a distribuição mensal dos vídeos da categoria Films & Animation…….13

Tabela 6 Valor da métrica Silhouette para um dado número de clusters dos vetores representando a distribuição mensal dos vídeos da categoria Music……..17

Tabela 7 Valor da métrica Silhouette para um dado número de clusters dos vetores representando a distribuição mensal dos vídeos da categoria News & Politics ……22

LISTA DE GRÁFICOS

Gráfico 1 Representação t-SNE do espaço de 32 dimensões para os vetores de distribuição de vídeos por categoria em um mês, com agrupamentos entre os vetores em 3 clusters obtidos pelo algoritmo K-Means……..…….3

Gráfico 2 Representação t-SNE do espaço de 26 dimensões para os vetores de distribuição de vídeos por categoria por intervalos mensais, com agrupamentos entre os vetores em 2 clusters obtidos pelo algoritmo K-Means……...……8

Gráfico 3 Representação t-SNE do espaço de 26 dimensões para os vetores de distribuição de vídeos por categoria por intervalos mensais, com agrupamentos entre os vetores em 3 clusters obtidos pelo algoritmo K-Means…..……..9

Gráfico 4 Representação t-SNE do espaço de 26 dimensões para os vetores de distribuição de vídeos por categoria por intervalos mensais, com agrupamentos entre os vetores em 4 clusters obtidos pelo algoritmo K-Means……..10

Gráfico 5 Representação t-SNE do espaço de 26 dimensões para os vetores de distribuição de vídeos por categoria por intervalos mensais, com agrupamentos entre os vetores em 5 clusters obtidos pelo algoritmo K-Means.…….11

Gráfico 6 Representação t-SNE do espaço de 27 dimensões para os vetores de distribuição de visualizações e likes dos vídeos da categoria Films & Animation em intervalos mensais, com agrupamentos entre os vetores em 2 clusters obtidos pelo algoritmo K-Means……..14

Gráfico 7 Representação t-SNE do espaço de 27 dimensões para os vetores de distribuição de visualizações e likes dos vídeos da categoria Films & Animation em intervalos mensais, com agrupamentos entre os vetores em 3 clusters obtidos pelo algoritmo K-Means…….…….15

Gráfico 8 Representação t-SNE do espaço de 27 dimensões para os vetores de distribuição de visualizações e likes dos vídeos da categoria Films & Animation em intervalos mensais, com agrupamentos entre os vetores em 4 clusters obtidos pelo algoritmo K-Means……..16

Gráfico 9 Representação t-SNE do espaço de 27 dimensões para os vetores de distribuição de visualizações e likes dos vídeos da categoria Films & Animation em intervalos mensais, com agrupamentos entre os vetores em 5 clusters obtidos pelo algoritmo K-Means.…….17

Gráfico 10 Representação t-SNE do espaço de 27 dimensões para os vetores de distribuição de visualizações e likes dos vídeos da categoria Music em intervalos mensais, com agrupamentos entre os vetores em 2 clusters obtidos pelo algoritmo K-Means……..18

Gráfico 11 Representação t-SNE do espaço de 27 dimensões para os vetores de distribuição de visualizações e likes dos vídeos da categoria Music em intervalos mensais, com agrupamentos entre os vetores em 3 clusters obtidos pelo algoritmo K-Means……...19

Gráfico 12 Representação t-SNE do espaço de 27 dimensões para os vetores de distribuição de visualizações e likes dos vídeos da categoria Music em intervalos mensais, com agrupamentos entre os vetores em 4 clusters obtidos pelo algoritmo K-Means……..…….20

Gráfico 13 Representação t-SNE do espaço de 27 dimensões para os vetores de distribuição de visualizações e likes dos vídeos da categoria Music em intervalos mensais, com agrupamentos entre os vetores em 5 clusters obtidos pelo algoritmo K-Means……..…….21

Gráfico 14 Representação t-SNE do espaço de 27 dimensões para os vetores de distribuição de visualizações e likes dos vídeos da categoria News & Politics em intervalos mensais, com agrupamentos entre os vetores em 2 clusters obtidos pelo algoritmo K-Means…….…….22

Gráfico 15 Representação t-SNE do espaço de 27 dimensões para os vetores de distribuição de visualizações e likes dos vídeos da categoria News & Politics em intervalos mensais, com agrupamentos entre os vetores em 3 clusters obtidos pelo algoritmo K-Means……..23

Gráfico 16 Representação t-SNE do espaço de 27 dimensões para os vetores de distribuição de visualizações e likes dos vídeos da categoria News & Politics em intervalos mensais, com agrupamentos entre os vetores em 4 clusters obtidos pelo algoritmo K-Means…….…….24

Gráfico 17 Representação t-SNE do espaço de 27 dimensões para os vetores de distribuição de visualizações e likes dos vídeos da categoria News & Politics em intervalos mensais, com agrupamentos entre os vetores em 5 clusters obtidos pelo algoritmo K-Means……..25

SUMÁRIO

RESUMO ..... **III**

ABSTRACT ..... **IV**

LISTA DE FIGURAS ..... **V**

LISTA DE SIGLAS ..... **VI**

LISTA DE TABELAS ..... **VII**

LISTA DE GRÁFICOS ..... **VIII**

1 INTRODUÇÃO **1**

2 CONTEXTUALIZAÇÃO E TRABALHOS RELACIONADOS **2**

3 DESENVOLVIMENTO DO TRABALHO **5**

3.1 Adaptação e entendimento do banco de dados 5
3.2 Análise de distribuição de visualizações e likes entre os países 5
3.3 Análises de comportamento das categorias 11
3.3.1 Clusters para categoria Films & Animation 13
3.3.2 Clusters para categoria Music 17
3.3.3 Clusters para categoria News & Politics 22

4 RESULTADOS E DISCUSSÃO **26**

5 CONCLUSÕES **27**

6 REFERÊNCIAS **29**

1 INTRODUÇÃO

Vídeos são uma das principais formas de conteúdo da sociedade contemporânea; seja em redes sociais ou grupos de mensagens, vídeos são uma forma central de compartilhar conteúdo. Talvez a maior responsável pela popularização dos vídeos foi a plataforma YouTube, cujo principal recurso é justamente a criação e postagem de vídeos, e que criou um dos primeiros grandes mercados digitais. Embora tenha visto uma perda na dominância devido a criação de redes sociais como Instagram e Twitter por serem mais dinâmicas, o YouTube continua o grande nome de plataforma voltada para vídeos, e continua a se adaptar e implementar novos recursos.

Já estabelecido há alguns anos, o recurso dos vídeos "Em Alta" do YouTube é sem dúvida uma das principais características do site. Ele coloca em destaque os vídeos mais populares no momento ao analisar um conjunto de características, como número de visualizações, quão rápido o vídeo está gerando visualizações, dentre outros fatores. Todos os criadores de conteúdo procuram ter seus vídeos na categoria Em Alta, pois assim atraem mais atenção para seus canais, e com mais visualizações, mais renda é retornada para eles.

Cada país possuí seu próprio ranking de vídeos Em Alta. Deste modo, o Em Alta pode ser usado para identificar conteúdo que foi popular em um dado país, e descrever os gostos de um país de acordo com conteúdo dos vídeos.

O objetivo deste trabalho é justamente usar os vídeos para agrupar e diferenciar países de um banco de dados de vídeos Em Alta no YouTube, procurando identificar como os conteúdos de vídeos assistidos de cada país são similares entre si.

2 CONTEXTUALIZAÇÃO E TRABALHOS RELACIONADOS

Analisar as redes sociais e como seu conteúdo se distribui é fundamental para entender como conteúdo é consumido no século XXI. Não sendo uma exceção, o YouTube é uma das plataformas centrais na idade contemporânea, com mais de um bilhão de horas de vídeos sendo assistidos por dia no mundo (YOUTUBE IN NUMBERS, YOUTUBE FOR PRESS). A rede social mostra um retrato da sociedade, como as pessoas se relacionam entre si, o que elas assistem, o que é popular nesta época. Isto é especialmente verdade no momento em que os dados deste estudo foram coletados, no final de 2017 até meados de 2018.

O recurso Em Alta é central em analisar esta distribuição de conteúdo por representar justamente o que é mais popular em cada dia. Ele revela o que a população gosta, de uma maneira geral, e quais assuntos são os mais relevantes.

Em um mundo em que a globalização nunca foi tão grande, e conteúdo é tão fácil de ser compartilhado, entender quão forte são as relações entre países se mostra extremamente interessante. Podemos nos aproximar de um entendimento de como os efeitos destes avanços da sociedade e fatores históricos aproximam países, por meio desta análise proposta da plataforma do YouTube.

O artigo *You are What you Eat (and Drink): Identifying Cultural Boundaries by Analyzing Food & Drink Habits in Foursquare* foi uma excelente referência de trabalho de análise cultural qualitativa, no qual eram detectadas semelhanças entre países e cidades em relação a tipos de comida e bebida consumidos. O trabalho tinha intuito parecido com o desenvolvido neste relatório, e por isto

No texto, foram tiradas ideias de representação dos dados, métricas para computar similaridades, métodos para representação dos dados e sua análise, como algoritmos de clusterização.

Na primeira parte deste estudo, foi feita uma análise da similaridade das distribuições gerais das categorias dos vídeos entre os países. Primeiramente utilizamos similaridade de cossenos para avaliar semelhanças entre países, usando a métrica entre as representações vetoriais entre cada par país por país. Esta primeira análise mostrou que países possuíam uma relação bastante alta entre si de modo geral, o que preocupou um pouco os resultados futuros do estudo.

