Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Aplicadas
Departamento de Computação e Sistemas

Inferência de similaridades culturais usando dados extraídos de plataformas de propaganda baseadas em redes sociais

Gabriel Felipe Cordeiro Freire

TRABALHO DE CONCLUSÃO DE CURSO

ORIENTAÇÃO:
Filipe Nunes Ribeiro

COORIENTAÇÃO:
Alexandre Magno de Souza

Dezembro, 2019
João Monlevade–MG

Gabriel Felipe Cordeiro Freire

Inferência de similaridades culturais usando dados extraídos de plataformas de propaganda baseadas em redes sociais

Orientador: Filipe Nunes Ribeiro

Coorientador: Alexandre Magno de Souza

Monografia apresentada ao curso de Engenharia de Computação do Instituto de Ciências Exatas e Aplicadas, da Universidade Federal de Ouro Preto, como requisito parcial para aprovação na Disciplina "Trabalho de Conclusão de Curso II".

Universidade Federal de Ouro Preto

João Monlevade

Dezembro de 2019

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

F866i Freire, Gabriel Felipe Cordeiro .
Inferência de similaridades culturais usando dados extraídos de plataformas de propaganda baseadas em redes sociais. [manuscrito] / Gabriel Felipe Cordeiro Freire. - 2020.
57 f.: il.: color., gráf., tab., mapa.

Orientador: Prof. Dr. Filipe Nunes Ribeiro.
Coorientador: Prof. Me. Alexandre Magno de Souza.
Monografia (Bacharelado). Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Aplicadas. Graduação em Engenharia de Computação .

1. Computação - Pesquisa social. 2. Demografia. 3. Redes sociais on-line. 4. Internet - Aspectos sociais. I. Ribeiro, Filipe Nunes. II. Souza, Alexandre Magno de. III. Universidade Federal de Ouro Preto. IV. Título.

CDU 316.472.4

Bibliotecário(a) Responsável: Flavia C. M. Reis - CRB6-2431

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E APLICADAS
DEPARTAMENTO DE COMPUTACAO E SISTEMAS

FOLHA DE APROVAÇÃO

Gabriel Felipe Cordeiro Freire

Inferência de similaridades culturais usando dados extraídos de plataformas de propaganda baseadas em redes sociais

Membros da banca

Alexandre Magno de Souza (Coorientador) - Mestre - Universidade Federal de Ouro Preto
George Henrique Godim da Fonseca - Doutor - Universidade Federal de Ouro Preto
Gilda Aparecida de Assis - Doutora - Universidade Federal de Ouro Preto

Versão final
Aprovado em 19 de dezembro de 2020

De acordo

Filipe Nunes Ribeiro

Documento assinado eletronicamente por **Filipe Nunes Ribeiro**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 20/04/2020, às 15:03, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Rafael Frederico Alexandre**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 14/05/2020, às 13:09, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0050317** e o código CRC **F37AFE38**.

Referência: Caso responda este documento, indicar expressamente o Processo nº 23109.003381/2020-28 SEI nº 0050317

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35400-000
Telefone: - www.ufop.br

Dedico este trabalho à pessoa mais forte e magnífica que eu conheço, minha mãe, que mesmo com todas as dificuldades do mundo nas suas costas, deu sempre 1000 por cento de si para criar a mim e a meu irmão da melhor forma possível. Dedico também a meu pai, meu irmão, meus amigos, minha família, que me ajudaram nessa caminhada tão importante na minha vida. Dedico a Deus, que me provou que milagres existem sim, que me ajudou e me ajuda na minha jornada. Tudo isso junto foi o que me construiu.

Agradecimentos

Agradeço ao meu orientador Filipe Nunes Ribeiro, que aceitou me orientar e me sugeriu este tema, um profissional extremamente dedicado e competente que foi me guiando até o fim desta etapa final do curso. Agradeço também ao Alexandre Magno, tanto como coorientador como professor, outro exemplo de profissionalismo e dedicação. Com uma dupla dessas não tinha como eu querer mais na orientação deste trabalho, o meu muitíssimo obrigado aos dois. Agradeço também aos meus amigos Olimpio e Victor Lucas, vulgo Toruska, que ao longo do curso me inspiraram a ser mais organizado, e tentar fazer diferente às vezes, nas disciplinas, na faculdade, na vida.

"To see a world in a grain of sand, and heaven in a wild flower. Hold infinity in the palm of YOUR hand. And ETERNITY in an hour..."

— Willian Blake (1757 – 1827),
in: Auguries of Innocence.

Resumo

Hoje em dia, estudos de diversas áreas utilizam dados provenientes da internet para as mais diversas aplicações. Estudos sociais na área de dados online têm crescido nos últimos tempos, uma vez que essa fonte de dados é mais barata e o procedimento às vezes é mais prático do que a coleta de dados através de questionários presenciais. Como fonte destes dados online, o *Facebook* se mostra uma boa plataforma para extração de informações que podem ser usadas para diversos tipos de estudos sociais. Neste contexto, através de dados disponibilizados pelo *Facebook*, o presente trabalho pretende inferir semelhanças culturais entre 16 países de diferentes regiões do mundo utilizando metodologias de clusterização. Tais metodologias identificam quais países são mais próximos culturalmente entre si, tendo como base nas inferências que a rede social faz sobre cidadãos de cada local. Inicialmente estes grupos sociais são montados através de dados sobre comida e bebida, mesma metodologia usada em um trabalho que foi referência para a monografia aqui apresentada. Posteriormente os dados para inferências culturais foram expandidos para mais tipos de informações, como dados sobre política e religião, além de serem filtrados pelo nível de redundância que eles apresentaram. Sendo assim, teve-se como objetivo, comparar os resultados das inferências culturais deste trabalho que usa dados do *Facebook*, com os resultados de outros trabalhos, que usam fontes de dados diferentes. Os resultados mostraram boa correlação entre este trabalho e o trabalho do *World Value Survey* (WVS) que é realizado por um grupo mundial de cientistas sociais, em que este por sua vez em um dos seus estudos monta *clusters* culturais de vários países desde 1981. Por fim, tem-se que a estrutura resultante neste trabalho pode ser usada para outras pesquisas sobre diferenciação de grupos de usuários ou locais, sobre dados extraídos do *Facebook*.

Palavras-chaves: Computação Social. Demografia. Redes Sociais.

Abstract

Nowadays, studies from many areas use internet data to a variety of applications. Social studies on the online data field has expanded in the last years , once this source of data is cheaper and sometimes the proceedings is simpler if compared to data from surveys. As online data source, Facebook has shown to be a good data extraction point, that can be used to a diversity of social studies. In this context, by leveraging the Facebook data, the present work aims to infer cultural similarities among 16 countries from different spots around the world, using clusters. These clusters identify which countries are culturally closer to each other, based on Facebook inferences about its users preferences. Initially these social groups are built using data about food and drink, the same methodology used on a similar work, discussed later on. Later, the cultural inference data was expanded to more themes, like politics and religion, and in the same time they are filtered by its redundancy. In this case, the goal here is to compare cultural inferences from this work, with cultural inferences from another works that have different sources of data. The results showed some correlation between this work and WVS's work that is done by a world wide social scientists group. At last, the structure built here can be used for another researches about users or locations clustering based on Facebook data.

Keywords: Social Computing. Demography. Social Networks.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

CB *Census Bureau*

API *Application Programming Interface*

CDF *Cumulative Distribution Function*

FMA *Facebook Marketing API*

FMP *Facebook Marketing Platform*

GPS *Global Positioning System*

PCA *Principal Component Analysis*

WHO *World Health Organization*

WVS *World Value Survey*

Sumário

1 Introdução

1.1 O problema

Estudos demográficos dizem respeito a informações sobre uma sociedade em diversos aspectos, como aspectos econômicos, educacionais e políticos. Geralmente estes estudos têm como fonte dados oriundos de questionários e entrevistas pessoais presenciais feitas com um certo número de pessoas. Em um certo trabalho é apresentado um estudo social que mostra a influência de religião, engajamento político e nível de escolaridade sobre a democracia de alguns países, com o uso de questionários presenciais para obter as informações.(INGLEHART; WELZEL, 2010). Isso implica tempo e dinheiro, necessários para a realização de tal, porém hoje, com as redes sociais há a possibilidade do uso das suas plataformas de *marketing* para coleta de dados que podem ser usados para inferências demográficas. Segundo CRANSHAW et al. (2012, p. 59) dados de sistemas compartilhadores de locais estão se tornando cada vez mais disponíveis para pesquisadores [...] que estão encontrando novas formas de extrair várias informações sobre as relações de interações na sociedade. Neste contexto, muitas metodologias têm sido utilizadas para inferir dados demográficos dos usuários da internet, em particular, trabalhos recentes têm explorado plataformas de propagandas de redes sociais. ARAÚJO et al. (2017) apresentam um estudo que infere os números de pessoas que possuem algumas doenças, baseada em dados da internet e compara com dados de pesquisas estatísticas de órgãos de pesquisa renomados. Em Zagheni et al. (2018), dados da rede social *Facebook* são usados como base para inferir o número de pessoas nascidas no México que moram na Califórnia.

1.2 Solução proposta

Neste trabalho, foi explorada a plataforma de propaganda do *Facebook* para inferir características culturais de alguns países. Foi utilizada a *Application Programming Interface* (API) de *marketing* do *Facebook*, que permite seus usuários pesquisarem por dados demográficos sobre pessoas de quase todos os países do mundo. Foram escolhidos 16 países com intuito de se testar os resultados das formações de grupos culturais entre eles, sendo estes um subconjunto dos 60 países em um dos estudos no trabalho WVS (2019). Sequencialmente comparou-se os resultados do presente trabalho com os resultados em WVS, que usa de questionários presenciais para obter os dados destes países. Inicialmente a metodologia aqui adotada foi a mesma do trabalho de Silva et al. (2017), que escolheu os mesmos 16 países para os mesmos tipos de inferências culturais, porém fez do uso de dados pertencentes à rede social *Foursquare* em que os dados eram apenas sobre comida e

bebida. Por fim, no presente trabalho foram feitas análises adicionais, em que se incluiu o uso de dados relacionados também a política, educação, religião, entre outros. Ainda sobre as análises adicionais os dados foram também filtrados com base nas correlações altas entre alguns deles, já que foi possível identificar redundâncias nos mesmos.

1.3 Justificativa

O uso de dados online para pesquisas sociais tem como principais vantagens a coleta de dados barata, em que esta é também mais rápida e mais escalável. Tem-se também como outra vantagem a possibilidade da atualização dos dados com uma frequência maior. Já o estudo social em si deste trabalho e de trabalhos similares têm os objetivos de verificar o quão países se tornam similares nas questões culturais. Similaridades que se dão por conta dos impactos da imigração, política, religião, economia, educação e costumes que perduram muitos anos nas sociedades contra os novos costumes mais recentes. Por fim, para uma empresa ou grupo, as metodologias apresentadas aqui também podem ser usadas para descrição de grupos de consumidores dentro de mercados específicos. Dentro destes mercados é possível indicar quais locais são mais associáveis a perfis de interesses dos clientes de uma marca, justificativa também presente em Silva et al. (2017).

1.4 Objetivos

Este trabalho visa inferir similaridades ou diferenças culturais entre alguns países de regiões distintas do mundo, com base em dados da plataforma de marketing rede social *Facebook*.

Tem-se como objetivos específicos:

1. Utilizar a metodologia explicitada em Silva et al. (2017) para as inferências de similaridades culturais entre 16 países, para os separar em 7 grupos culturais distintos e propor uma metodologia para obtenção dos dados através do *Facebook*.

2. Viabilizar a coleta dos dados do *Facebook*, uma vez que a plataforma de marketing da rede social tem duas formas de coleta, a primeira é a forma manual, em que apenas um dado demográfico é coletado por vez pela interface web da plataforma. A segunda, é automatizada, através da API, que possibilita uma coleta mais rápida e robusta.

3. Comparar os resultados dos grupos culturais formados no presente trabalho, com os resultados em Silva et al. (2017) e WVS (2019).

1.5 Organização do trabalho

No Capítulo 2 são apresentados os principais trabalhos relacionados, tanto trabalhos que utilizaram dados de redes sociais, quanto o WVS, trabalho que utilizou dados de questionários presenciais. O Capítulo 3 mostra as metodologias e as tecnologias utilizadas durante o trabalho, assim como a metodologia do trabalho similar Silva et al. (2017), já que a primeira metodologia utilizada para inferências foi baseada neste trabalho. Na seção 3.2 são mostradas as tecnologias e a metodologia do presente trabalho, das primeiras análises de similaridade cultural, a subseção 3.3.3 mostra a segunda metodologia para inferências culturais e utiliza filtros por eliminação de dados redundantes, que possuem alta correlação. Logo após, no Capítulo 4, os resultados das metodologias das seções 3.2 e 3.3.3 são apresentados. Por fim, o Capítulo 5, contem as discussões e conclusão sobre o trabalho, assim como propostas para trabalhos futuros.

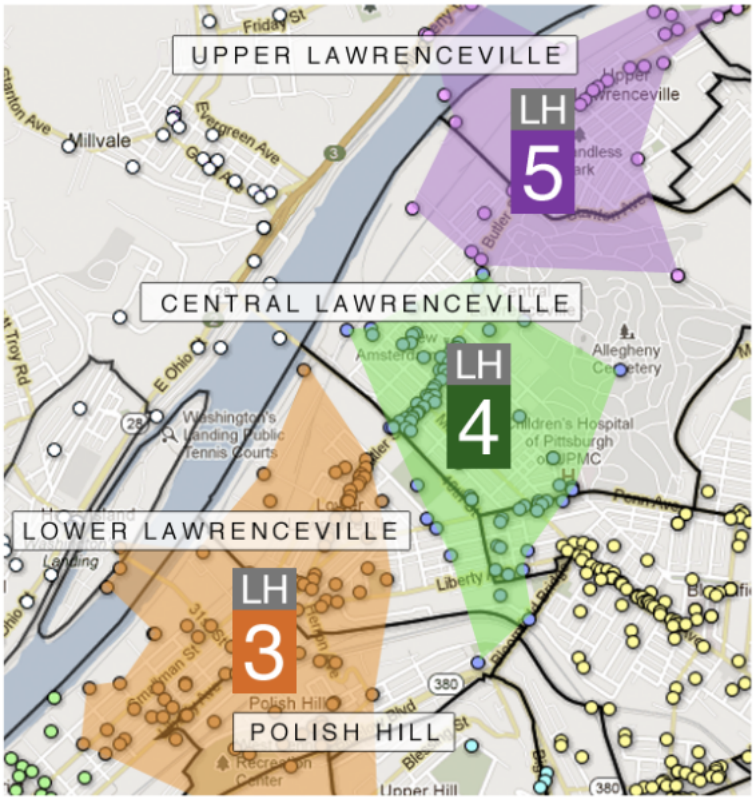
2 Revisão da literatura

Este capítulo apresenta os principais estudos correlatos ao o presente trabalho. A seção 2.1 apresenta estudos que inferem similaridades culturais sobre locais a partir de *check-ins* em plataformas de pesquisa e descoberta de locais. Na seção 2.2 é possível ver estudos relacionados, que utilizam a plataforma de marketing do *Facebook* para estudos sociais. Na seção 2.3 o trabalho WVS (2019) é apresentado, o qual faz a mesma clusterização cultural que o presente estudo, porém com dados oriundos de questionários respondidos presencialmente em vários países do mundo. A seção 2.4 contém as considerações sobre a revisão bibliográficas, faz também a comparação entre as propostas apresentadas com as propostas do presente trabalho.