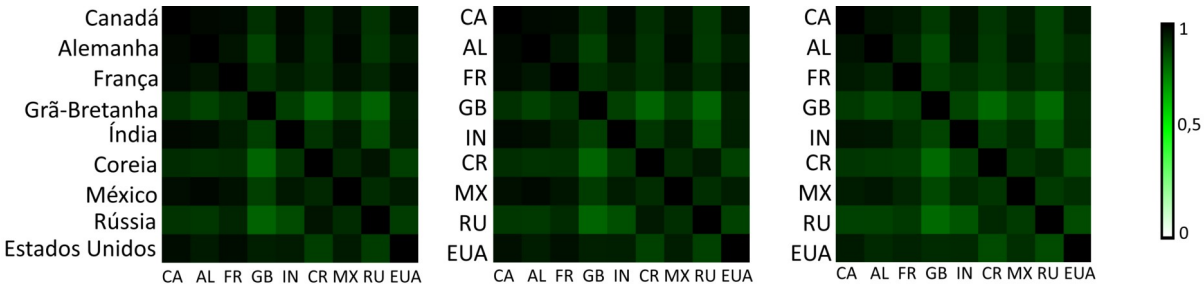

Figura 1: Representações das matrizes de similaridade entre os países. Na esquerda, a matriz para representação da distribuição dos vídeos em todo o intervalo (seis meses); no meio, para intervalos mensais; na direita, para intervalos diários

A segunda parte do estudo utilizou representações vetoriais similares para os países, sendo distribuições de vídeos por suas categorias, para cada país. Utilizamos também separações por diferentes intervalos de tempo para diferentes estudos, sendo elas o intervalo completo (sete meses), intervalos mensais e intervalos diários. Para cada um dessas escolhas de intervalo, agrupamos cada um dos vetores dos países utilizando o algoritmo K-Means, de modo a estudar os agrupamentos criados. Para cada agrupamento, utilizamos a métrica Silhouette para avaliar sua força, e o método t-SNE para gerar as representações bidimensionais dos espaços vetoriais.

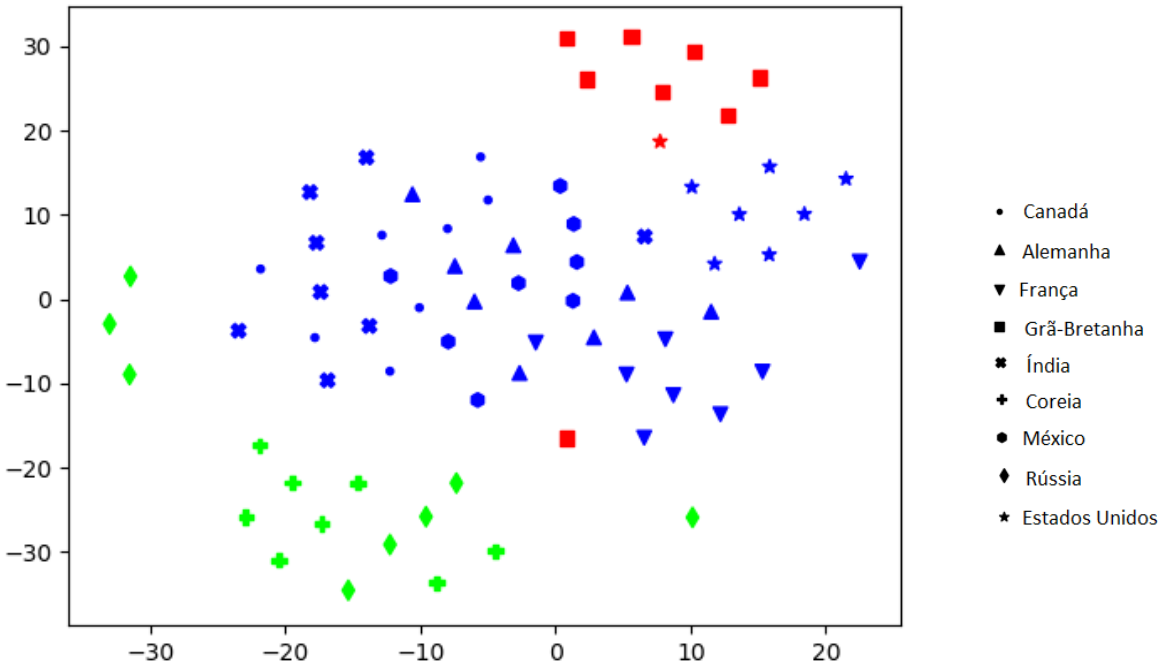

Gráfico 1: Representação t-SNE do espaço de 32 dimensões para os vetores de distribuição de vídeos por categoria em um mês, com agrupamentos entre os vetores em 3 clusters obtidos pelo algoritmo K-Means

A primeira parte deu familiaridade com as ferramentas necessárias para a continuidade do estudo. As análises desta segunda etapa teriam formato similar às feitas utilizando clusterização com o K-Means, mas adaptadas para algumas variações de contexto e representações. A métrica Silhouette foi empregada mais uma vez como forma de avaliar a força dos clusters, e também do estudo de forma geral, junto com o t-SNE para ajudar na interpretação dos estados em um espaço bidimensional. Isto permitiu um maior número de análises nesta segunda parte.

3 DESENVOLVIMENTO DO TRABALHO

3.1 Adaptação e entendimento do banco de dados

O banco de dados utilizado para este trabalho continha dados de dez países: Canadá, Estados Unidos, México, Grã-Bretanha, França, Alemanha, Índia, Rússia, Coreia e Japão. Com exceção do Japão, todos os países tinham o mesmo intervalo de dias, indo de 14 de Novembro de 2017 até 14 de Junho de 2018, com as datas 10/01/2018, 11/01/2018, 08/04/2018 até 13/04/2018 ausentes de suas tabelas. O Japão possuía mais intervalos ausentes em sua tabela, além de ter vários dias que possuíam um número muito baixo de vídeos, o que levou a decisão de não analisarmos o país, pelo menos nesta primeira parte do estudo.

O banco de dados tem origem do Kaggle, a plataforma de compartilhamento online de dados e projetos. Eles foram baixados em formato CSV e convertidos e importados para uma tabela pelo software MySQL Workbench para uso local. Como as tabelas do banco de dados eram de países bastante diferentes, a diferença nos caracteres causou alguns problemas no uso dos dados. Isto afetava, porém, somente dados qualitativos que não eram de interesse nesta parte do estudo, sendo eles: nome do vídeo, nome do criador do vídeo, descrição do vídeo, etc. Já que nossa análise é temporal e se refere à categoria dos vídeos e suas distribuições de visualizações e likes, pudemos descartar os dados que causaram problemas no uso das tabelas para focar nos dados de interesse.

3.2 Análise de distribuição de visualizações e likes entre os países

A primeira análise feita neste trabalho teve o intuito de detectar como cada país se comportava em número de visualizações e de likes para os vídeos Em Alta.

Os intervalos utilizados para o número de visualizações e de likes foi criada e acordo com padrões gerais de avaliação do YouTube e também para serem significativos. Na tabela, o início do intervalo é fechado, e o final, aberto.

Início do intervalo	Fim do intervalo	Início do intervalo	Fim do intervalo
0	10.000	500.000	750.000
10.000	25.000	750.000	1.000.000
25.000	50.000	1.000.000	2.000.000
50.000	75.000	2.000.000	3.000.000
75.000	100.000	3.000.000	5.000.000
100.000	250.000	5.000.000	-
250.000	500.000		

Tabela 1: Definição dos intervalos criados para números de visualizações

Início do intervalo	Fim do intervalo	Início do intervalo	Fim do intervalo
0	100	2.000	5.000
100	250	5.000	10.000
250	500	10.000	25.000
500	750	25.000	50.000
750	1.000	50.000	100.000
1.000	2.000	100.000	-
2.000	3.000		

Tabela 2: Definição dos intervalos criados para números de likes

O primeiro elemento indica que, em todo o intervalo, um certo país teve uma quantidade de vídeos no Em Alta com número de visualizações entre 0 e 10.000. O segundo elemento indica que neste intervalo de tempo o país teve uma quantidade de vídeos no Em Alta entre 10.000 visualizações e 25.000, etc. Até que acabam os intervalos de visualizações e começam os de likes, onde a representação é análoga. Com treze intervalos para visualizações e likes, os intervalos de tempo de cada um dos países foram representados como vetores de 26 elementos.

Foi escolhido olhar para intervalos mensais de distribuições destes valores de likes e visualizações, pois se mostrou o mais interessante na primeira parte do estudo. As distribuições eram normalizadas de acordo com o número de vídeos coletados naquele intervalo. Este número varia pois cada mês tem diferente número de dias, mas para cada dia em específico foram utilizados 100 vídeos do banco de dados.

O algoritmo K-Means foi escolhido para realizar os agrupamentos deste momento do trabalho. Ele agrupa os vetores de cada intervalo de cada país em diferentes grupos, e buscamos

enxergar quais países se uniam e quais se diferenciavam, além de procurar consistências do agrupamento de vetores de um mesmo país.

A avaliação da força das estruturas dos clusters foi realizada pela análise da métrica Silhouette. Como estamos analisando poucos países, não é necessário realizar nenhum método mais robusto para encontrar o número ótimo de clusters, e podemos apenas obter o valor Silhouette dentre para o número de clusters de dois (mínimo) até nove, e observar qual das estruturas têm maior força.

Nº Clusters	Silhouette
2	0.41922840982906684
3	0.34648886319852323
4	0.3562298482318066
5	0.37750965369390366
6	0.36563643911480415
7	0.34859844429422066
8	0.3385572060636888

Tabela 3: Valor da métrica Silhouette para um dado número de clusters dos vetores representando a distribuição de vídeos em todo o intervalo de dados

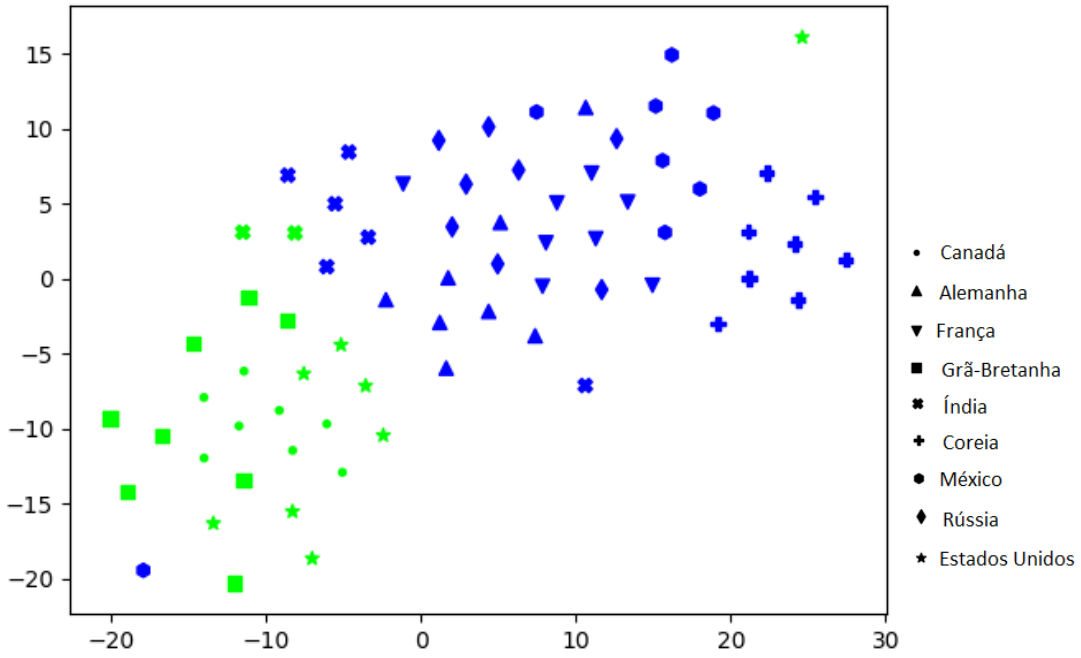

Gráfico 2: Representação t-SNE do espaço de 26 dimensões para os vetores de distribuição de vídeos por categoria por intervalos mensais, com agrupamentos entre os vetores em 2 clusters obtidos pelo algoritmo K-Means

Este foi o cluster com maior força detectado pelo Silhouette. Ele agrupou a Grã-Bretanha, Estados Unidos e Canadá, contra os demais países. Alguns meses da Índia também estão presentes neste grupo menor, e vemos que a tendência de agrupamento favoreceu países falantes do inglês, embora apenas parcialmente para a Índia. Esta análise esperava agrupar os países de populações e tamanho territorial similares, mas isto de fato não ocorreu aqui. Uma explicação plausível para este comportamento é que os vídeos em inglês ganham mais atenção mundial, pois é uma língua mais falada, além de compartilhar conteúdo entre si mais facilmente. Vale lembrar porém, que este estudo não permite definir as causas dos agrupamentos, mas somente eles em si e especulações.

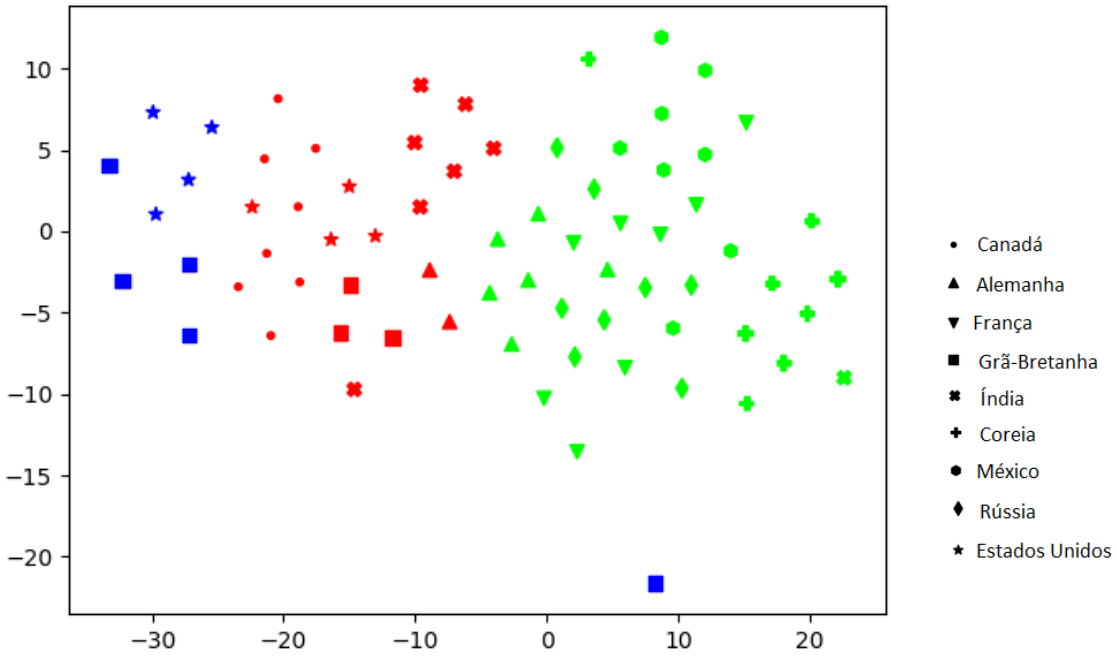

Gráfico 3: Representação t-SNE do espaço de 26 dimensões para os vetores de distribuição de vídeos por categoria por intervalos mensais, com agrupamentos entre os vetores em 3 clusters obtidos pelo algoritmo K-Means

Vemos que o agrupamento para 3 clusters separou o grupo dos falantes de inglês em dois grupos, com pouca alteração no grupo dos demais. Grã-Bretanha e Estados Unidos compõe um grupo com metade de seus elementos, definindo uma proximidade entre si; e novamente no segundo grupo criado, onde também são juntados ao Canadá, mais elementos da Índia, e alguns meses da Alemanha. O terceiro e último grupo é composto pelos demais países novamente, com as mudanças de menos elementos da Índia e da Alemanha o compondo.

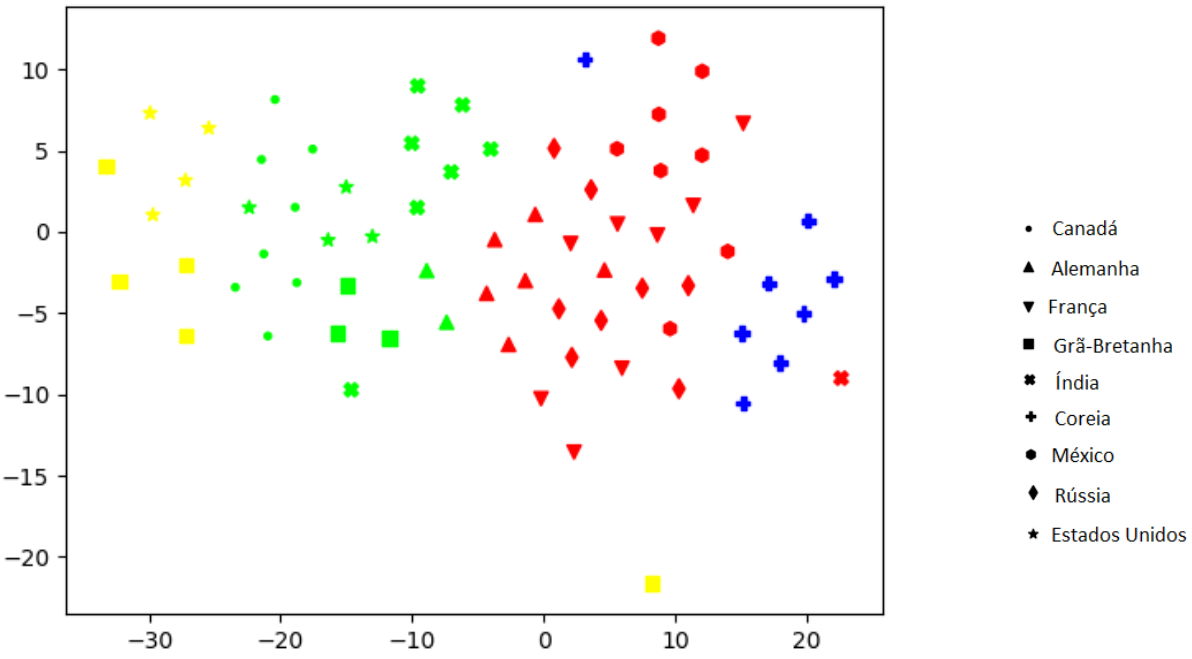

Gráfico 4: Representação t-SNE do espaço de 26 dimensões para os vetores de distribuição de vídeos por categoria por intervalos mensais, com agrupamentos entre os vetores em 4 clusters obtidos pelo algoritmo K-Means

O agrupamento utilizando 4 clusters levou a uma separação da Coreia dos dos demais países, com todos seus elementos separados. Vemos uma falta de alterações nos outros países e grupos. Agora aparece um comportamento mais relacionado a tamanho de país e consistência interna. O cluster para 4 tem menos força do que para 2, mas mais do que para 3, o que indica um comportamento interno.

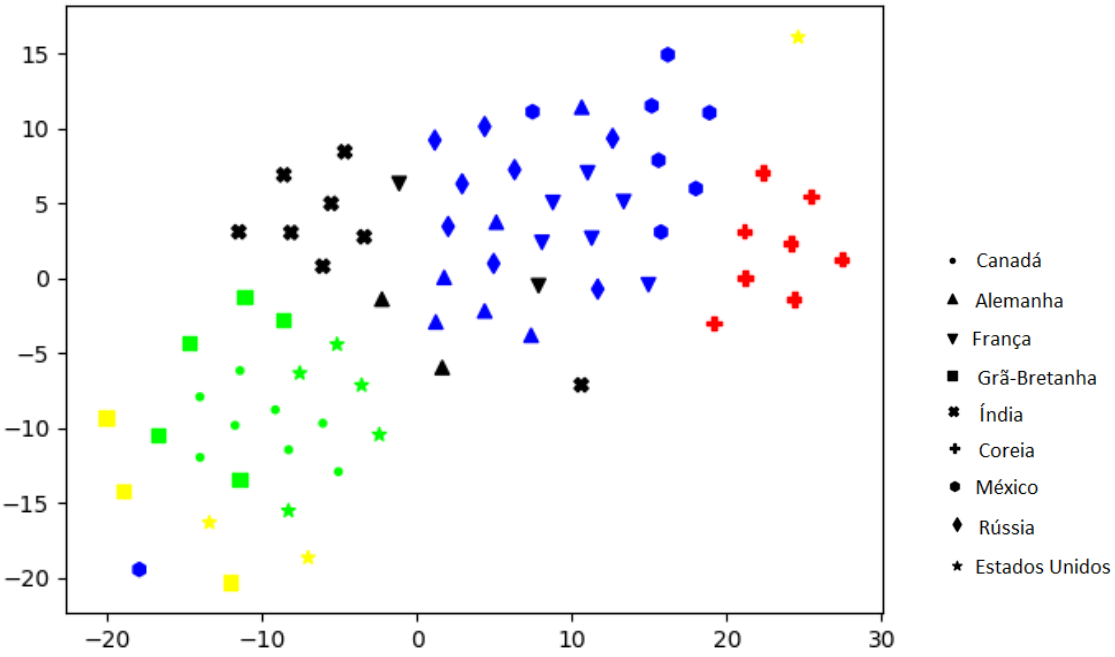

Gráfico 5: Representação t-SNE do espaço de 26 dimensões para os vetores de distribuição de vídeos por categoria por intervalos mensais, com agrupamentos entre os vetores em 5 clusters obtidos pelo algoritmo K-Means

Para agrupamentos utilizando 5 clusters, vemos poucas alterações novamente, com a exceção de um novo país separado: a Índia. Alguns elementos de França e Alemanha compões este grupo, o que indica alguma similaridade comportamental entre os países. O novo cluster tem mais força do que o de 4 grupos, o que aumenta a ideia de uma consistência interna entre países.