2.1 Inferências de similaridades culturais com dados online

Os autores de CRANSHAW et al. (2012) apresentaram neste trabalho classificações de *clusters* comportamentais em regiões da cidade de *Pittsburgh*, na Pensilvânia, fazendo uso de dados da rede social *Foursquare*. Essa é uma rede social baseada em local, ou seja, o usuário vai até um estabelecimento e faz um *check-in* ali, portanto ele compartilha a informação de onde esteve. Sendo assim, eles usam 18 milhões de *check-ins*, de diversos estabelecimentos para verificar as preferências dos usuários da rede social, que moravam em algumas regiões da cidade, sobre quais tipos de locais e estabelecimentos os cidadãos costumavam frequentar. Com base nessas preferências, utilizaram uma metodologia para inferir as características similares e diferentes de algumas regiões para, por fim, delimitar fronteiras entre elas. Além dessas análises, os envolvidos neste trabalho entrevistaram 27 pessoas de diferentes locais da cidade para validar e comparar as informações das duas diferentes fontes de dados. O que foi percebido foi que as regiões culturais inferidas, tinham uma boa correlação com a demografia de cada bairro, como por exemplo, pela situação econômica dos moradores dentro das regiões, ou dominância étnica, entre outros. É possível ver algumas regiões culturais, definidas pelos autores na Figura 1. As bordas pretas na figura são os limites das vizinhanças intramunicipais oficiais e as áreas sombreadas coloridas são as regiões culturais inferidas. Por fim, os autores concluíram também que dos 3 bairros maiores na imagem, as fronteiras culturais, que são as fronteiras inferidas, se aproximaram das fronteiras intramunicipais, assim puderam apresentar a correlação entre os limites das vizinhanças e os limites culturais nestas áreas.

Figura 1 – *Livehoods*: *clusters* geográficos de cultura das vizinhanças.



Fonte: CRANSHAW et al. (2012).

Em Silva et al. (2017), o trabalho base para a metodologia de inferência de similaridade cultural do presente trabalho, também foram usados *check-ins* em estabelecimentos cadastrados na rede social *Foursquare*. Foram selecionados diferentes grupos de usuários que moravam em diferentes países e os dados de *check-ins* deles foram coletados por alguns dias. O intuito era no final agrupar os locais observados de acordo com as preferências destes usuários. De forma que os países em um mesmo grupo são considerados culturalmente próximos.

Os *check-ins* coletados em Silva et al. (2017), são de vários tipos de estabelecimento, como pode ser visto na Figura 2, que apresenta os nomes das categorias de estabelecimentos do lado esquerdo e alguns exemplos de sub categorias de estabelecimentos do lado direito. Entretanto o trabalho fez a escolha de usar apenas dados sobre comida e bebida para formar os grupos culturais.

A escolha por usar dados apenas sobre comida e bebida, ou ainda apenas as subcategorias da categoria "*Food*" vista na Figura 2, foi baseada em testes empíricos sobre os *check-ins*.

Antes de escolher usar dados apenas sobre comida e bebida, os autores, represen-

Figura 2 – Todas as categorias de estabelecimentos do *Foursquare* em 2014.

Table 1: Foursquare categories.

Name	Subcategories examples
Arts & Entertainment	Comedy Club, Movie Theater, Casino
College & University	College Lab, Fraternity House, Student Center
Residences	Home, Residential Building, Trailer Park
Professional & Other Places	Factory, Laboratory, Art Studio
Outdoors & Recreation	Baseball Field, Surf Spot, Park
Nightlife Spots	Bar, Rock Club, Nightclub, Strip Club
Shop & Service	Shoe Store, Nail Salon, Bike Shop
Food	Chinese Restaurant, Bakery, Pizza Place
Travel & Transport	Airport, Hotel, Pier
Event	Christmas Market, Festival, Parade

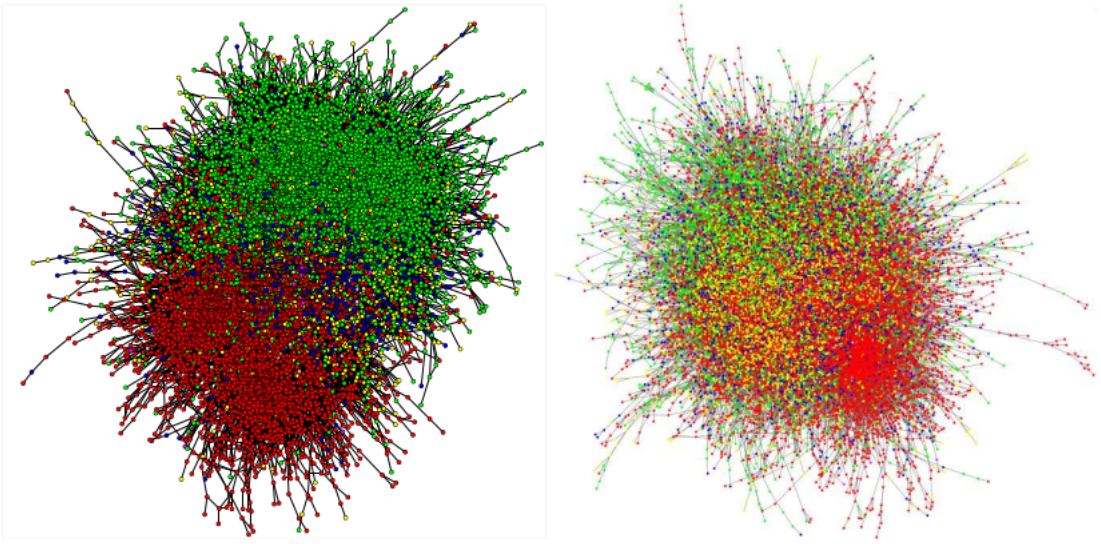
Fonte: Silva et al. (2017).

taram as informações coletadas com os *check-ins*, em forma de um grafo, em que cada vértice representa um grupo de usuários da rede social. Ainda no grafo, cada aresta não direcionada entre dois vértices, representa uma proximidade de preferências entre dois grupos de usuários. Os vértices da mesma cor, representam grupos de usuários do mesmo país, ou países geograficamente próximos, os vértices azuis por exemplo são grupos da Europa, e os amarelos da América Central e do Sul. O grafo do lado esquerdo, Figura 3 (a), foi construído com dados de *check-ins* apenas de estabelecimentos da categoria "*Food*", já o grafo do lado direito, Figura 3 (b), foi construído com dados de todas as categorias de estabelecimentos. Os autores verificaram que o grafo (a) tem mais ligações entre grupos geograficamente próximos, o que segundo eles pode indicar uma inferência melhor sobre similaridades culturais, uma vez que países geograficamente próximos, tendem a ter culturas mais próximas. Para reforçar as conclusões sobre o uso de dados de comida e bebida, eles citam o trabalho Cochrane e Bal (1990), como um dos que reforçam o poder de hábitos alimentares como fator de diferenciação cultural. Assim, a partir deste teste, os autores utilizaram apenas dados sobre comida e bebida para as futuras inferências.

Por fim, após as formações dos grupos culturais, o trabalho também comparou os seus resultados com os grupos culturais em WVS. O WVS, por sua vez, também forma *clusters* culturais entre países, e terá mais detalhes apresentados na seção 2.3.

Como o presente trabalho inicialmente usou esta mesma metodologia e faz comparações com Silva et al. (2017), informações mais técnicas sobre tal trabalho são apresentadas em 3.2.1, e os resultados nas seções 4.1.1 e 4.1.2.

Figura 3 – Rede de similaridade.



(a) Largest component of G_s^1 (b) largest component of G_s^2

Figure 4: [Better in color] Similarity network for the 0.65-network. Node colors: Africa (Pink), Asia (Red), Central and South America (Yellow), Europe (Blue), North America (Green), Oceania (Grey).

Fonte: Silva et al. (2017).

2.2 Utilização de plataformas de marketing de redes sociais para inferências de dados demográficos

As plataformas das redes sociais têm evoluído bastante e têm proporcionado ambientes que permitem aos anunciantes criar propagandas para certos grupos muito específicos de pessoas. Como as redes sociais têm acesso a um volume grande de dados dos seus usuários, elas conseguem inferir muitas informações sobre eles, como interesses em determinados jornais, determinadas comidas, figuras políticas, entre outros assuntos. Para permitir uma propaganda de qualidade, ou permitir que os anunciantes atinjam grupos bem específicos, estas plataformas permitem que os anunciantes escolham sua audiência, ou seja, seu público alvo. As plataformas permitem também que os anunciantes tenham acesso a vários filtros demográficos, como faixa etária, escolaridade, cidade, estado, país de residência, entre outros.

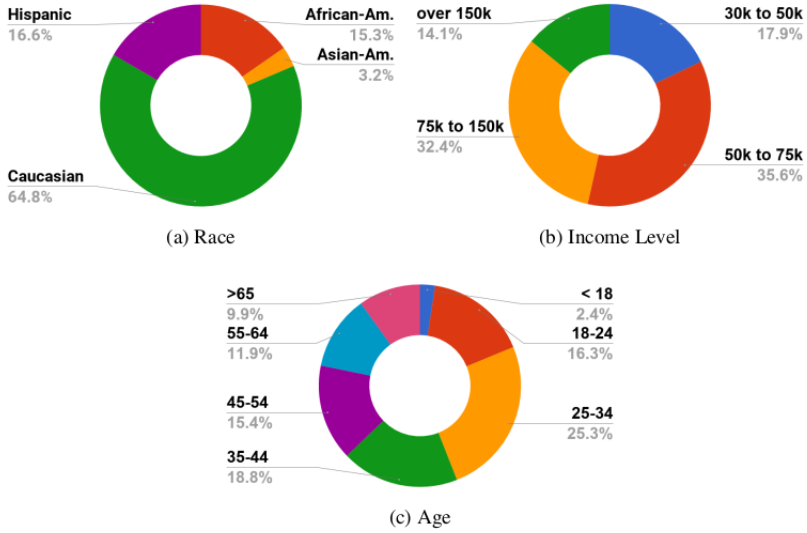
Sendo assim, trabalhos que se utilizam dessas ferramentas para obtenção de dados, utilizam o número de audiência, ou seja, o número de pessoas interessadas em determinados assuntos, para inferir alguma informação sobre isto. Neste contexto, por exemplo, se um anunciante quiser criar uma propaganda direcionada a pessoas que moram no estado de São Paulo e estão interessadas culinária italiana, ele pode filtrar o público alvo pelo local de residência, São Paulo, e que estão interessadas no interesse denominado "culinária italiana". A rede social pode, por exemplo, retornar que há 100.000 pessoas que moram em São Paulo que têm este interesse. Portanto, um trabalho que colete este dado numérico, pode tentar inferir algo a respeito da população de São Paulo sobre costumes italianos, em geral os trabalhos utilizam estes números de audiência como base para inferências.

Os trabalhos presentes nesta subseção utilizam a *Facebook Marketing API* (FMA), ferramenta da plataforma de propaganda do *Facebook* que permite a coleta programada de dados disponíveis para anunciantes. Os detalhes desses dados serão mostrados na subseção 3.3.1.

Na tese de RIBEIRO (2019) foram conduzidos diferentes estudos de caso com dados do *Facebook* nos Estados Unidos. Um destes estudos foi a utilização da rede social como um censo demográfico. Assim sendo, foram escolhidos sete atributos para que fossem mostrados as suas distribuições a nível nacional, estadual e municipal. Estes atributos são: distribuição por raça, idade, renda, nível escolar, opinião política e último país de residência. Esses valores de distribuição foram então comparados com dados de órgãos oficiais. Um dos achados dessas comparações foi que a distribuição por raça era muito similar para com a distribuição real deste atributo, em todos os níveis, nacional, estadual e municipal. Em contraste a isso, o nível de escolaridade inferidos pelos dados online eram maiores do que os das companhias estatísticas. De acordo com o autor, isso provavelmente se dá porque alguns usuários podem informar o nível de escolaridade errado. Ainda assim, os números de usuários com formação em pós graduação e ensino médio estavam de acordo com o real.

Na Figura 4 é apresentada uma amostra dos dados da rede social, coletados na tese em que é possível ver as distribuições por raça, renda e idade do usuários do *Facebook* que moram dos Estados Unidos.

Figura 4 – Distribuição demográfica dos usuários do *Facebook* nos Estados Unidos.



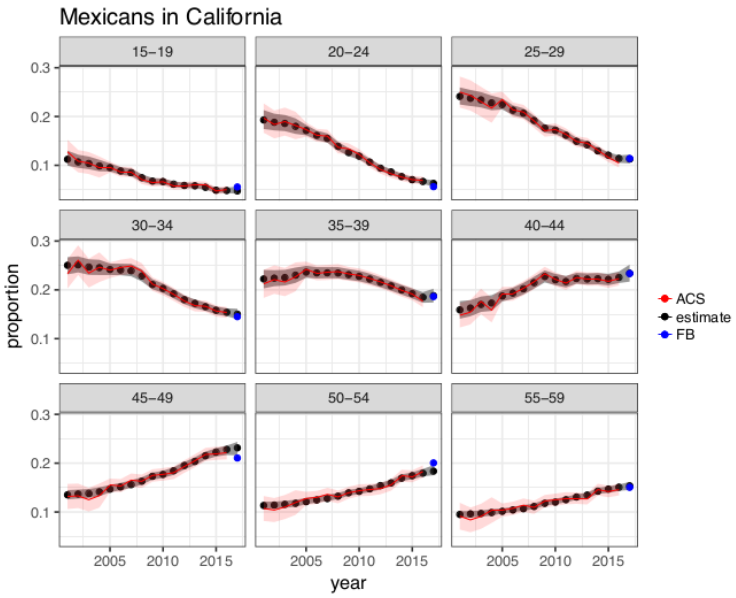
Fonte: RIBEIRO (2019, p. 30).

No trabalho de ARAÚJO et al. (2017), a FMA foi usada para verificar correlações entre o número de usuários do *Facebook*, interessados em assuntos relacionados a doenças ligadas a estilo de vida, como diabetes e obesidade, com os números estatísticos de portadores destas doenças. Foi descoberto que havia uma boa correlação entre os valores de audiência extraídos e os números reais destas doenças em alguns países selecionados. Assim, foi encontrada uma correlação razoável entre os dados da pesquisa e os dados estatísticos da *World Health Organization* (WHO) com relação aos números de diabéticos. Segundo o trabalho, apesar disso, outros tipos de doenças que estão menos ligadas a estilo de vida, como doenças hereditárias, tiveram correlações pequenas entre as duas fontes de informação demográficas. Um outro resultado destacado, foi por exemplo que o número de homens interessados em tabaco, no *Facebook*, é maior que o número de mulheres interessadas no mesmo assunto. Por fim foi concluído que essa relação de interesse no componente do cigarro, distribuída entre os dois sexos se alinha com dados estatísticos de que há mais fumantes do sexo masculino que feminino nos países observados.

Em Zagheni et al. (2018), é proposta uma mistura dos dados da plataforma de *marketing* do *Facebook*, com dados do *Census Bureau* (CB), principal agência governamental do sistema estatístico nos Estados Unidos. Isso foi proposto para que o resultado dessa combinação sejam informações mais confiáveis e atualizadas o possível. Combinou-se a taxa alta de atualização dos dados de uma rede social, com a confiabilidade de uma pesquisa estatística com um histórico longo dos dados. Foi usado um modelo hierárquico Bayesiano cujos resultados parciais pode ser visto na Figura 5. A faixa vermelha é o intervalo de

confiança dos dados do CB, a linha fina vermelha são os dados da agência. Os pontos pretos são o resultado do modelo hierárquico Bayesiano que combina os dados do CB e *Facebook*. O ponto azul é o valor apenas do *Facebook* no momento da medição dos dados em 2016. Por fim, o ponto mais à direita da cor preta, em cada gráfico, seria o resultado quase que em tempo real da predição Bayesiana. Cada imagem representa o número de pessoas nascidas no México que moram na Califórnia e que tem a idade dentro da faixa etária destacada no cabeçalho de cada gráfico. Por algum motivo, o valor coletado no *Facebook* se distancia mais do valor predito na faixa etária de 45-49, assim os autores dizem que esse valor seria menos confiável, já que não acompanha mais de perto a progressão histórica dos dados.

Figura 5 – Nascidos no México que moram na Califórnia.



Fonte: Zagheni et al. (2018).

2.3 World Value Survey

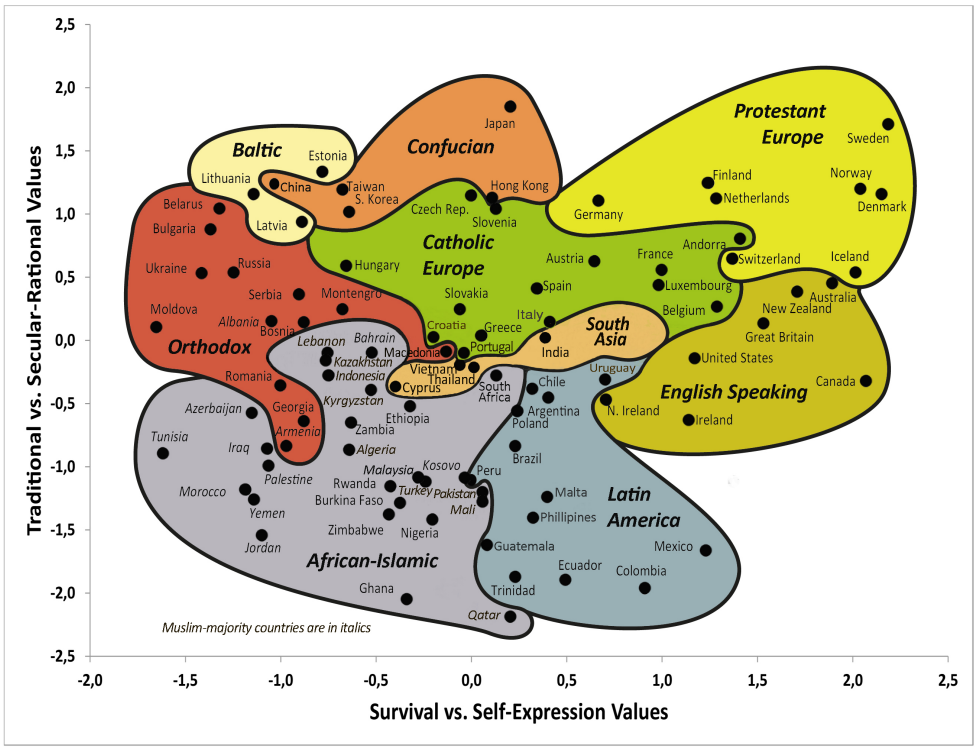
Segundo WVS (2019)[1], o *World Value Survey* é: " uma rede global de cientistas sociais, estudando valores culturais em mudança e seus impactos na vida social e política, liderado por um time internacional de estudiosos". O WVS faz várias análises com os dados coletados nos mais de 60 países alvo, como por exemplo inferir estatisticamente o quanto a população confia em religião, confia em políticos, quais valores acham mais importantes para os filhos, entre outros. Um pedaço do trabalho dessa associação é transformado em um mapa cultural. Este mapa, por sua vez, separa países do mundo todo em *clusters*, definidos

[1] http://www.worldvaluessurvey.org/WVSContents.jsp

pelos seus valores políticos, características econômicas, sociais, entre outros. Esses mapas são atualizados, aproximadamente, a cada 4 anos. Em cada um dos países pesquisados, um questionário presencial é utilizado por pesquisadores. Esse questionário é respondido por algumas pessoas e, a partir daí, as regiões culturais são inferidas. Na Figura 6 é possível ver o mapa mais recente do trabalho. É possível ver que alguns países geograficamente próximos, se encontram nos mesmos grupos, como Brasil, Argentina e Chile. Também é possível ver outros países que estão juntos no grupo cultural, porém distantes fisicamente, como Austrália e Estados Unidos.

O presente trabalho tem 16 países destes na Figura 6, eles se encontram nos 7 *clusters* denominados no WVS como, América Latina, Africano-islâmico, Ortodoxo, Confucionista, Falantes de língua inglesa, Europa Católica e Ásia do Sul.

Figura 6 – Mapa Cultural Mundial de 2014 pelo *World Value Survey*.



Fonte: WVS (2019).

2.4 Considerações finais

Neste capítulo foram apresentados alguns trabalhos que usam dados da rede social *Foursquare*, outros do *Facebook*, assim como um que utiliza dados de entrevistas presenciais. Destes trabalhos, os mais correlatos com o presente trabalho são Silva et al. (2017) e WVS (2019), pois se baseiam em agrupamentos de locais com critério de semelhança cultural. Assim foi possível verificar que os grupos formados nos três trabalhos são relativamente diferentes, porém isso pode mostrar características complementares uns sobre os outros. Sendo assim cada trabalho tem vantagens e desvantagens aparentes uns so. A vantagem do trabalho WVS (2019) é que nos seus questionários as perguntas não têm respostas binárias, elas são respondidas em um formato que permite o entrevistado ter uma liberdade maior para expressar o grau de concordância com as opções de resposta. A desvantagem do WVS é que cada país tem somente cerca de 1000 entrevistados. A vantagem do trabalho de Silva et al. (2017) é que ele tem milhões de *check-ins* como base de dados, podendo ser uma característica melhor em algum sentido se comparado ao WVS. Uma possível desvantagem que ele tem consiste no fato de que clientes que fazem os *check-ins* nos estabelecimentos não representam com muita precisão os cidadãos de cada país. O presente trabalho tem como vantagem o uso de dados sobre muitos usuários do *Facebook*, rede que atingiu aproximadamente 2,449 bilhões de usuários ativos mensalmente no mundo no fim de 2019 [2]. Porém o *Facebook* tem o mesmo problema que qualquer rede social: seus dados não representam completamente a população de um país. Além disso a maneira como o *Facebook* infere os dados sobre seus usuários não é aberta, portanto não é possível saber a sua precisão através da rede.

[2] https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

3 Metodologia

Este capítulo apresenta as metodologias utilizadas para inferência de similaridades culturais que utilizam da comparação de correlação por diferença de cosseno e por clusterização. A seção 3.1 detalha o funcionamento das ferramentas utilizadas para inferências. Na seção 3.2 são apresentadas a metodologias para inferências de similaridade e as características dos dados em Silva et al. (2017), uma vez que a sua metodologia foi usada como base em parte do presente trabalho. A seção 3.3 mostra a coleta de dados, características dos dados e metodologia que foi usada para as primeiras inferências de similaridades culturais dos países deste trabalho. Ainda na seção 3.3 é descrita a segunda coleta de dados, que abrange outros tipos de assuntos além de comida e bebida, que por fim mostra também uma abordagem para filtragem de dados redundantes, ou altamente correlacionados.

3.1 Ferramentas para inferências

Esta seção apresenta as 3 ferramentas principais para as análises e inferências das similaridades entre os 16 países. A subseção 3.1.1 apresenta a operação diferença de cosseno, a subseção 3.1.2 descreve brevemente o *Principal Component Analysis* (PCA), e a subseção 3.1.3 mostra a função do algoritmo *k-means*.

3.1.1 Diferença de cosseno

No presente trabalho, na primeira análise de similaridades, os países são comparados par a par, de forma que cada país é representado por um vetor. Cada par de países é então passado como parâmetro de entrada para uma função diferença de cosseno, cuja saída é um número que varia entre 0 e 1, em que 0 significa altíssima correlação e 1 significa baixíssima correlação. A diferença de cosseno é definida como sendo 1 - X, em que X é a similaridade de cosseno, que por sua vez é uma operação matemática que permite medir o quão dois vetores de n dimensões são próximos, com base na direção de cada vetor. Como esta técnica usa o cosseno do ângulo, caso os dois vetores tenham um ângulo nulo entre eles, o valor da similaridade de cosseno resulta no valor 1, de forma que a diferença de cosseno resulta no valor zero. A representação equacional da diferença de cosseno é apresentada na Equação 3.1.

$$1 - \cos(v1, v2) = 1 - \frac{\sum_{i=1}^{n} v1_i v2_i}{\sqrt{\sum_{i=1}^{n} (v1_i)^2}\sqrt{\sum_{i=1}^{n} (v2_i)^2}}$$

(3.1) : Diferença de cosseno entre os vetores, ou países, v1 e v2. Fontes: Elaborado pelo autor e (DANGETI, 2017, p. 281).

3.1.2 *Principal Component Analysis* (PCA)

O PCA é uma técnica de transformação dos dados que permite reduzir as dimensões de um conjunto de informações. Segundo Dangeti (2017) "Entender a estrutura de dados com centenas de dimensões pode ser difícil, portanto, reduzindo as dimensões para 2D ou 3D, observações podem ser visualizadas facilmente" e segundo Jain (1991) "o PCA produz um grupo de valores resultantes ... em que estes valores provêm a máxima discriminação das novas dimensões". Logo, o PCA serve para discriminar melhor os dados, levando a uma melhor clusterização, além disso serve para visualização dos dados com mais de 3 dimensões com o mínimo de perda possível de informações sobre os mesmos.

O trabalho Silva et al. (2017) utiliza o PCA para auxiliar em uma melhor clusterização dos países, já que os dados tem muitas dimensões, cada país é representado por um vetor de 101 dimensões. O presente trabalho tem vetores que variam de 95 até 10.282 dimensões, assim o PCA também foi utilizado.

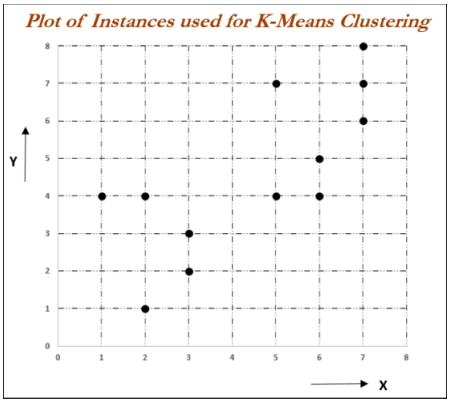
3.1.3 Algoritmo *K-means*

Como parte final da metodologia de inferências de similaridades culturais, foi preciso agrupar os países por afinidade cultural, deixou-se os que têm culturas mais próximas, nos mesmos *clusters*. Dado um conjunto de dados, pontos/vetores, de n dimensões, o algoritmo *k-means* tem como entradas principais os dados a serem agrupados e o número k desejado de grupos a serem formados. A saída do algoritmo é a classificação de cada ponto/vetor como pertencente a um dos k grupos calculados por ele.

Na Figura 7 é exemplificado um conjunto de dados de duas dimensões em um espaço cartesiano, na Figura 8 é mostrado o resultado do *k-means* com $k = 2$, ou seja, dois grupos são formados. Um grupo está em vermelho, o outro na cor verde.

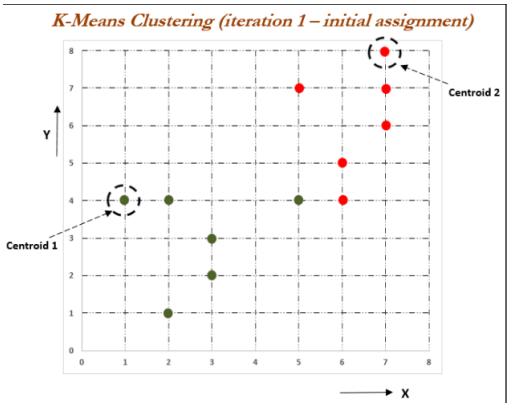
Após o uso do PCA, o *k-means* foi aplicado tanto no presente trabalho, quanto em Silva et al. (2017), em cada vetor correspondente aos países. Isso significa que, cada local foi representado por um ponto de n dimensões, e depois classificado como pertencente a algum grupo.

Figura 7 – Conjunto de pontos antes da clusterização com *k-means*.



Fonte: Dangeti (2017, p. 307).

Figura 8 – Conjunto de pontos agrupados com *k-means*, $k = 2$.



Fonte: Dangeti (2017, p. 309).

3.2 Metodologia e dados do trabalho base

Neste trabalho pretende-se reproduzir a metodologia utilizada em Silva et al. (2017), para inferências culturais, foram comparados os resultados desta referência com os resultados do presente trabalho.

Pretende-se comparar os resultados aqui com os dados do trabalho em WVS (2019) do mapa cultural mundial de 2008. Como os dados deste trabalho e do trabalho base são de duas redes sociais diferentes, o aspecto qualitativo e quantitativo destes dados são diferentes. Estas características dos dados do trabalho base serão mostradas nas subseções 3.2.1 e 3.2.2.