3.3 Análises de comportamento das categorias

Cada vídeo no YouTube possui uma categoria atrelada a ele, de acordo com a tabela acima. A representação de cada país é a contagem de quantos vídeos cada uma destas categorias teve nos cem primeiros vídeos de cada dia. As contagens eram agrupadas em diferentes intervalos, estas sendo: todo o intervalo, intervalos mensais, intervalos diários. Para a segunda parte deste trabalho, foi feita uma análise de agrupamentos dos países usando algoritmos de clusterização.

Para cada categoria escolhida, analisamos os vídeos agrupados por meses do intervalo, e representamos os país como múltiplos vetores normalizados de distribuição de visualizações e de

likes, assim como já havíamos usado na primeira etapa. Usamos também mais uma variável para este estudo, que foi o número de vídeos da categoria de interesse naquele intervalo.

Identificador	Categoria	Identificador	Categoria	Identificador	Categoria
1	Film&Animation	24	Entertainment	36	Drama
2	Autos & Vehicles	25	News&Politics	37	Family
10	Music	26	Howto&Style	38	Foreign
15	Pets&Animals	27	Education	39	Horror
17	Sports	28	Science&Tech	40	Sci-Fi/Fantasy
18	Short Vehicles	30	Movies	41	Thriller
19	Travel & Events	31	Anime/ Animation	42	Shorts
20	Gaming	32	Action/ Adventure	43	Shows
21	Videoblogging	33	Classics	44	Trailers
22	People & Blogs	34	Comedy*		
23	Comedy	35	Documentary		

Tabela 4: Categorias dos vídeos disponíveis no YouTube com seus identificadores. Marcado em vermelho as categorias escolhidas para análise.

As categorias escolhidas para análise foram Films & Animation, Music, e News & Politics. Elas possuíam um maior número de vídeos registrados de maneira geral, e pareceram as mais interessantes para um estudo.

Mais uma vez o algoritmo K-Means foi escolhido para realizar os agrupamentos. Ele agrupou os vetores de cada intervalo de cada país em diferentes grupos, e buscamos enxergar quais países se uniam e quais se diferenciavam, além de procurar consistências do agrupamento de vetores de um mesmo país.

A avaliação da força das estruturas dos clusters também foi realizada pela análise da métrica Silhouette. Como estamos analisando poucos países, não é necessário realizar nenhum método mais robusto para encontrar o número ótimo de clusters, e podemos apenas obter o valor Silhouette dentre para o número de clusters de dois (mínimo) até nove, e observar qual das estruturas têm maior força.

Para a representação do espaço de 27 dimensões (uma para cada intervalo de visualizações e de likes, além da dimensão com o número de vídeos daquela categoria) utilizamos o algoritmo t-SNE, que facilita para a compreensão humana como os vetores estão no espaço. Geramos representações no espaço de duas dimensões.

A biblioteca *sklearn* para Python foi fundamental no desenvolvimento desta parte do trabalho, pois já contém as implementações dos algoritmos K-Means e t-SNE, além de calcular a métrica Silhouette. Seu uso facilitou o cumprimento do objetivo do trabalho, da análise dos dados, ao providenciar uma estrutura de dados robusta e completa.

3.3.1 Clusters para categoria Films & Animation

Ao olhar as distribuições do número de visualizações e de likes para a categoria Films & Animation para intervalos mensais de cada país. Os valores da métrica Silhouette encontrados para os agrupamentos foram as seguintes:

Nº Clusters	Silhouette
2	0.5513498288545833
3	0.5678527908285417
4	0.6008619954895629
5	0.6135211176976235
6	0.4083094585887614
7	0.3142629689114915
8	0.3964946523903567

Tabela 5: Valor da métrica Silhouette para um dado número de clusters dos vetores representando a distribuição mensal dos vídeos da categoria Films & Animation

Vemos que dois e três números de clusters mostram Silhouettes maiores que os demais, e por isto possuem estruturas mais fortes. Iremos nos focar nas estruturas mais fortes detectadas pelo K-Means e representá-las usando o t-SNE.

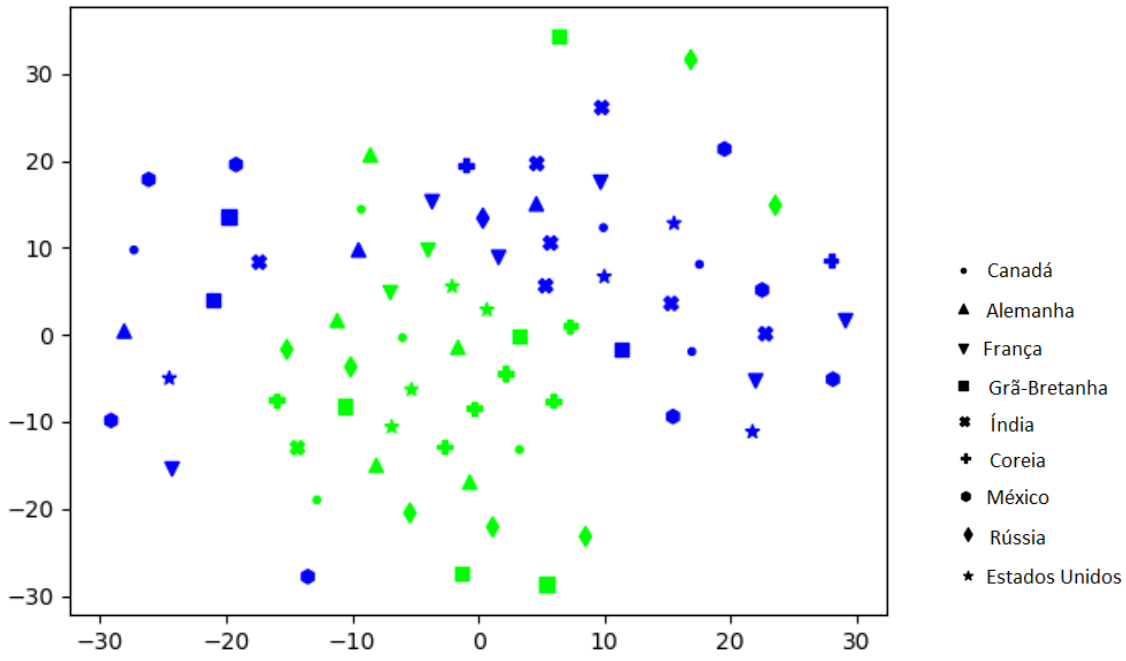

Gráfico 6: Representação t-SNE do espaço de 27 dimensões para os vetores de distribuição de visualizações e likes dos vídeos da categoria Films & Animation em intervalos mensais, com agrupamentos entre os vetores em 2 clusters obtidos pelo algoritmo K-Means

Esta primeira representação gerado mostra grupos bastante heterogêneos, mas que ainda possuem alguma concordância interna. Alemanha, Grã-Bretanha, Coreia e Rússia aparecem em mesmo grupo com a maior parte de seus elementos. Índia e México compõe o outro grupo também com bastante concordância. Abaixo, a relação entre países, o mês e o cluster assinalado para aquele mês para aquele país.

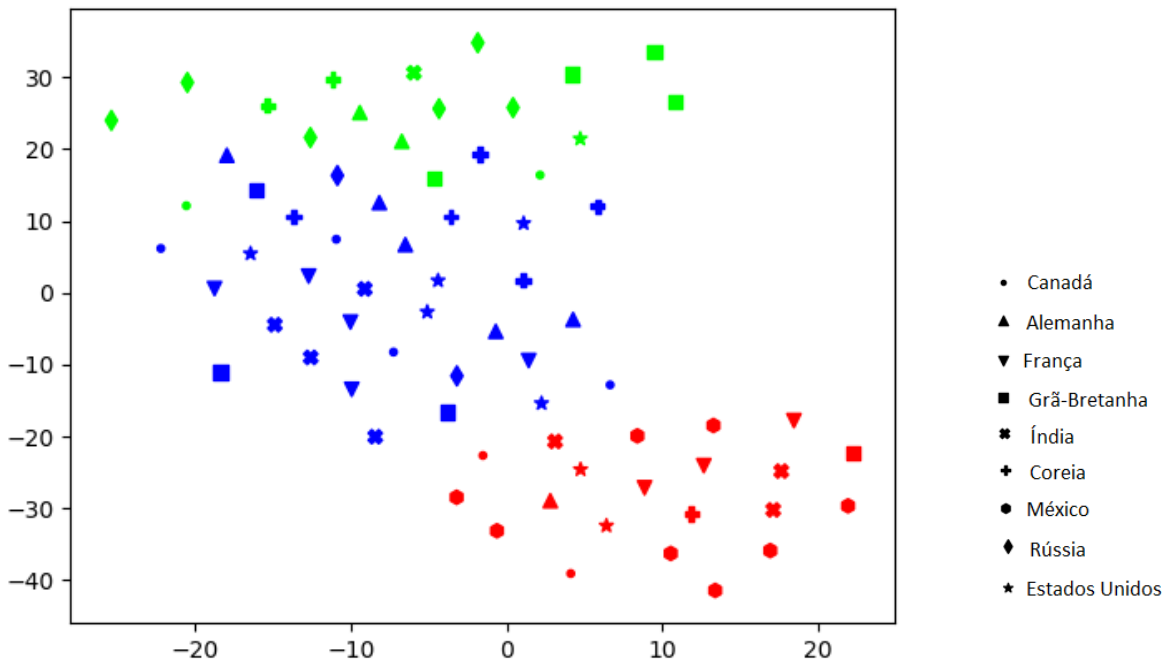

Gráfico 7: Representação t-SNE do espaço de 27 dimensões para os vetores de distribuição de visualizações e likes dos vídeos da categoria Films & Animation em intervalos mensais, com agrupamentos entre os vetores em 3 clusters obtidos pelo algoritmo K-Means

Mais uma vez o México, Rússia e Estados Unidos aparecem como países internamente consistentes. Os grupos ainda são bastante heterogêneos, se dividindo bastante entre os clusters. Estados Unidos tende a se agrupar um pouco com os países França, Índia e Coreia; enquanto Rússia e Grã-Bretanha aparecem em outro grupo. A França também se relaciona bastante México, ambos se dividindo entre os grupos.