3.2.1 Dados do *Foursquare*

A rede social *Foursquare* é do tipo de compartilhamento de local, assim um usuário da rede, usa o *Global Positioning System* (GPS), vai até um estabelecimento cadastrado na rede, e pode fazer um *check-in* no local, ou seja, mostrar que ele estava lá. Este *check-in* é então compartilhado no *Twitter*, e pode ser coletado pelo próprio *Foursquare*. Os estabelecimentos cadastrados são classificados como pertencentes a uma categoria e subcategoria. Um exemplo de categoria é a denominada de *"Food"*, que tem diversas subcategorias, dentre elas a classificada como *"American Restaurant"* e a *"Brazilian cuisine"*. Todas as categorias e suas derivadas podem ser acessadas no site da *Foursquare* para desenvolvedores [1]. Se um usuário faz *check-in* em um estabelecimento classificado como "*Italian cuisine*", por exemplo, segundo Silva et al. (2017), isso indica um interesse, por parte do usuário, nesse tema. Logo, se por meio dos dados da rede social é possível extrair conjuntos de preferências e associar as pessoas à locais, é possível inferir similaridades/diferenças culturais entre locais.

3.2.2 Metodologia

A metodologia apresentada em Silva et al. (2017) escolheu 101 subcategorias de estabelecimentos da categoria *"Food"*, para se verificar o número de *check-ins* nos estabelecimentos destas subcategorias, em alguns países. São escolhidos 16 países para clusterização e comparação com o mapa cultural mundial de 2008 do WVS. Os 16 países são apresentados na Tabela 1.

Tabela 1 – Os 16 países das análises de similaridade cultural.

Nome do país
Argentina
Austrália
Brasil
Chile
Inglaterra
França
Indonésia
Japão
Coreia do Sul
Malásia
México
Rússia
Singapura
Espanha
Turquia
Estados Unidos

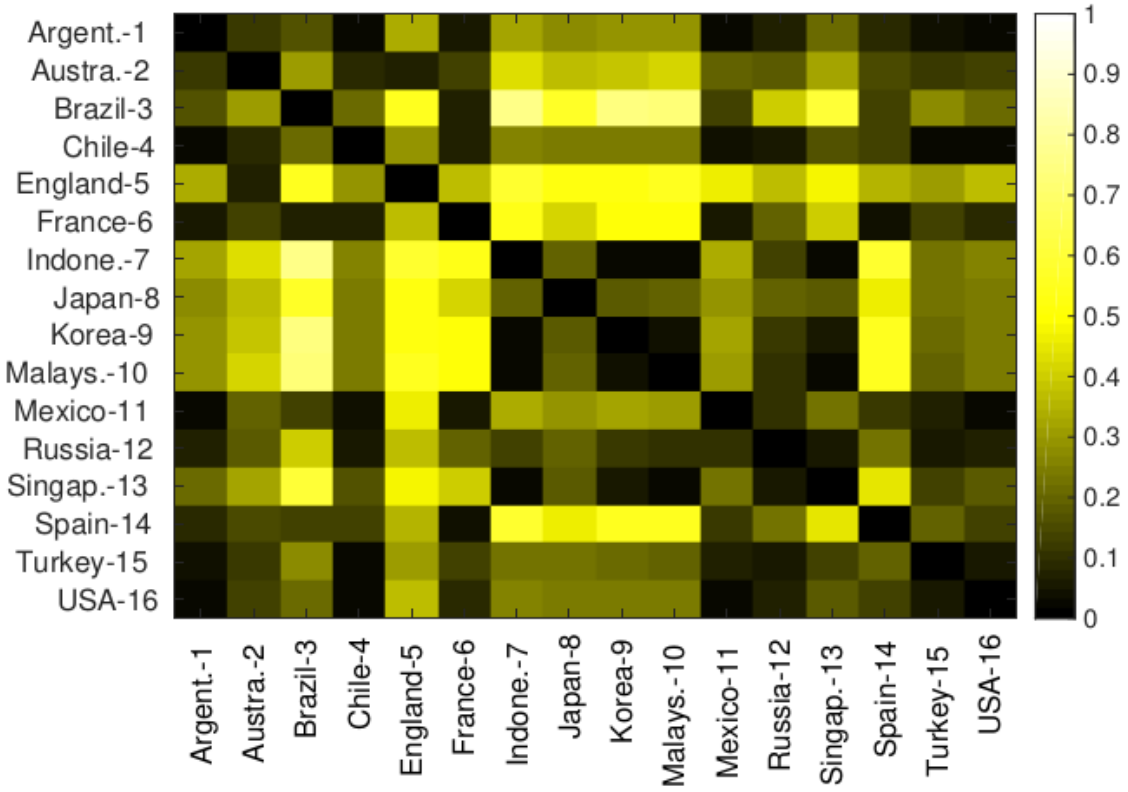
Desta forma são obtidos 16 vetores, com 101 colunas cada, de forma que cada coluna representa o número de *check-ins* feitos em um determinado tipo de estabelecimento. Cada

[1] https://developer.foursquare.com/docs/resources/categories

valor de *check-in* é o somatório do número de *check-ins* nos países no período de 19 dias. Foram usados pouco mais de 2 milhões de *check-ins*, de quase 1 milhão de usuários diferentes, em aproximadamente 600 mil estabelecimentos distintos . Logo, um vetor referente a um país "a" se dá da seguinte forma $C^a = c_1^a, c_2^a..., c_{101}^a$ de forma que c_1^a representa o número de *check-ins* na subcategoria "1", da categoria "*Food*", de estabelecimentos no país "a", do período da coleta. Após a coleta de todos os dados, estes foram normalizados, assim um vetor de um país "a" é representado por $F^a = f_1^a, f_2^a..., f_{101}^a$ em que $f_1^a = c_1^a / máx(C^a)$.

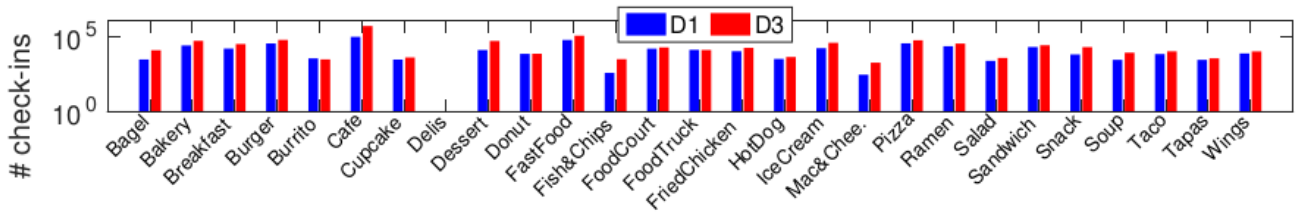
Após os dados serem transformados, cada um dos vetores de 101 subcategorias são divididos em 3 subvetores, $F_{Drink} = f_1^a, ...f_{21}^a$, $F_{FastFood}$ $f_{22}^a, ...f_{48}^a$, e $F_{SlowFood} = f_{49}^a, ...f_{101}^a$. O conjunto dos dados chamado de *Drink*, contém as subcategorias referentes apenas a bebidas, *FastFood* apenas a subcategorias relacionadas a *Fast-food* e comidas prontas; por fim, *SlowFood* às subcategorias de comidas regionais, como por exemplo "Culinária Brasileira". Esta divisão dos dados nestes 3 subconjuntos foi feita com intuito de fazer uma pré análise com mapas de calor, antes de agrupar os países. Essa pré análise foi baseada na diferença de cosseno, então os vetores dos 16 países foram comparados 2 a 2, o que resultou em uma matriz, ou mapa de calor, em que os valores mais altos (com células de cor clara) representam menos similaridade entre dois países, e valores mais baixos (com células de cor escura), o inverso. As análises foram feitas separadamente para cada subvetor. Na Figura 9 por exemplo, é possível observar o mapa de calor gerados pelos pares de diferença de cosseno na qual cada célula da matriz representa um valor resultado da diferença de cosseno entre o país da linha e da coluna. É possível observar por exemplo que a célula da linha Austrália e coluna Inglaterra tem uma cor escura, o que significa a inferência de uma forte proximidade cultural, o que condiz com resultados de outras fontes apresentadas no penúltimo capítulo deste trabalho. Na Figura 10 são mostrados os valores de *check-in* para cada uma das subcategorias de estabelecimentos contidos no subvetor *FastFood*. Os dados utilizados em Silva et al. (2017) são apenas os valores das barras vermelhas. Os valores das barras azuis são de uma base de dados que teve um período de falha em coletas de dados e por isso esta base não foi usada para inferências.

Figura 9 – Mapa de calor por diferença de cosseno do subvetor *Drink*.



Fonte: Silva et al. (2017, p. 19).

Figura 10 – Números de *check-ins* para cada subcategoria dentro do subvetor *FastFood*.



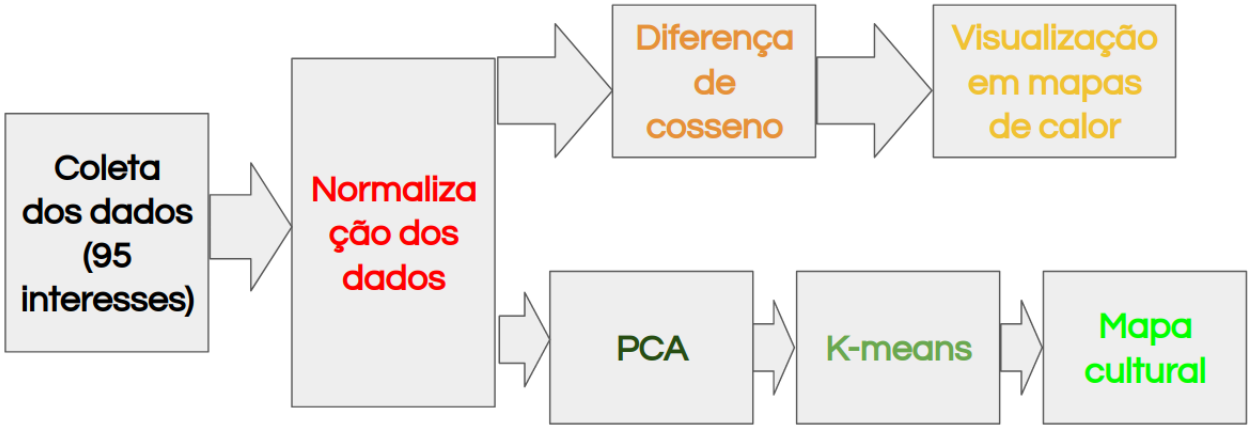
(b) Fast Food

Fonte: Silva et al. (2017, p. 14).

Por fim, foram formados *clusters* para a comparação com o trabalho em WVS (2019) que utilizou a PCA e *k-means* para a clusterização. O parâmetro "k" do *k-means* tem valor 7 na análise principal, pois no WVS, os 16 países escolhidos por Silva et al. (2017) no mapa mundial cultural, estão separados em 7 *clusters*.

Para sintetizar os passos feitos pelos autores de Silva et al. (2017), a Figura 11

Figura 11 – Fluxograma simplificado da metodologia em Silva et al. (2017).



mostra um fluxograma simplificado até as metodologias de análise de similaridade.

3.3 Metodologia e dados do presente trabalho

A fim de comparar a metodologia em Silva et al. (2017), porém com os dados do *Facebook*, primeiro foi preciso mapear as 101 subcategorias de estabelecimentos do *Foursquare*, em dados correlatos na rede social *Facebook*.

3.3.1 Coleta dos dados

Os dados utilizados vêm da *Facebook Marketing Platform* (FMP), que possibilita a coleta do número aproximado de pessoas que estão interessadas em um determinado assunto. Ela geralmente é utilizada por anunciantes, para que eles possam verificar o tamanho do público alvo de algum anúncio. Assim como a rede social *Foursquare* tem uma lista com categorias de estabelecimentos, o *Facebook* tem uma lista com palavras ou frases, que representam interesses que podem ser filtrados por local, idade do público alvo, escolaridade, entre outros. Nas figuras 12, 13, 14 e 15 é possível ver alguns elementos da interface web da plataforma de marketing. Na Figura 12, é apresentada a possibilidade da escolha do local de onde se quer obter as informações. A Figura 14 mostra alguns dos filtros principais do público alvo, que são: a faixa etária, gênero e idioma falado. O dado de entrada principal é preenchido no campo mostrado na Figura 15, que é o nome do interesse que se quer obter a informação, que pode ser uma palavra, ou frase. Por fim, na Figura 13, é mostrada a saída, ou seja, o número aproximado de pessoas que moram naquele local que estão interessadas no assunto e que passaram nos filtros demográficos.

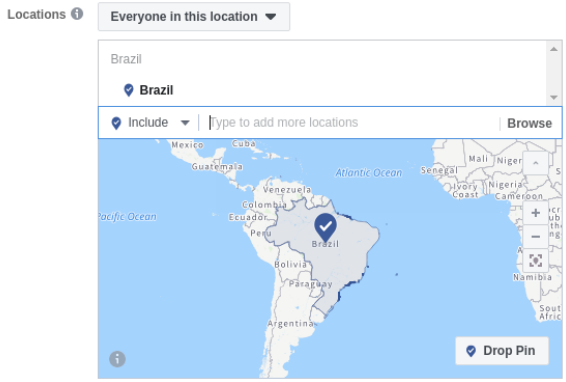


Figura 12 – Interface web da FMP: escolha de local.

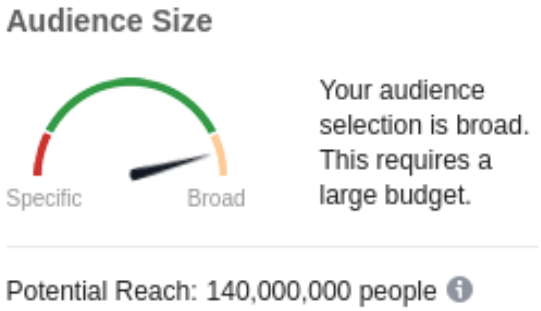


Figura 13 – Interface web da FMP: valor de audiência.

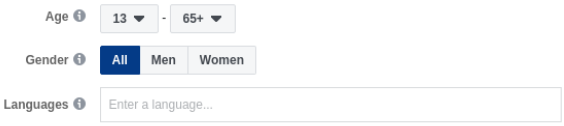


Figura 14 – Interface web da FMP: filtros demográficos básicos.

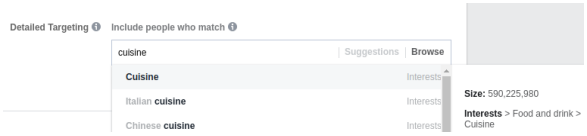


Figura 15 – Interface web da FMP: interesses para filtragem.

Apesar da possibilidade de usar a interface web da FMP para a coleta de dados, para uma coleta muito grande, isso é inviável, logo, foi utilizada a API da FMP (FMA), permitindo coletas por software. A FMA só permite coletar os dados dos interesses através de identificadores ao invés dos nomes, sendo que cada interesse tem um identificador único de 13 dígitos.

Para comparar *clusters* que usaram dados das 101 sub categorias de estabelecimentos no trabalho Silva et al. (2017), foram procurados interesses no *Facebook* em que os nomes eram correspondentes aos nomes das 101 subcategorias. Foram encontrados interesses com nomes correspondentes a 95 das 101 subcategorias, segundo RIBEIRO (2019): "...interesses ausentes podem ocorrer porque o *Facebook* não tem interesse em criar ou manter os mesmos". Assim as primeiras análises foram feitas com os 95 interesses correspondentes encontrados. Para achar os identificadores de cada interesse, foi utilizada a plataforma FacebookInterestsExplorer (2019), que foi criada como parte do trabalho de RIBEIRO (2019). Nela é possível pesquisar um interesse na sua forma textual e obter o identificador correspondente. Na Figura 16 é possível ver um pedaço da página web que permite pegar o código de um interesse [2], o identificador de 13 dígitos é retirado do link da página, no exemplo, para o interesse *" Brazilian cuisine"* o identificador é "6003460935625". É possível ver também na Figura 17, alguns interesses e seus respectivos identificadores, entre os 95 mapeados.

Após mapear os interesses, a FMA foi instalada para o interpretador da linguagem

[2] http://blackbird.dcc.ufmg.br/interest$_s$tudy/app.php?query = 6003259680957name =

Figura 16 – Pedaço de página do site *"Facebook Interests Explorer"*.

Figura 17 – Alguns interesses mapeados e seus identificadores.

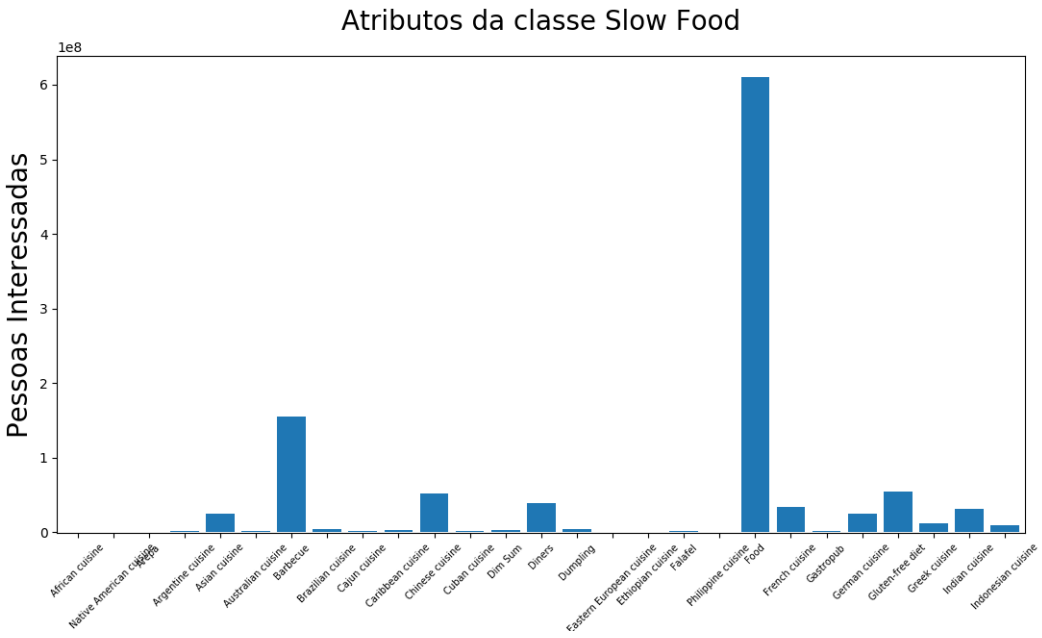
```
['Brazilian cuisine', '6003460935625'],
['Cajun cuisine', '6003326947654'],
['Caribbean cuisine', '6003706238946'],
['Chinese cuisine', '6003030029655'],
['Cuban cuisine', '6003381728404'],
['Dim Sum', '6003353270460'],
['Diners', '6003243058188'],
['Dumpling', '6003287254476'],
['Eastern European cuisine', '6003836415252'],
['Ethiopian cuisine', '6003372392375'],
['Falafel', '6002910631572'],
['Philippine cuisine', '6003214636861'],
['Food', '6003266061909'],
['French cuisine', '6003420024431'],
```

Python, versão 2.7, que era a versão mais estável da API, até o momento desta primeira parte do trabalho (início de 2019). Assim todos os dados foram coletados para os 16 países e armazenados em disco. Os números de pessoas interessadas em alguns interesses do subvetor *SlowFood*, pode ser visto na Figura 18, de forma que cada valor apresentado na figura é o somatório das audiências dos 16 países. Para cada assunto específico, a escala vertical tem como valor máximo 10 milhões de pessoas. É possível observar nesta última figura que o interesse de maior valor sobre todos os países é o interesse denominado "*Food*". Na próxima seção, serão mostrados outros dados sobre este interesse em específico.

3.3.2 Reprodução da metodologia do trabalho base

Similar à estrutura dos dados em Silva et al. (2017), um país "a" pertencente aos 16 países escolhidos tem um vetor correspondente de 95 colunas. Logo $I^a = i_1^a, i_2^a ..., i_{95}^a$, em que i_1^a corresponde ao número de pessoas interessadas no assunto "1" que moram no

Figura 18 – Números totais de pessoas interessadas em uma parte do subvetor *SlowFood*.



país "a". Também da mesma forma os dados foram normalizados, portanto um vetor de um país "a" passou a ser representado por $F^a = f_1^a, f_2^a..., f_{95}^a$, em que $f_1^a = i_1^a / máx(I^a)$.

Na Tabela 2, são mostrados os valores de audiência, após normalização, de alguns países. Na primeira coluna à esquerda temos os nomes dos países em duas linhas seguidas, a primeira linha mostra os valores normalizados dos 5 primeiros interesses, entre os 95, de maior valor. A segunda linha com o nome repetido do país, mostra os nomes dos interesses de maior valor de cada país, que têm seus valores correspondentes sempre uma célula acima. É possível identificar que o interesse "*Food*" é o de maior valor em todos os países, por conta disso, e pelo nome ser algo que aparente ser muito genérico, no Capítulo 4 os grupos com e sem este atributos foram comparados. Como curiosidade, dá para visualizar também que os interesses "*Pizza*" e "*Desserts*" (sobremesas), são alguns dos interesses mais populares entre os países.

Após os dados coletados e normalizados, mapas de calor com diferença de cosseno foram construídos e comparados com os do trabalho Silva et al. (2017), também dividindo os vetores principais nos subvetores *Drink*, *FastFood* e *SlowFood*. Tais mapas podem ser observados na subseção 4.1.1.

Em paralelo aos mapas de calor, os dados receberam uma transformação com a técnica PCA, por fim, sendo aplicado ao *k-means* para a clusterização, com $k = 7$. O valor de k foi o mesmo valor principal escolhido em Silva et al. (2017). Os resultados podem ser vistos na subseção 4.1.2, nas figuras 44 e 45.

Tabela 2 – Os interesses de mais altos valores, normalizados, em alguns dos 16 países.

País	1º maior	2º maior	3º maior
Argentina	1	0,417391304347826	0,41304347826087
	Food	*Pizza*	*Desserts*
Austrália	1	0,430769230769231	0,415384615384615
	Food	*Desserts*	*Barbecue*
Brasil	1	0,419354838709677	0,376344086021505
	Food	*Pizza*	*Desserts*
Chile	1	0,373737373737374	0,353535353535353
	Food	*Pizza*	*Desserts*
França	1	0,5	0,329166666666667
	Food	*Tea*	*Desserts*
Grã Bretanha	1	0,375	0,34375
	Food	*Desserts*	*Veganism*
Indonésia	1	0,263157894736842	0,25
	Food	*Tea*	*Desserts*
Japão	1	0,438888888888889	0,361111111111111
	Food	*Coffeehouse*	*Japanese cuisine*

3.3.3 Demais análises e testes

Nesta segunda aproximação, os dados não são mais os 95 interesses relacionados a comida e bebida, e sim variações de subgrupos de dados, que juntos são 10.282 interesses. Foi coletada, para todos os 16 países, uma parte, dos mesmos interesses coletados no trabalho Speicher et al. (2018), que coletou aproximadamente 240.000 interesses do *Facebook*, em que se usou uma técnica de "bola de neve". Essa técnica consiste em informar à rede social um interesse qualquer e obter como resposta os interesses relacionados a ele. A partir desses novos interesses relacionados, também pedir os interesses relacionados a cada um deles, se faz isso recursivamente, desta forma é construida uma espécie de árvore cada vez mais profunda e larga, de interesses coletados.

Após esta coleta, alguns países apresentaram números maiores de interesses do que outros. Os Estados Unidos por exemplo teve 175.712, já a Argentina 73.813. Dessa maneira foi preciso filtrar os interesses em comum nos 16 países, deixou-se apenas a intersecção entre os atributos, restando em 23.329 deles. Antes de prosseguir, os dados foram filtrados mais uma vez, foram retirados os interesses que tinham valores abaixo de 10.000 pessoas interessadas, em pelo menos 1 país, o que resultou no valor final de 10.282 interesses. Assim, cada um dos 16 países possui os mesmos 10.282 interesses, em que todos eles têm valores acima de 10.000 pessoas interessadas. Dessa vez, antes de prosseguir com a metodologia para clusterização, foi feita uma verificação de normalidade dos dados, fez-se o teste máx/min dos valores de audiência de todos os países juntos, concluindo que os dados não eram normalmente distribuídos, pois a razão máx/min resultou em um valor com mais de duas ordens de grandeza. Isso implica que os dados precisam de uma transformação

que diminua a diferença geral entre eles, dessa forma a transformação logarítmica na base 10 foi utilizada, que consiste em aplicar a cada um dos valores de audiência dos 10.282 interesses, o logaritmo na base 10. Segundo Jain (1991) "...A transformação logarítmica é útil apenas se a razão máx/min é grande...". Como a distribuição dos dados não é do tipo normal, isso significa que há muitos interesses correlacionados entre eles, o que pode dificultar uma boa clusterização. O *Facebook* inclusive classifica cada um dos interesses como pertencente a um tópico. Todos os 15 tópicos contidos nos dados usados podem ser vistos na Tabela 3.

Tabela 3 – Tópicos e seus respectivos números de interesses.

Nome do tópico	**Número de interesses correspondentes**
News and entertainment	2117
None	42
Shopping and fashion	400
Lifestyle and culture	563
Sports and outdoors	540
Technology	315
-	533
Fitness and wellness	218
Hobbies and activities	1074
Travel, places and events	925
Business and industry	1896
Food and drink	514
Family and relationships	51
Education	229
People	865

Após a transformação logarítmica, a fim de diminuir a alta correlação que atrapalha uma boa clusterização, primeiro foi verificada a distribuição cumulativa de correlação entre os dados, de forma que para cada tópico foi calculada a correlação de *spearman* par a par dos interesses. Esta correlação classifica dois vetores de dados como tendo correlações fortes ou fracas. Sua saída é um valor entre -1 e +1, em que valores próximos de zero representam pouca correlação e valores altos, em módulo, alta correlação. Foi usada uma *Cumulative Distribution Function* (CDF) para a visualização das distribuições de correlação para cada tópico, com objetivo de verificar o quanto, cada valor dos coeficientes de *spearman*, representa nos dados, em termos de porcentagem. Foi verificado que em todos os 15 tópicos, os coeficientes com valores acima de 0.5 ou abaixo de -0.5, ou seja, correlação razoável ou alta, sempre representam entre 60% e 70% dos dados, isso reforça que a distribuição não é normal. Na Figura 19 é apresentada a CDF dos coeficientes de *spearman* entre os interesses do tópico *News and entertainment*. Essa figura mostra que por volta de 70% dos dados do tópico têm coeficientes de *spearman* com módulo acima de 0.5, e isso não é muito diferente nas CDF's dos outros 14 tópicos, expostos da Figura 20 até a Figura 33.

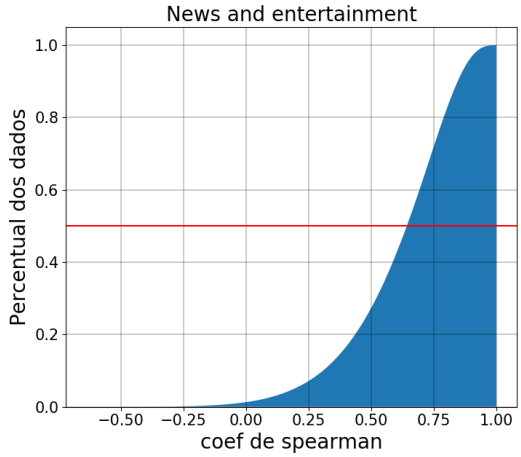


Figura 19 – CDF dos coeficientes de *spearman* dos interesses do tópico *News and entertainment*.

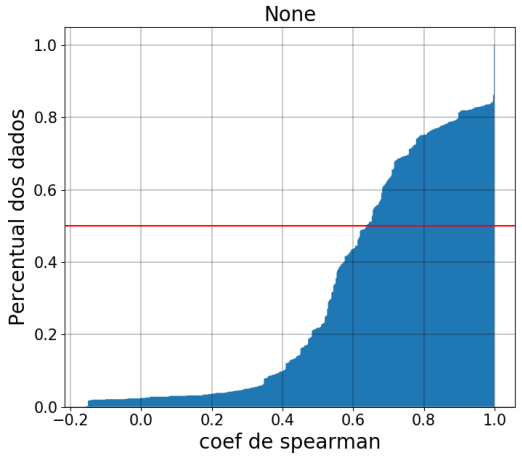


Figura 20 – CDF dos coeficientes de *spearman* dos interesses do tópico *None*.

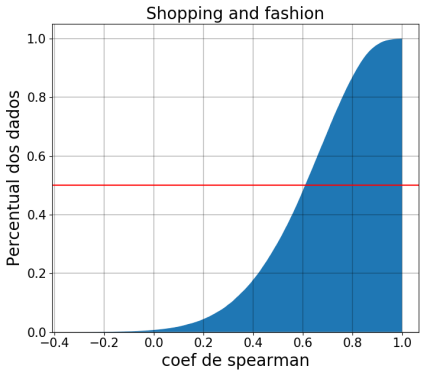


Figura 21 – CDF dos coeficientes de *spearman* dos interesses do tópico *Shopping and Fashion*.

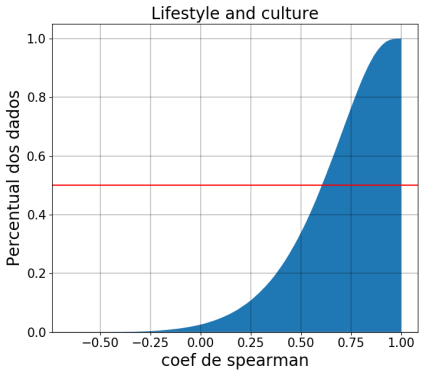


Figura 22 – CDF dos coeficientes de *spearman* dos interesses do tópico *Life and Culture*.