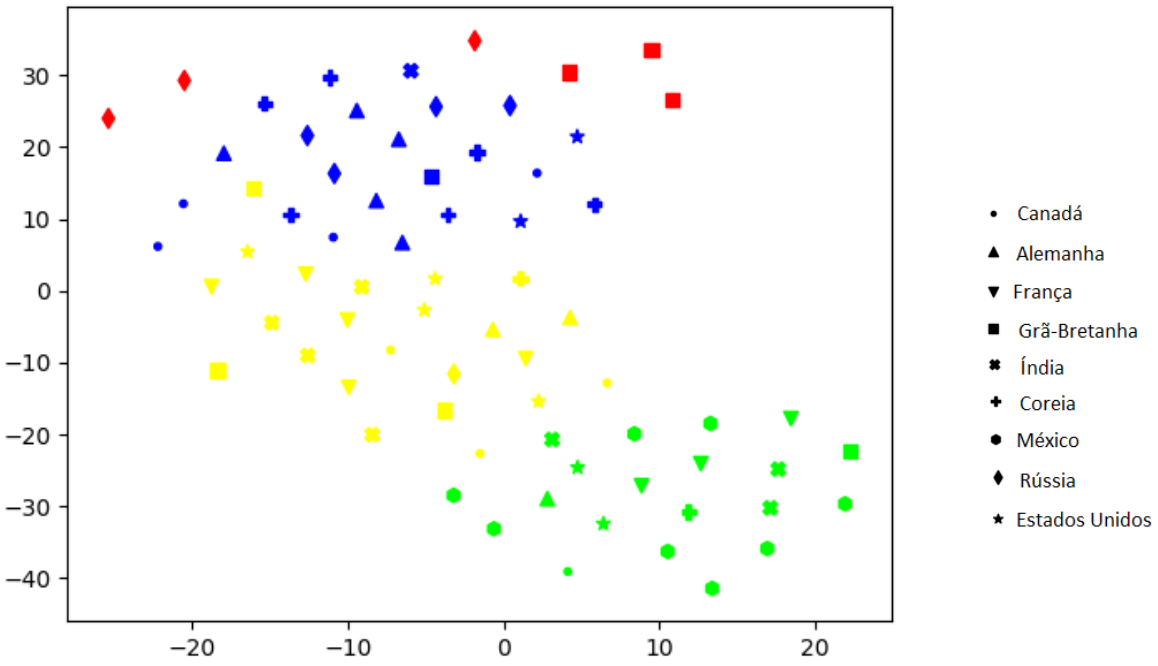

Gráfico 8: Representação t-SNE do espaço de 27 dimensões para os vetores de distribuição de visualizações e likes dos vídeos da categoria Films & Animation em intervalos mensais, com agrupamentos entre os vetores em 4 clusters obtidos pelo algoritmo K-Means

Rússia e Grã-Bretanha aparecem mais uma vez relacionadas em um grupo pequeno. Ambos porém se encontram bastante divididos, a Grã-Bretanha especialmente aparecendo próxima dos outros europeus Alemanha e França. Estados Unidos compõe outro grupo com o resto dos meses europeus, e também russos.

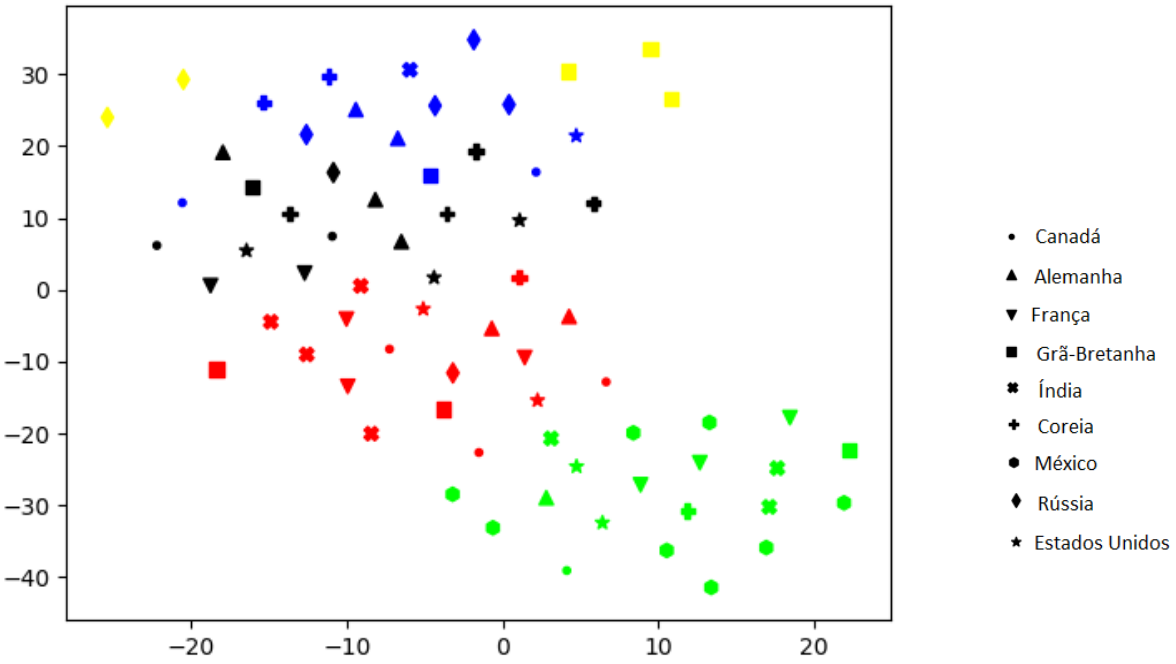

Gráfico 9: Representação t-SNE do espaço de 27 dimensões para os vetores de distribuição de visualizações e likes dos vídeos da categoria Films & Animation em intervalos mensais, com agrupamentos entre os vetores em 5 clusters obtidos pelo algoritmo K-Means.

O novo grupo formado separa alguns elementos da Alemanha, França, Estados Unidos, Canadá, além de poucos representantes de Rússia e Coreia. Os demais grupos não sofreram mudanças, deixando assim definida a estrutura mais forte de acordo com a métrica Silhouette.

3.3.2 Clusters para categoria Music

Nº Clusters	Silhouette
2	0.7894692698562689
3	0.6434105165075826
4	0.6445038647431428
5	0.58847262531526
6	0.5824634893838582
7	0.6015229717287345
8	0.6009166663926548

Tabela 6: Valor da métrica Silhouette para um dado número de clusters dos vetores representando a distribuição mensal dos vídeos da categoria Music

A métrica Silhouette indicou que as estruturas de dois a quatro clusters se mostraram mais interessantes para uma análise, tendo valores indicativos excelentes. Mais uma vez focaremos nossa análise para clusters de tamanho até cinco.

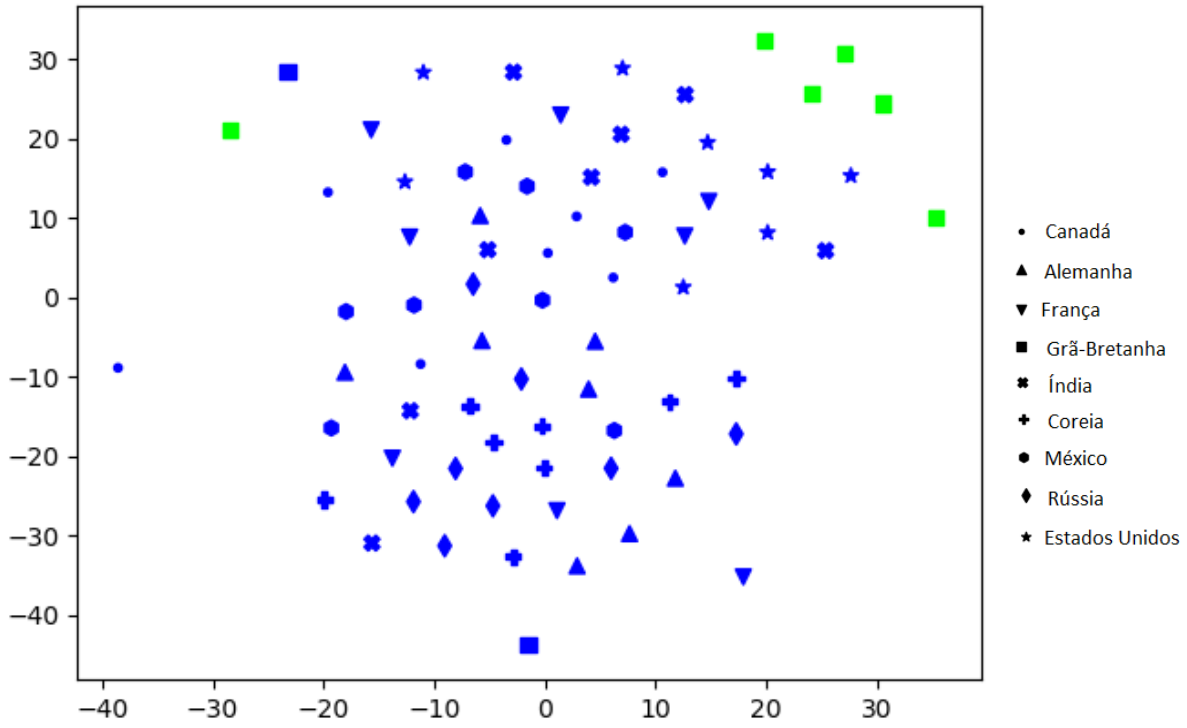

Gráfico 10: Representação t-SNE do espaço de 27 dimensões para os vetores de distribuição de visualizações e likes dos vídeos da categoria Music em intervalos mensais, com agrupamentos entre os vetores em 2 clusters obtidos pelo algoritmo K-Means

A primeira análise mostra que a Grã-Bretanha possui comportamento bastante diferente dos demais, resultado que apareceu algumas na primeira etapa deste trabalho. Os demais países compõe um outro grande grupo, e o fato deste cluster ter o maior Silhouette dentre os demais reforça a ideia da Grã-Bretanha ter interesse diferenciado para vídeos de música.