Após a verificação da correlação alta, foi feita uma eliminação de redundância, assim como em Sousa, Almeida e Figueiredo (2019). Porém utilizou-se um critério exploratório, sendo este baseado em interesses que têm maior número de correlações altas com outros interesses que estão dentro do mesmo tópico.

Para este filtro de redundância, o primeiro passo foi representar os dados como grafos. Assim, cada tópico equivale a um vértice e, caso um interesse "m" tenha com o interesse "n" um coeficiente de *spearman* com módulo maior ou igual a 0.5 por exemplo, então terá uma aresta não orientada entre estes dois vértices. Por conseguinte, com a análise dos graus de conectividade dos vértices é possível identificar os interesses que tem correlação igual ou acima do valor estabelecido e que estão conectados com muitos outros interesses. Isso significa que se um vértice, interesse, tem grau muito alto no grafo, logo ele tem uma correlação alta com uma parte grande do grafo. Assim podendo indicar que

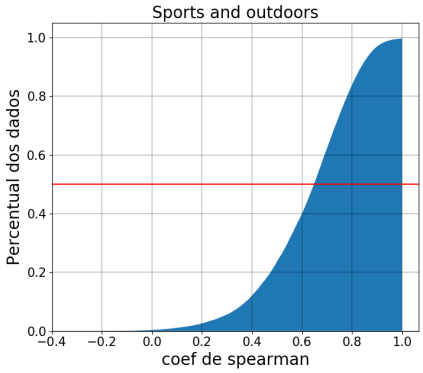


Figura 23 – CDF dos coeficientes de *spearman* dos interesses do tópico *Sports and Outdoors*.

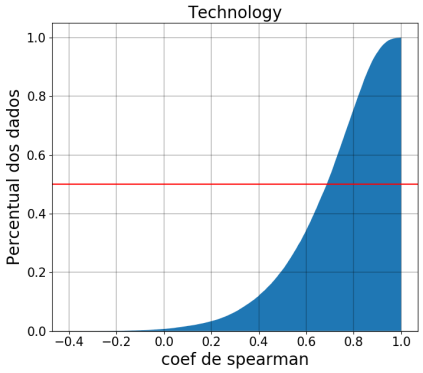


Figura 24 – CDF dos coeficientes de *spearman* dos interesses do tópico *Technology*.

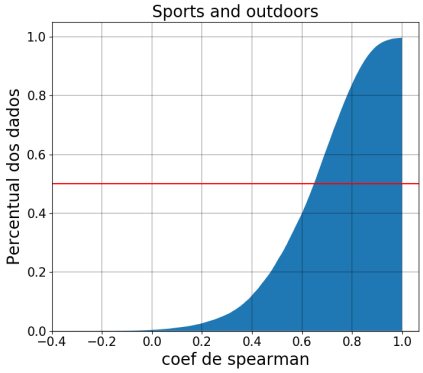


Figura 25 – CDF dos coeficientes de *spearman* dos interesses do tópico -.

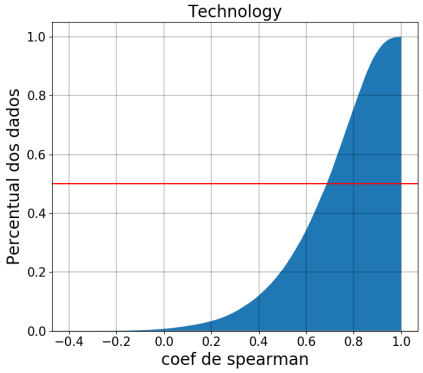


Figura 26 – CDF dos coeficientes de *spearman* dos interesses do tópico *Fitness and Wellness*.

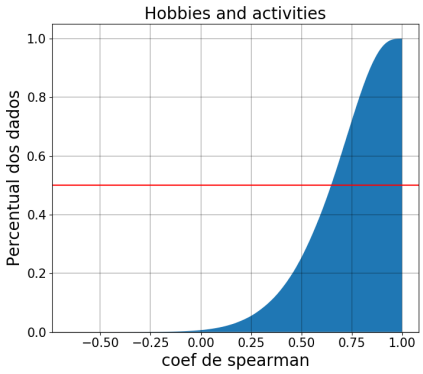


Figura 27 – CDF dos coeficientes de *spearman* dos interesses do tópico *Hobbies and Activities*.

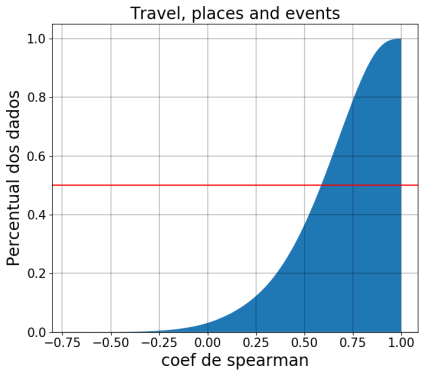


Figura 28 – CDF dos coeficientes de *spearman* dos interesses do tópico *Travel, places and events*.

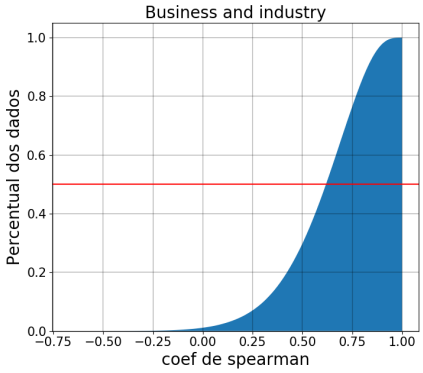


Figura 29 – CDF dos coeficientes de *spearman* dos interesses do tópico *Business and industry*.

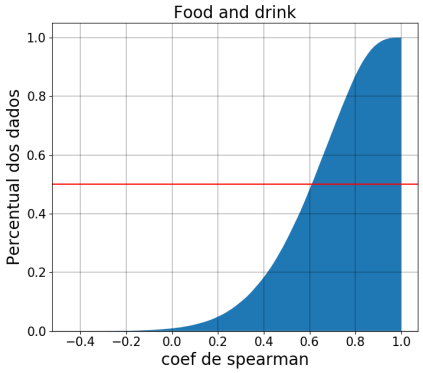


Figura 30 – CDF dos coeficientes de *spearman* dos interesses do tópico *Food and Drink*.

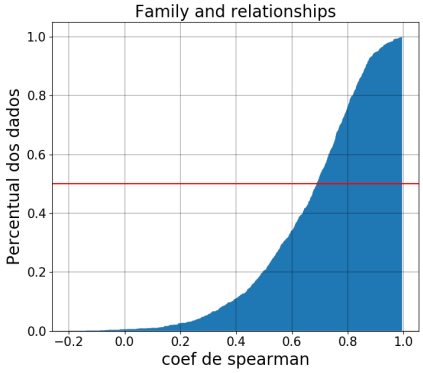


Figura 31 – CDF dos coeficientes de *spearman* dos interesses do tópico *Family and relationships*.

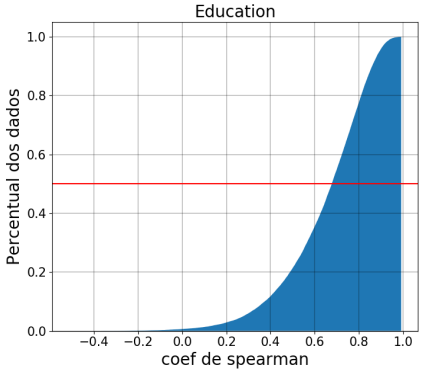


Figura 32 – CDF dos coeficientes de *spearman* dos interesses do tópico *Education*.

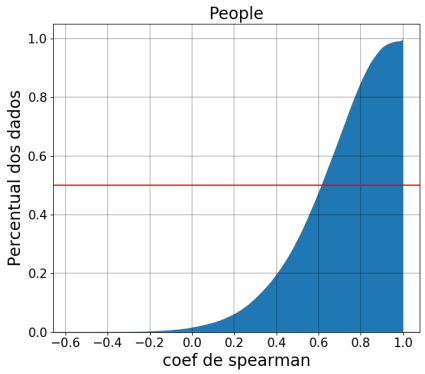
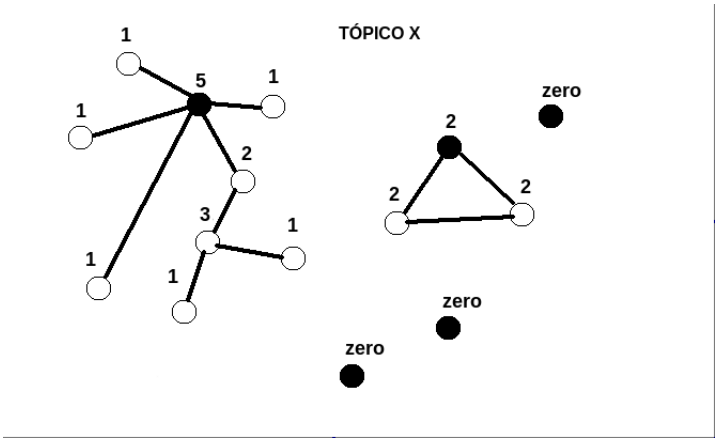


Figura 33 – CDF dos coeficientes de *spearman* dos interesses do tópico *People*.

aquele interesse representa uma boa parte dos dados, portanto ele sozinho vale mais para a discriminação dos dados do que os outros interesses de menor grau que estão conectados a ele.

A Figura 34 mostra alguns interesses em um tópico qualquer. Os vértices que tem arestas os ligando representam um par de interesse que tem coeficiente de *spearman* acima do módulo de um limiar. Cada vértice tem o seu grau de conectividade indicado logo acima do mesmo, sendo possível identificar alguns sub grafos desconexos do restante e alguns vértices isolados. A ideia é ordenar todos os vértices em ordem decrescente de grau, com a hipótese de que o vértice de maior grau de todos, carrega mais representação dos dados, assim todos os vértices ligados a ele são eliminados. Na Figura 35 por exemplo, logo após a eliminação dos vértices conectados naquele de grau 5 (X vermelho), prossegue-se com os cortes, de forma que o próximo vértice de maior grau ainda não visitado é o vértice de grau 3, em que este tem todos os seus vizinhos eliminados também (X amarelo). Os vértices são visitados na ordem decrescente de grau, porém os vértices de mesmo grau são visitados na ordem que os interesses foram coletados, portanto em caso de empate de grau aquele vértice encontrado primeiro é o que permanece, e os outros são eliminados. Esta eliminação de vértices do mesmo grau é o que acontece com os dois vértices do triângulo na Figura 35 (X azul claro). Os cortes foram feitos até chegar nos vértices de grau zero, caso estes existam, de forma que os vértices de grau zero nunca são eliminados.

Figura 34 – Interesses e correlações de *spearman* pertencentes a um tópico, representados em grafo, antes do corte.



Após a aplicação desse filtro, os países foram agrupados novamente, os resultados estão na seção 4.2. Finalizando esta seção, a Figura 36 mostra um fluxograma representando as etapas das metodologias com as duas bases de dados que foram apresentadas.

Figura 35 – Interesses e correlações de *spearman* pertencentes a um tópico, representados em grafo, após o corte.

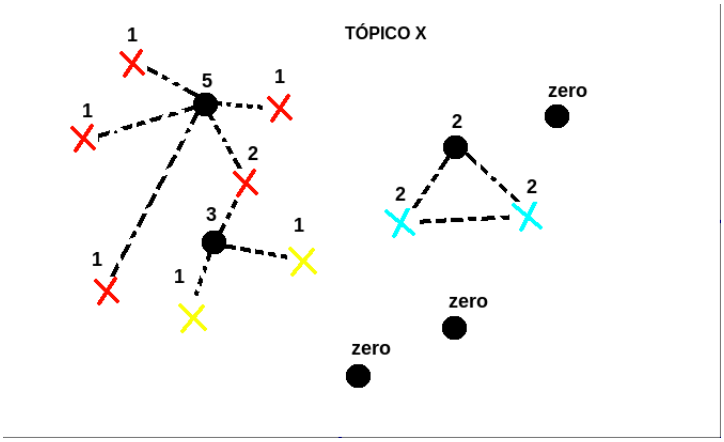
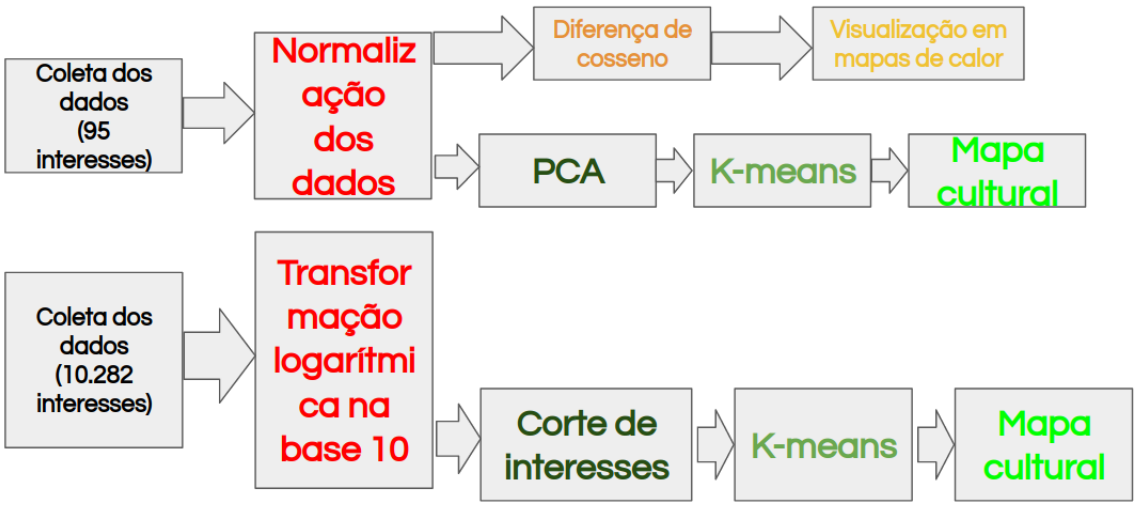


Figura 36 – Fluxogramas simplificados das metodologias



4 Resultados

Neste capítulo são apresentados os resultados das metodologias e análises sobre os *clusters* culturais formados neste trabalho, assim como algumas análises de similaridade por diferença de cosseno. Na seção 4.1 são apresentados os resultados para o primeiro conjunto de dados coletados, apenas sobre comida e bebida, já na seção 4.2 são apresentados os resultados para o segundo conjunto de dados, os 10.282 interesses sobre diversos assuntos, como política, religião, entretenimento, educação, entre outros.

4.1 Inferência de similaridade por comida e bebida

Para o primeiro conjunto de dados, foram geradas algumas comparações entre os 16 países, foram utilizados mapas de calor baseados em diferenças de cosseno, comparou-se os resultados com os mapas de calor do trabalho Silva et al. (2017) presentes na subseção 4.1.1. Além disso foram formados *clusters* culturais, apresentados na subseção 4.1.2, juntamente com os *clusters* culturais em Silva et al. (2017) e WVS (2019).

4.1.1 Mapas de calor e comparações com Silva et al. (2017)

Na Figura 37 é possível ver o mapa de calor para uma parte do vetor de 101 subclasses de estabelecimentos do trabalho (SILVA et al., 2017), intitulado *Slow Food*. Na Figura 38 é possível ver o mapa de calor deste trabalho, para um pedaço do vetor de 95 interesses, também do *Slow Food*, incluindo o interesse denominado *"Food"*. Sobre este interesse em específico é apresentada na Tabela 2 que ele tem o maior valor normalizado de todos os interesses em todos os países, porém o nome dele indica algo muito genérico, por este motivo na na Figura 39 é apresentado o mapa de calor que não utiliza esse interesse. É possível identificar que o mapa de um modo geral fica mais claro, isto significa que agora os países se diferenciam melhor e é possível observar com mais contraste algumas similaridades, como por exemplo a Austrália com a Grã Bretanha e a Singapura com a Malásia.

Foi possível observar alguns pontos em comum, de similaridade entre pares de países, entre os dois trabalhos. Os pares de correlação Inglaterra e Austrália, e Singapura e Malásia são pontos similares. Além disso, no presente trabalho é possível observar outras correlações que condizem com *clusters* do WVS, como as correlações entre Austrália e Estados Unidos, e Inglaterra e Estados Unidos. É importante relembrar que o aspecto qualitativo dos dados das duas redes sociais são diferentes, apresentados nas subseções 3.2.1 e 3.3.

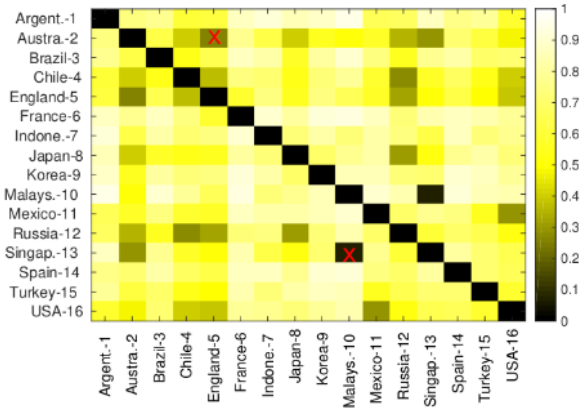


Figura 37 – Mapa de calor *Slow Food* Silva et. al.

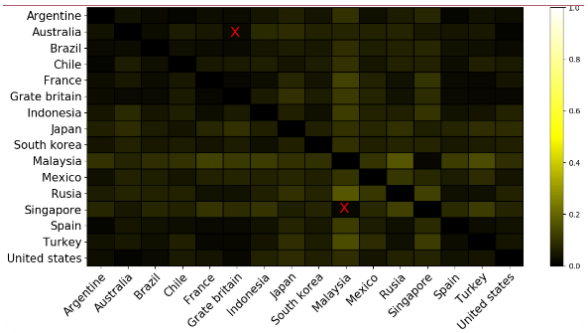
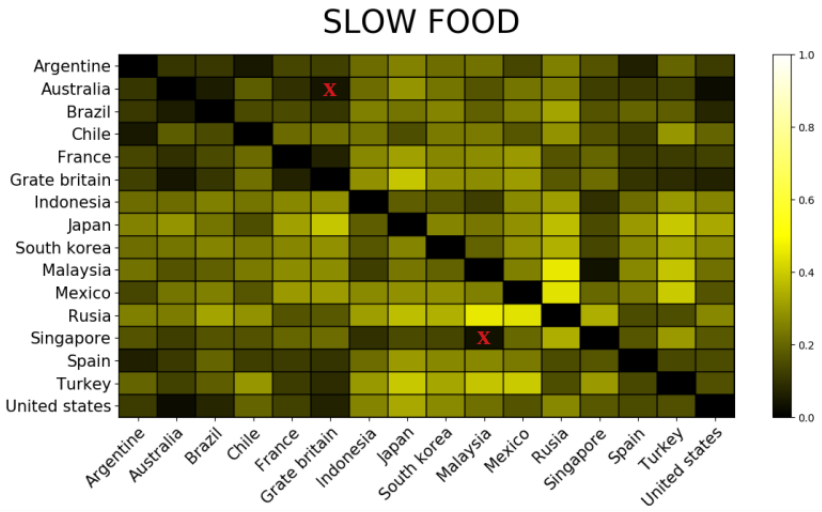


Figura 38 – Mapa de calor *Slow Food* deste trabalho, com o interesse *"Food"*.

Figura 39 – Mapa de calor *Slow Food* deste trabalho, sem o interesse *"Food"*.

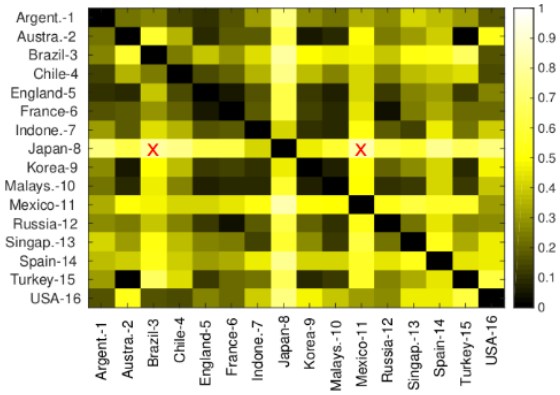


Nas Figuras 40 e 41, são apresentados os mapas de calor dos dois trabalhos, para o subvetor *Fast Food*. Aqui também é possível verificar alguns pontos em comum, entre as duas inferências, com destaque das relações do Japão para com os demais países, com ainda mais destaque com o Brasil e México. Apesar disso é possível observar novamente que o mapa de calor neste trabalho é mais escuro, porém desta vez este mapa de calor já está sem o interesse *"Food"*. O que indica que os dados de *fast-food* do *Facebook* para os 16 países, têm correlações mais altas entre si naturalmente se comparado aos dados da rede social *Foursquare*, ou que há outros interesses redundantes que influenciam nesta diferenciação.

De maneira geral os mapas de calor de ambos os trabalhos apresentam relações entre pares de países que condizem com os grupos culturais formados no WVS como vistos

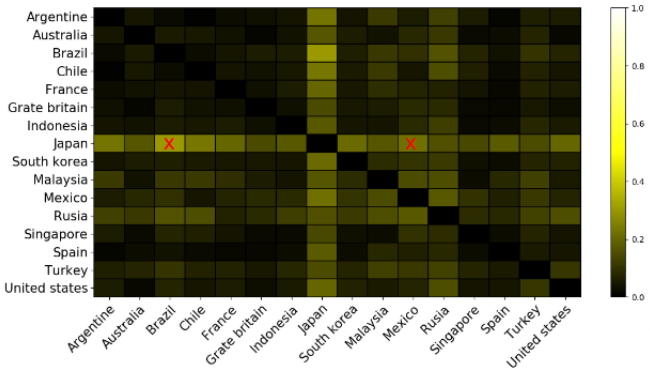
na próxima seção.

Figura 40 – Mapa de calor *Fast Food* Silva et. al.



Fonte: Silva et al. (2017, p. 19).

Figura 41 – Mapa de calor *Fast Food* deste trabalho, sem o interesse *"Food"*.



4.1.2 Clusterização dos países com o uso de PCA e *k-means*

Após a análise com os mapas de calor feitos com valores de diferença de cosseno, foi usado o PCA para transformação dos dados. Em seguida foram montados *clusters* através do algoritmo *k-means* e comparados com os *clusters* em Silva et al. (2017) (Figura 43) e WVS (2019) (Figura 42). É possível observar na Figura 43 que o trabalho base Silva et al. (2017) tem 11 dos 16 países em grupos em que se encontra pelo menos mais um outro país de um *cluster* correspondente no mapa do WVS na Figura 42. É possível observar no trabalho base por exemplo que o trio Argentina, Brasil e Chile, também está no mesmo grupo no WVS.

No presente trabalho, entre os 95 interesses utilizados nesta etapa, foi possível verificar que o assunto com maior número de pessoas interessadas em todos os 16 países era o interesse "*Food*". Por este nome ser muito genérico foram testados *clusters* sem (Figura 44) e com (Figura 45) este atributo.

Nesta seção todos os gráficos dos grupos culturais apresentam no seu eixo horizontal os valores transformados pelo PCA do primeiro componente principal dos respectivos atributos, já o eixo vertical apresenta valores associados ao segundo componente principal.

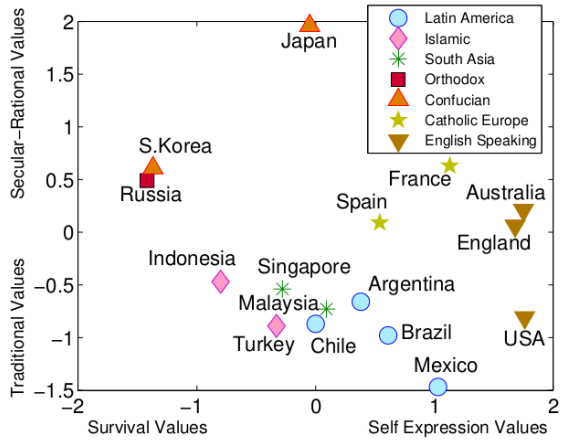
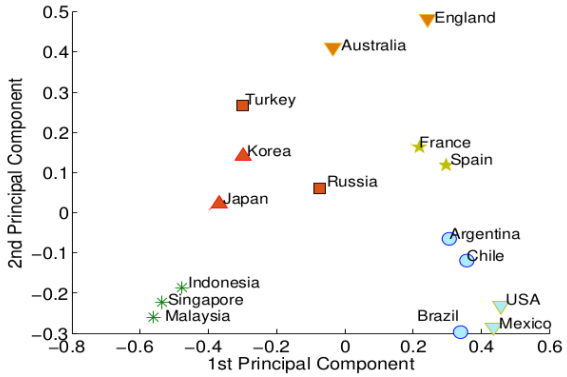


Figura 42 – Mapa cultural mundial por WVS.

Fonte: Silva et al. (2017, p. 33).



(a) $k = 7$

Figura 43 – Mapa cultural mundial por SILVA et al.

Fonte: Silva et al. (2017, p. 41).

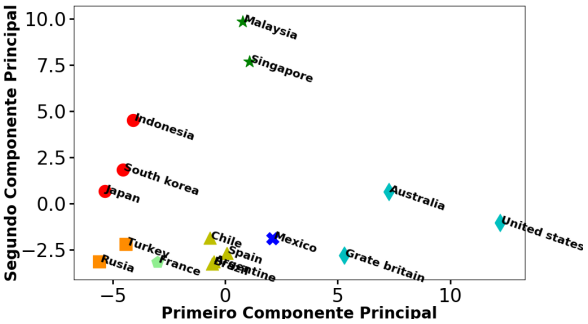


Figura 44 – Mapa cultural mundial do presente trabalho, sem o atributo *"Food"*.

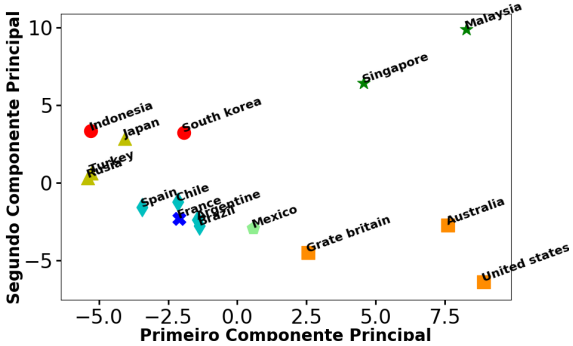


Figura 45 – Mapa cultural mundial do presente trabalho, com o atributo *"Food"*.

O *cluster* com o atributo *"Food"* (Figura 45) tem 8 países em grupos, que em cada um deles há pelo menos mais um outro país, que está no mesmo *cluster* correspondente no trabalho do WVS (Figura 43). Enquanto o grupo sem o mesmo atributo (Figura 44), tem 10 países em *clusters*, em que há pelo menos mais um país que se encontra no mesmo grupo no WVS.

O resultado sem o atributo *"Food"* pode ser visualizado em forma de tabela e comparado com o resultado em WVS na Tabela 4.

4.2 Inferências de similaridades com o uso de interesses gerais

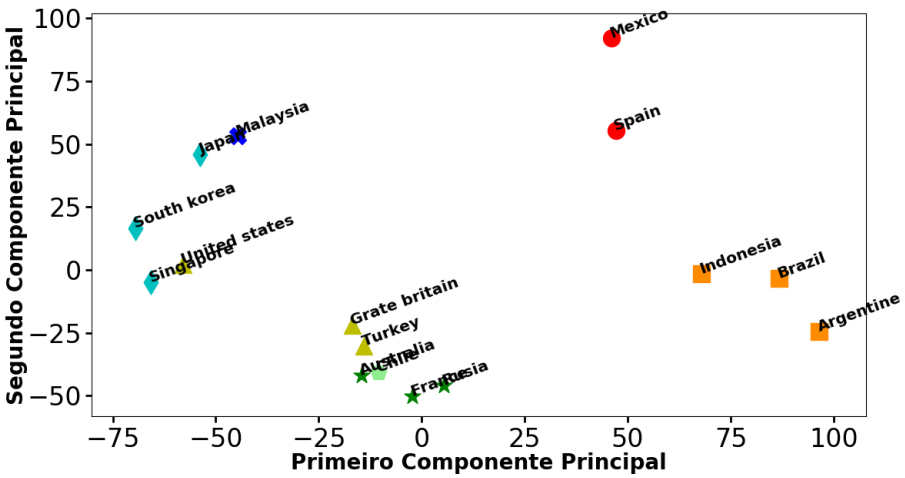
A Figura 46 apresenta o *cluster* formado com exatamente a mesma metodologia em Silva et al. (2017), porém com os 10.282 interesses da segunda coleta, referentes a diversos assuntos. Foi possível notar, que com o uso de todos estes dados, apenas 6 dos 16

Tabela 4 – Clusters do WVS e do presente trabalho e seus respectivos países integrantes.