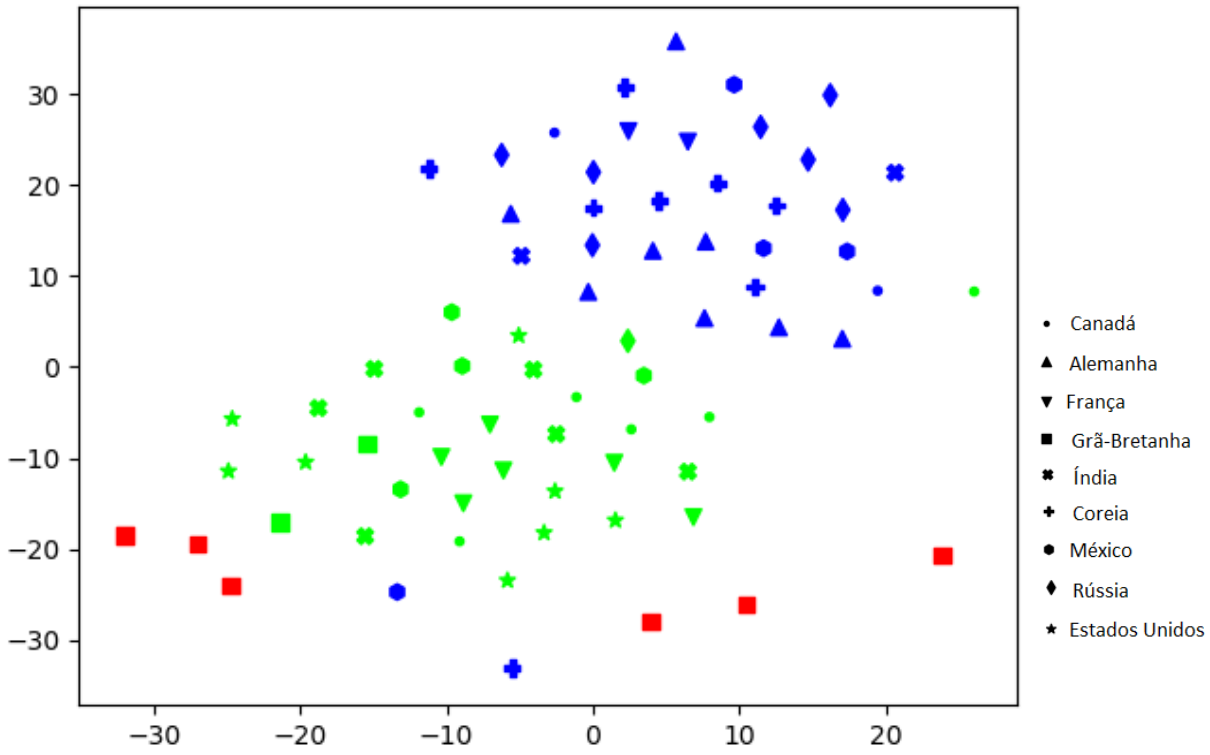

Gráfico 11: Representação t-SNE do espaço de 27 dimensões para os vetores de distribuição de visualizações e likes dos vídeos da categoria Music em intervalos mensais, com agrupamentos entre os vetores em 3 clusters obtidos pelo algoritmo K-Means

A Grã-Bretanha continua aparecendo como um grupo separado, reforçando sua diferença dos demais grupos. O grupo maior, no entanto, se divide em dois: França, Estados Unidos e Canadá de um lado; Alemanha, Rússia e Coreia em outro. Os demais países aparecem divididos entre os dois grupos.

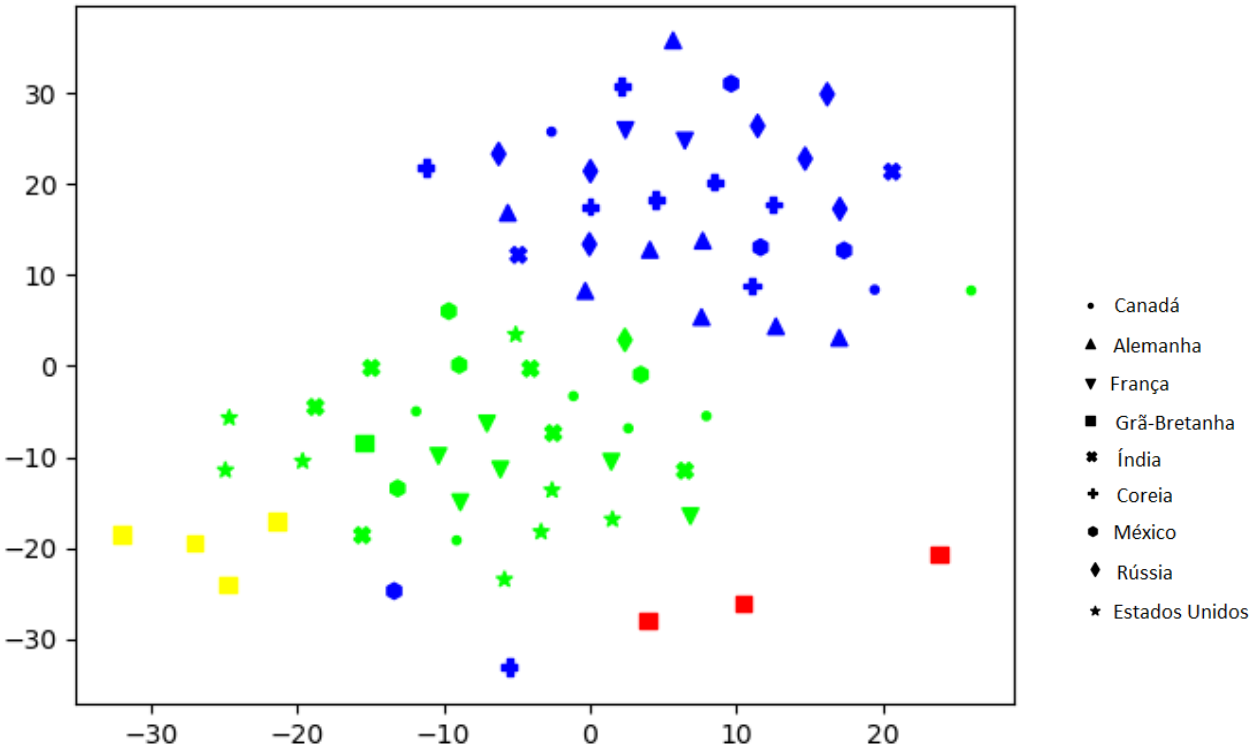

Gráfico 12: Representação t-SNE do espaço de 27 dimensões para os vetores de distribuição de visualizações e likes dos vídeos da categoria Music em intervalos mensais, com agrupamentos entre os vetores em 4 clusters obtidos pelo algoritmo K-Means

Aqui, de maneira inesperada, Grã-Bretanha se divide em dois grupos, deixando o grupo novo para três clusters inalterado. O país apresentou alguma variação no conteúdo entre os meses que levou a uma nova separação interna, o que é extremamente interessante e não observado em outros casos.

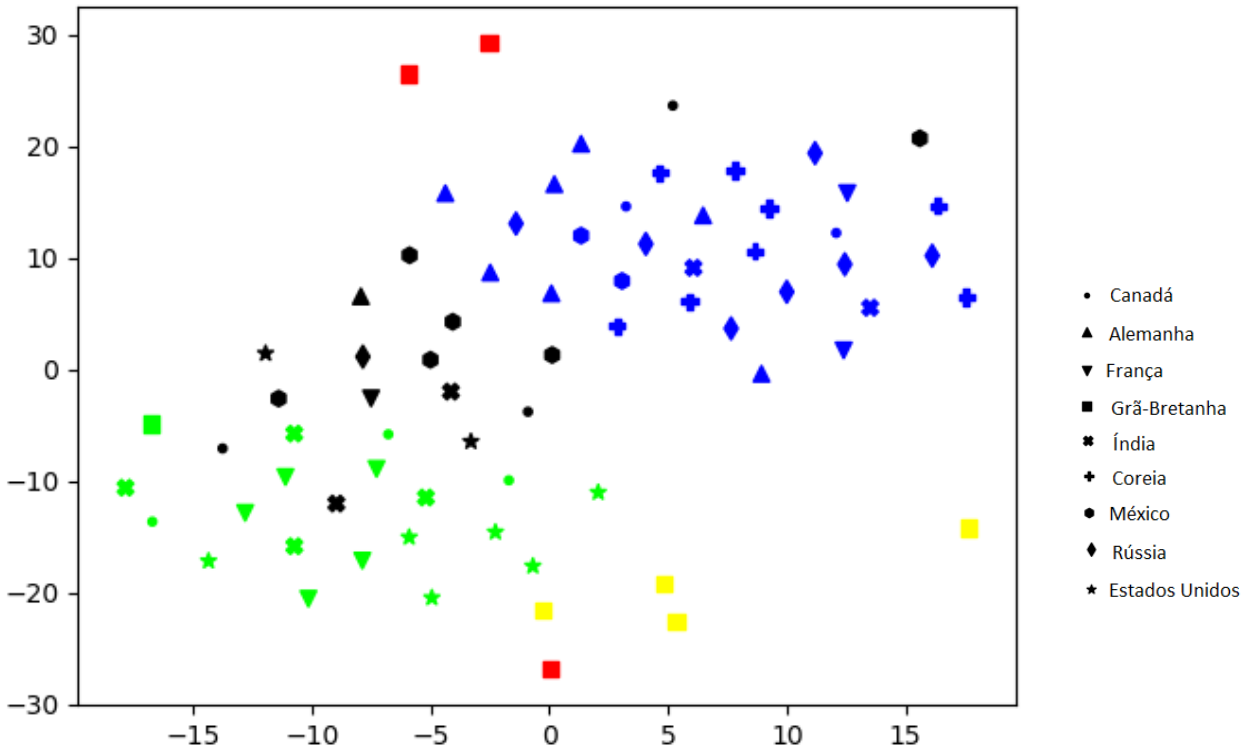

Gráfico 13: Representação t-SNE do espaço de 27 dimensões para os vetores de distribuição de visualizações e likes dos vídeos da categoria Music em intervalos mensais, com agrupamentos entre os vetores em 5 clusters obtidos pelo algoritmo K-Means

Para os cinco clusters, um novo país se mostra diferenciado: o México. Alguns meses de outros países se unem a este novo grupo, mas o México é o único que aprece com uma consistência. Os antigos dois grupos foram divididos para criação deste novo quinto, algo que não ocorreu com tanta frequência em transição de quatro para cinco clusters. Os grupos da Grã-Bretanha permaneceram inalterados.