World Value Survey	**Presente trabalho**	**Correspondência**
Japão, Coreia do Sul	Grupo 1: (Japão, Turquia e Rússia), Grupo 2: (Coreia do Sul e Indonésia)	Nenhuma
Malásia, Singapura	Grupo 3: (**Malásia, Singapura**)	**Total**
Brasil, Argentina, Chile e México	Grupo 4: (Espanha, **Argentina, Brasil e Chile**), Grupo 5: (México)	Parcial
Austrália, Estados Unidos e Inglaterra	Grupo 6: (**Austrália, Estados Unidos e Grã Bretanha**)	**Total**
Turquia e Indonésia	Grupo 1 e Grupo 2	Nenhuma
Rússia	Grupo 1	Nenhuma
França e Espanha	Grupo 3 e Grupo 7: (França)	Nenhuma

países se encontravam em *clusters* relativamente correlacionados com os no WVS.

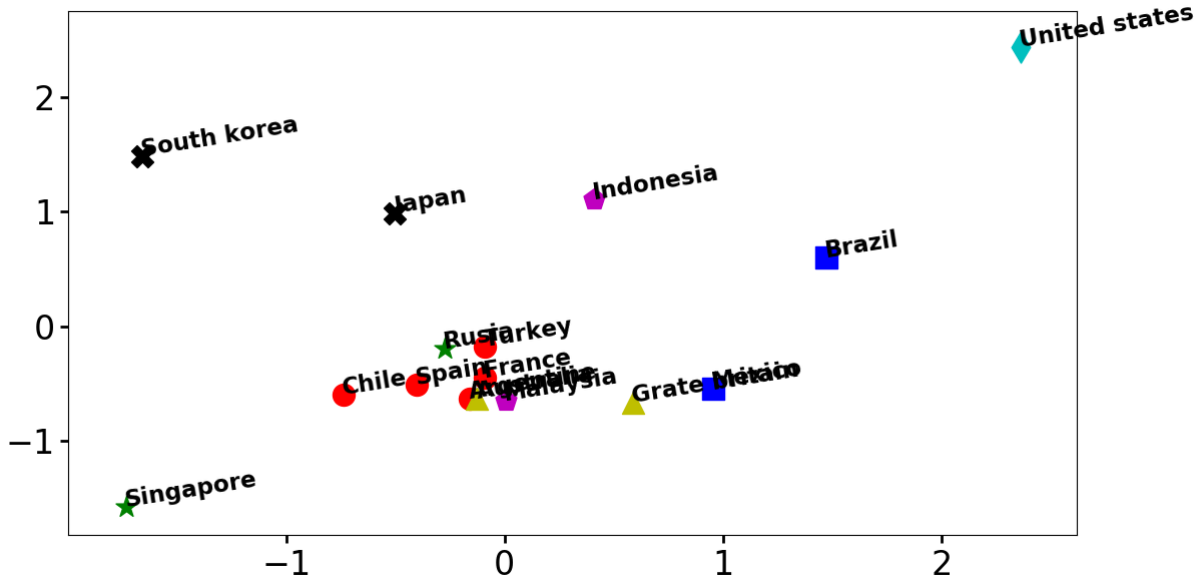
Figura 46 – *Cluster* que usa a mesma metodologia que em Silva et al. (2017), com 10.282 interesses.



Sobre os resultados dos *clusters* feitos com a eliminação de interesses altamente correlacionados. A Figura 47 apresenta um *cluster* em que o critério de eliminação foi a por interesses de maior grau de conectividade, com módulos dos coeficientes de *spearman* acima de 0.4. Sobre a escolha do valor limiar do módulo de 0.4 do coeficiente, foram testados valores entre 0.2 e 0.9, teve-se como melhor resultado um critério de 0.4 que apresentou um grafo com uma densidade intermediária. Uma vez que um grafo muito denso resultou em pouco interesses após a filtragem, que foram em torno de 180 interesses, já os grafos pouco conexos, com critérios próximos a 0.9 ainda deixava muita redundância nos dados, restava em torno de 5000 interesses. Este *cluster* da Figura 47 usou apenas o *k-means*, não usou o PCA uma vez que os dados não têm distribuição normal, uma vez que o PCA trabalha com a premissa de que os dados são normalmente distribuídos. Por este motivo os seis eixos verticais e horizontais são relativos a dois interesses quaisquer, que

foram escolhidos arbitrariamente, o foco da análise do *cluster* no presente trabalho, está em se o grupo contém os mesmos países que os grupos correspondentes do trabalho WVS.

Figura 47 – *Cluster* com filtro por interesses de maiores graus de conectividade, com os coeficientes de *spearman*.



Ainda na Figura 47 foi possível observar que 10 dos 16 países se encontram em *clusters* em que há pelo menos mais um outro país, que também está no mesmo grupo, no trabalho WVS (2019). Uma melhor visualização da região central da Figura 47 é apresentada na Figura 48.

Esses 10 países e grupos são apresentados na Tabela 5 em negrito. Nesta tabela cada linha da primeira coluna à esquerda representa um grupo dos 7 *clusters* no mapa da Figura 43. A coluna do meio tem em cada linha os grupos, do presente trabalho, que têm os países pertencentes ao grupo do WVS, da linha correspondente. Por fim a coluna da direita diz por método visual, o nível de correspondência entre cada grupo, nos dois trabalhos.

Figura 48 – *Cluster* com filtro por interesses de maiores graus de conectividade, com os coeficientes de *spearman* (Região central).

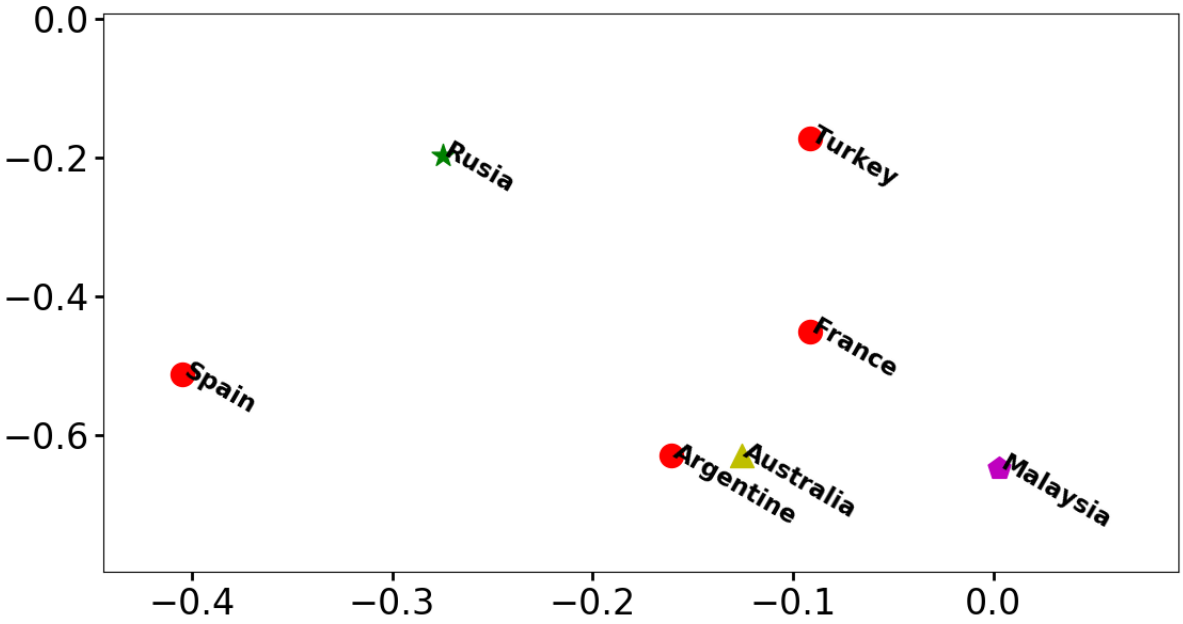


Tabela 5 – Clusters do WVS e do presente trabalho e seus respectivos países integrantes com o uso dos dados gerais.

World Value Survey	**Presente trabalho**	**Correspondência**
Japão, Coreia do Sul	Grupo 1: (**Japão, Coreia do Sul**)	**Total**
Malásia, Singapura	Grupo 2: (Malásia, Indonésia)	Nenhuma
Brasil, Argentina, Chile e México	Grupo 3: (**Brasil e México**), Grupo 4: (**Chile , Argentina**, Turquia, França, Espanha)	Parcial
Austrália, Estados Unidos e Inglaterra	Grupo 5: (**Austrália, Grã Bretanha**), Grupo 6: (Estados Unidos)	Parcial
Turquia e Indonésia	Grupo 7: (Singapura e Rússia), Grupo 4, Grupo 2	Nenhuma
Rússia	Grupo 7	Nenhuma
França e Espanha	Grupo4(**França, Espanha**, mais 3 países)	Parcial

5 Conclusão

Este trabalho apresenta uma nova metodologia para inferência de similaridades culturais entre diferentes regiões com base nos interesses regionais extraídos de dados coletados de plataformas de propagandas de redes sociais. Utilizou-se para isso dados da plataforma de *marketing* do *Facebook*, motivado pelo crescente uso de dados de redes sociais para estudos demográficos. Estes grupos culturais consistiram em 16 países distribuídos entre 7 *clusters* em que o fato de países estarem no mesmo grupo equivale à inferência de que eles são culturalmente próximos. Assim como também foram formados mapas de calor baseados em diferença de cosseno com finalidade de verificar pela cor de cada célula do mapa, o nível de afinidade cultural entre os locais. Como resultados, foram obtidas algumas inferências similares as inferências em WVS e inferências do trabalho base, assim o presente trabalho apresentou a relação que os dados do *Facebook* tem com as características demográficas de cada país. Além disso, as metodologias expostas aqui podem ser usadas para outros estudos de diferenciação de locais ou grupos específicos de usuários na rede social por meio de grupos culturais ou mapa de calor por afinidade de interesses. Metodologias estas, que também podem ser usadas para descrição de grupos de consumidores dentro de mercados específicos, para indicar quais locais por exemplo são mais associáveis a perfis de interesses dos clientes de uma marca, característica essa também presente em Silva et al. (2017).

Foi possível observar o quão úteis os dados podem ser para algumas análises demográficas; porém, com a incerteza sobre como são realizadas as inferências dos dados extraídos da rede social, ainda há algumas dificuldades, como por exemplo saber se o nicho social estudado tem dados muito precisos na rede social.

Apesar dos 7 *clusters* encontrados no presente trabalho não terem uma correlação maior com os *clusters* em WVS, é importante ressaltar que os aspectos qualitativos dos dados nos dois trabalhos são diferentes. Apesar da vantagem que as redes sociais trazem em relação à acessibilidade dos dados, um dos principais desafios dessa nova área de pesquisa, é a questão da validação dessas informações. Ainda também, apesar do uso dos dados de uma rede social para a inferência demográfica, segundo Zagheni et al. (2018) “O Facebook, por exemplo, pode ser visto como um enorme censo digital que é constantemente atualizado. Porém, seus usuários não representam a população subjacente... enquanto estes são tipicamente grandes volumes de dados com enormes amostras que frequentemente provêm informações que são qualitativamente diferentes daquelas obtidas por questionários... ”. Isso implica na necessidade de uma cautela ao comparar dois estudos com o mesmo objetivo porém que usam dados de características um pouco diferentes, como os dados online e aqueles provindos de questionários por exemplo.

O âmbito social do estudo em *World Value Survey* é polarizado para a política, religião e tradição versus novo, utiliza a metodologia clássica de questionários presenciais. Já os dados dos 10.282 interesses extraídos na segunda parte do trabalho não têm este tipo de polarização. Neste sentido trabalhos futuros podem apresentar metodologias para mapear questionários em consultas na plataforma de marketing do *Facebook*. Este mapeamento buscaria uma precisão maior nas escolhas dos interesses a fim de que estes representassem da melhor forma possível as perguntas de um questionário. Outro ponto de melhora é que como o presente trabalho não diferenciou os dados de cada país por faixa etária, renda, raça entre outros, essas diferenciações podem ser adicionadas. Assim, com um nível maior de granularidade dos dados é possível obter resultados mais completos sobre as inferências. Por fim, novas metodologias para corte de dados redundantes podem ser aplicadas e comparadas entre si, a fim de que este problema de agrupamento cultural/comportamental que utiliza dados do *Facebook* seja explorado mais profundamente.

Referências

ARAÚJO, M. et al. Using facebook ads audiences for global lifestyle disease surveillance: Promises and limitations. *In Proceedings of the ACM Conference on Web Science, WebSci'17, pages 253–257*, 2017. Citado 2 vezes nas páginas 16 e 24.

COCHRANE, R.; BAL, S. The drinking habits of sikh, hindu, muslim and white men in the west midlands: a community survey. *British Journal of Addiction*, 1990. Citado na página 21.

CRANSHAW, J. et al. The livehoods project: Utilizing social media to understand the dynamics of a city. *ICWSM'12*, v. 1, n. 1, p. 8, 2012. Citado 3 vezes nas páginas 16, 19 e 20.

DANGETI, P. *Statistics For Machine Learning*. [S.l.]: Packt Publishing, 2017. Citado 2 vezes nas páginas 29 e 30.

FACEBOOKINTERESTSEXPLORER. 2019. Disponível em: <http://blackbird.dcc.ufmg.br/interest_study/app.php?query=6003259680957&name=>. Citado na página 35.

INGLEHART, R.; WELZEL, C. Changing mass priorities: The link between modernization and democracy. *Perspectives on Politics*, v. 1, n. 1, p. 17, 2010. Citado na página 16.

JAIN, R. *The art of computer systems performance analysis*. [S.l.]: Whiley Professional Computing, 1991. Citado 2 vezes nas páginas 29 e 39.

RIBEIRO, F. N. *Inference of demographic data from digital advertising platforms based on social media*. Tese (Doutorado em Ciência da Computação) — UFMG, 2019. Citado 3 vezes nas páginas 23, 24 e 35.

SILVA, T. H. et al. A large-scale study of cultural differences using urban data about eating and drinking preferences. *ScienceDirect*, v. 1, n. 1, p. 45, 2017. Citado 25 vezes nas páginas 10, 11, 14, 16, 17, 18, 20, 21, 22, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 45, 47, 48, 49 e 52.

SOUSA, A. M. de; ALMEIDA, J. M.; FIGUEIREDO, F. Analyzing and modeling user curiosity in online content consumption: A lastfm case study. *ASONAM 2019*, 2019. Citado na página 40.

SPEICHER, T. et al. On the Potential for Discrimination in Online Targeted Advertising. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*'18)*. [S.l.: s.n.], 2018. Citado na página 38.

WVS. 2019. Disponível em: <http://www.worldvaluessurvey.org/wvs.jsp>. Citado 11 vezes nas páginas 16, 17, 19, 25, 26, 27, 30, 33, 45, 47 e 50.

ZAGHENI, E. et al. Combining social media data and traditional surveys to nowcast migration stocks. *Unece*, v. 1, n. 1, p. 45, 2018. Citado 4 vezes nas páginas 16, 24, 25 e 52.