3.3.3 Clusters para categoria News & Politics

Para o último tipo de agrupamento dos vídeos em intervalos, olhamos para as distribuições mensais da categoria News & Politics. Os valores de Silhouette obtidos para os agrupamentos usando estes vetores foram:

Nº Clusters	Silhouette
2	0.6840118358699276
3	0.6462576528819637
4	0.6306733038503771
5	0.6181679486236381
6	0.624484185279899
7	0.6276103944332562
8	0.6019899106812274

Tabela 7: Valor da métrica Silhouette para um dado número de clusters dos vetores representando a distribuição mensal dos vídeos da categoria News & Politics

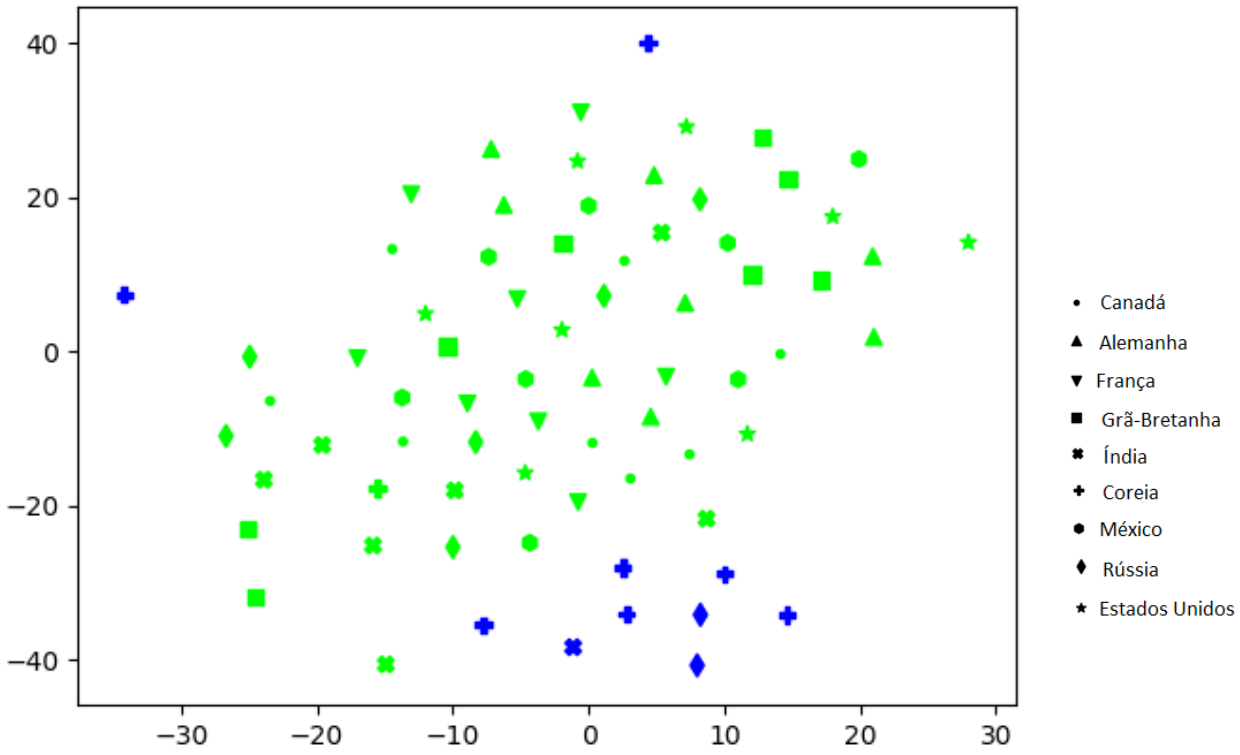

Gráfico 14: Representação t-SNE do espaço de 27 dimensões para os vetores de distribuição de visualizações e likes dos vídeos da categoria News & Politics em intervalos mensais, com agrupamentos entre os vetores em 2 clusters obtidos pelo algoritmo K-Means

Novamente para dois grupos aparece uma distinção entre um grupo pequeno e um grupo grande. Neste caso, porém, a Coreia compõe um grupo com poucos elementos de Rússia e Índia presentes. Os demais países formam um outro grande grupo, inclusive com a maioria dos vetores de Índia e Rússia. Embora este seja o Silhouette mais alto para a categoria, usar o K-Means para mais clusters deve retornar novos agrupamentos para estudar.

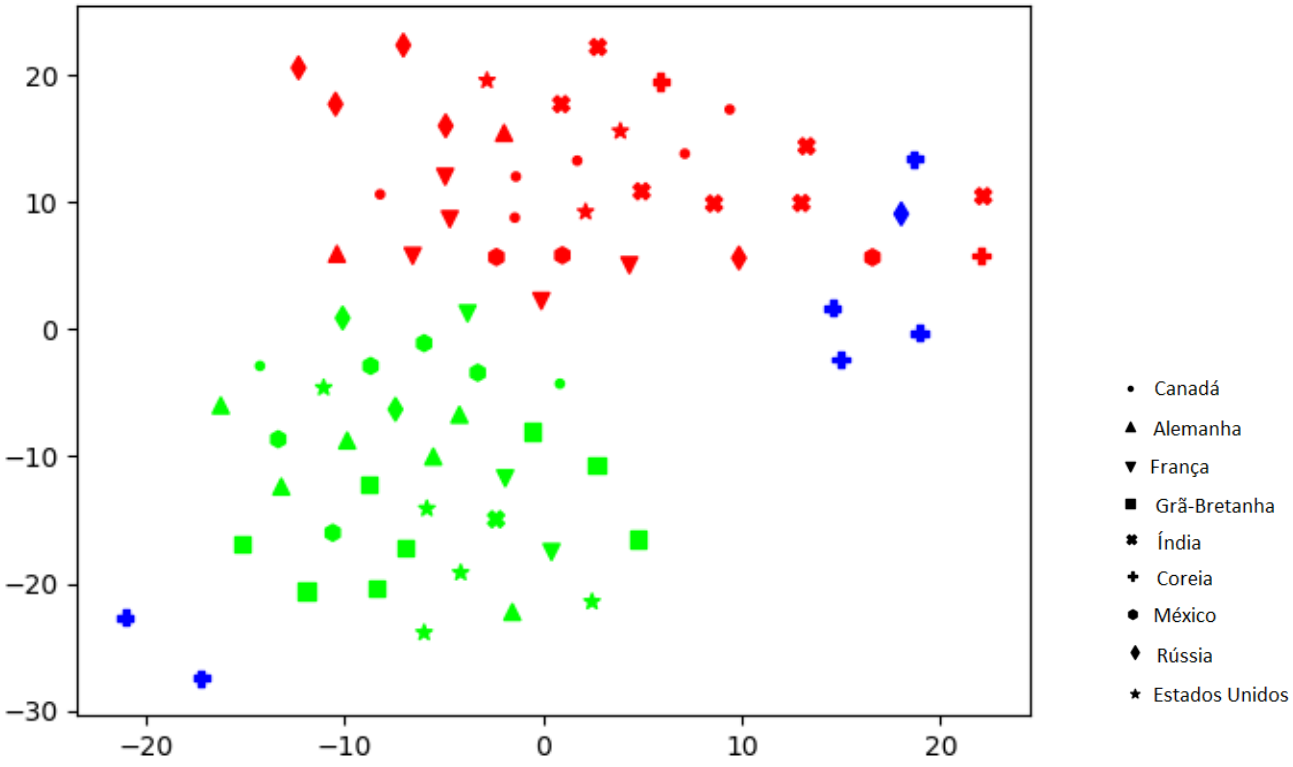

Gráfico 15: Representação t-SNE do espaço de 27 dimensões para os vetores de distribuição de visualizações e likes dos vídeos da categoria News & Politics em intervalos mensais, com agrupamentos entre os vetores em 3 clusters obtidos pelo algoritmo K-Means

Com três grupos, a Coreia agora aparece como um grupo mais fechado, apenas com um mês da Rússia junto. O grande grupo se divide em dois, um onde predomina Índia, Rússia, França e Canadá; e outro onde predominam México, Grã-Bretanha e Alemanha. A consistência interna dos países aparece um pouco fraca com divisões, mas mesmo assim permite encontrar relações entre eles.

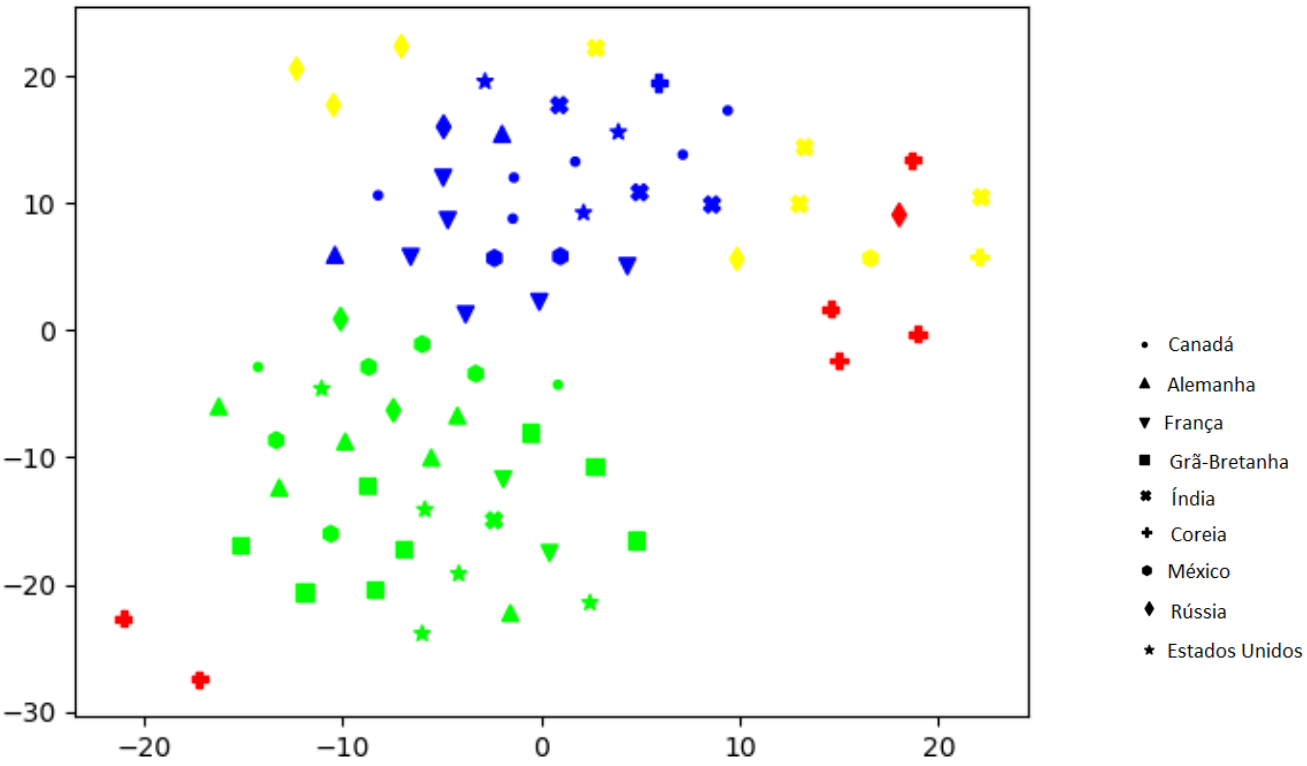

Gráfico 16: Representação t-SNE do espaço de 27 dimensões para os vetores de distribuição de visualizações e likes dos vídeos da categoria News & Politics em intervalos mensais, com agrupamentos entre os vetores em 4 clusters obtidos pelo algoritmo K-Means

Coreia e o grupo dos países México, Grã-Bretanha e Alemanha aparecem inalterados, mas o terceiro grupo recebe uma nova divisão. Rússia e Índia aparecem com elementos separados, gerando um novo perfil.

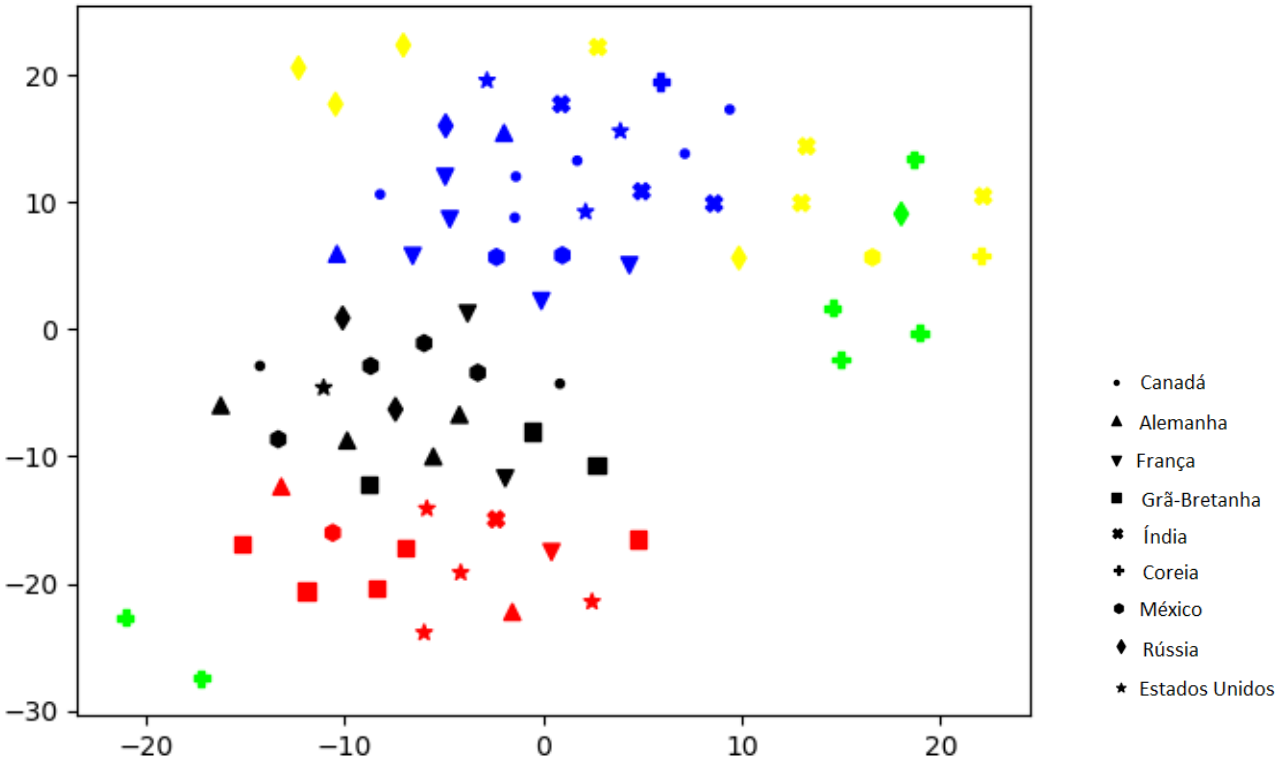

Gráfico 17: Representação t-SNE do espaço de 27 dimensões para os vetores de distribuição de visualizações e likes dos vídeos da categoria News & Politics em intervalos mensais, com agrupamentos entre os vetores em 5 clusters obtidos pelo algoritmo K-Means

Para esta última clusterização de News & Politics, os grupos anteriores mostraram nenhuma alteração, com exceção do grupo predominado pelos países México, Grã-Bretanha e Alemanha, que foi dividido em dois. Grã-Bretanha predominou uma nova parte junto com os elementos estadunidenses; enquanto México e Alemanha compuseram o novo grupo. Vale notar que a Coreia permaneceu unida em todos os clusters criados.

4 RESULTADOS E DISCUSSÃO

Os resultados indicaram que existe de fato grupos de países que possuem uma proximidade na preferência do conteúdo de vídeos assistidos no YouTube, estas proximidades variando de acordo com as categorias. Os valores da métrica Silhouette confirmam a força dos agrupamentos criados.

Os países orientais Rússia e Coreia foram agrupados de maneira consistente olhando para diferentes agrupamentos, o que indica a existência da relação entre os países, mesmo não sendo comumente associados de maneira geral. Isto indica que proximidade geográfica tem um efeito forte também no consumo de conteúdo de vídeos.

Os países ocidentais junto com a Índia compuseram o grupo de maneiras variadas para clusterização de dois grupos. Isto parece indicar um efeito da globalização e uma concordância alta entre cultura ocidental. Dependendo da categoria, os países europeus se mostraram mais unidos, separados dos países americanos; em alguns casos a distinção não ficava clara. O uso de mais características e mais dados pode ajudar a determinar melhor o comportamento dos grupos e de cada categoria de modo individual.

Um fator que é candidato para responder esta similaridade é o alfabeto: o alfabeto latino é utilizado pelos países do grupo. Canadá, Estados Unidos, México, Grã-Bretanha, França, Alemanha tem suas línguas nacionais no alfabeto latino, e a Índia faz forte uso do inglês.

Diferente da primeira parte deste trabalho, os países apresentaram maiores variações e menores concordâncias internas. Muitos clusters deixavam países divididos em dois ou três grupos, o que leva a curiosidade de como estas divisões são feitas.

Porém, vale notar que esta menor consistência interna dos países era esperada até certo ponto. Gostos populacionais tendem a variar durante o ano por uma série de razões. Logo, é normal que países apresentem um gosto diferenciado ao passar do tempo. Por exemplo, vídeos de política perto de épocas de eleição devem ser mais vistos, em contraste a outras épocas. Isto levaria diferentes meses a terem diferentes grupos de perfis, mesmo dentro de um país.

5 CONCLUSÕES

Assim como esperado, os países mostraram possuir proximidade cultural também no que se refere a conteúdo de vídeos assistidos no YouTube. A suposição que fatores históricos, proximidade geográfica e culturas semelhantes afetam este aspecto cultural não podem ser confirmadas, pois o estudo não consegue apontar por que as distribuições das categorias ocorreram desta maneira; mas os grupos formados pelos países correspondem a estas ligações históricas e culturais entre eles, o que dá força a ideia.

O que podemos afirmar é que os países próximos entre si analisados por este estudo possuem preferência similar no conteúdo que eles assistem, e no que é popular dentro deles. Embora a Rússia seja incrivelmente extensa, ela é o país mais próximo da Coreia. Os países da América do Norte apareciam no mesmo grupo na grande maioria dos casos, assim como os europeus estavam no mesmo grupo em várias ocasiões.

A Índia apareceu como uma exceção desta ideia de proximidade geográfica, pois foi agrupada junto aos países europeus e norte-americanos em todos os casos, tendo alta similaridade de conteúdo. Outras razões aproximam o país aos demais, e isto não prejudica os resultados obtidos. Isto indica que outros fatores afetam preferência de conteúdo de vídeos, mas é evidente que muitas outras variáveis têm papel importante nestes agrupamentos. Uma boa suposição para a similaridade da Índia com os países europeus é a proximidade linguística, já que o inglês é amplamente utilizado no país.

Outra semelhança que apareceu bastante neste estudo foi a da Grã-Bretanha com os Estados Unidos e Canadá, principalmente no que se refere a distribuição no número de visualizações e likes nos vídeos principalmente. Mais uma vez não podemos apontar uma causa específica, mas o primeiro fator que vem a mente para a definição é a língua do país ser o inglês, mas outros estudos teriam de ser feitos para apontar a causa deste agrupamento.

Comparado a primeira parte deste estudo, também foram vistos menos consistências internas para um país, sendo que muitos tiveram meses se juntando em diferentes grupos. Até certo ponto isto era mais esperado neste estudo, pois alguns assuntos tendem a ter mais relevância dependendo doa época do ano em que estamos observando. Foi bastante interessante ver que os países tendiam a se agrupar em certas épocas do ano em alguns casos, o que sugere esta variação na relevância dos assuntos dependendo da época.

Os valores da métrica Silhouette obtidos neste estudo foram muito bons, um aumento comparado à primeira parte do estudo. Isto foi um aprimoramento que trouxe mais força ao estudo, e que espero que atraia atenção para novos estudos da similaridade de conteúdo entre os países.

6 REFERÊNCIAS

Trending on YouTube. YouTube Help Center. Disponível em: <https://support.google.com/youtube/answer/7239739?hl=pt-BR>. Acesso em 24 jun 2019

YouTube in numbers. YouTube for Press. Disponível em: <https://www.youtube.com/intl/en-GB/yt/about/press/>. Acesso em 1 jul. 2019

Clustering. Scikit Learn. Disponível em: <https://scikit-learn.org/stable/modules/clustering.html>. Acesso em 1 jul. 2019

AGGARWAL, Charu C., REDDY, Chandan K.. *Data Clustering: Algorithms and Applications*

An Introduction to Cluster Analysis for Data Mining. Disponível em: <https://www-users.cs.umn.edu/~hanxx023/dmclass/cluster_survey_10_02_00.pdf>. Acesso em 25 jun. 2019