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

AvaliaGeo: Um sistema para geração e validação de bases de dados de notícias geocodificadas

Arthur Felipe de Freitas Domingues

TRABALHO DE CONCLUSÃO DE CURSO

ORIENTAÇÃO:
Bruno Rabello Monteiro

Dezembro, 2019
João Monlevade–MG

Arthur Felipe de Freitas Domingues

AvaliaGeo: Um sistema para geração e validação de bases de dados de notícias geocodificadas

Orientador: Bruno Rabello Monteiro

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

D671a Domingues, Arthur Felipe.
AvaliaGeo [manuscrito]: um sistema para geração e validação de bases de dados de notícias geocodificadas / Arthur Felipe Domingues. - 2019.

45f.:

Orientador: Prof. MSc. Bruno Rabello Monteiro.

Monografia (Graduação). Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Aplicadas. Departamento de Computação e Sistemas de Informação.

1. Banco de dados. 2. Aplicações web. 3. Avaliação. 4. Sistemas de informação geográfica. I. Monteiro, Bruno Rabello. II. Universidade Federal de Ouro Preto. III. Titulo.

CDU: 004.775

Catalogação: ficha.sisbin@ufop.edu.br

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

FOLHA DE APROVAÇÃO

Arthur Felipe de Freitas Domingues

AvaliaGeo: Um sistema para geração e validação de bases de dados de notícias geocodificadas

Membros da banca

Bruno Rabello Monteiro - Mestre - Decsi - Ufop
Gilda Aparecida de Assis - Doutora - Decsi - Ufop
Fernando Bernardes de Oliveira- Doutor - Decsi - Ufop

Versão final
Aprovado em 13 de dezembro de 2019

De acordo

Bruno Rabello Monteiro

Documento assinado eletronicamente por **Bruno Rabello Monteiro**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 16/01/2020, às 13:41, 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 **0032307** e o código CRC **0D5D07F0**.

Referência: Caso responda este documento, indicar expressamente o Processo nº 23109.000317/2020-95 SEI nº 0032307

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

Este trabalho é dedicado ao meu pai, Ermelindo, e à minha irmã, Marine.

Agradecimentos

Agradeço à minha família por todo o suporte durante toda a graduação.

Agradeço ao meu orientador, o professor Bruno Monteiro, por ter aceitado a orientação no desenvolvimento deste trabalho. O auxílio, lições e conselhos foram, sem dúvida, essenciais.

Agradeço aos amigos e colegas com os quais convivi durante toda a graduação. Em especial, aos membros do Triviais Work's: Arthur Bernardo, Edgar Alves e Rafael Oliveira, pela parceria e apoio durante toda a graduação.

Por fim, agradeço a comunidade do *stackoverflow*, sem vocês, este trabalho, com certeza, seria muito mais difícil.

"Algumas pessoas acham que foco significa dizer "Sim" para a coisa em que você irá se focar. Mas não é nada disso. Significa dizer "Não" às centenas de outras boas ideias que existem. Você precisa selecionar cuidadosamente."

— Steve Jobs (1955 – 2011)

Resumo

Há uma demanda crescente por aplicações que realizem a extração de informações geográficas em mídias textuais. Um dos fatores que impulsiona este tipo de aplicação é a possibilidade de desenvolver sistemas personalizados para grupos distintos de pessoas. Existem várias abordagens propondo soluções para este problema e problemas correlatos. Entretanto, essas utilizam bases de dados específicas para a construção dos algoritmos. Dessa forma, apesar de compartilharem problemas semelhantes, a comparação e cooperação entre as soluções propostas ficam comprometidas. Este trabalho teve como objetivo a construção de bases de dados confiáveis para facilitar o *benchmark* entre os algoritmos propostos na literatura. Foram utilizados textos de um portal de notícias para a criação da base de dados. A validação das informações geográficas presentes em cada notícia foi realizada por meio da participação dos usuários no sistema *web* desenvolvido. A participação foi realizada por meio de questionários sobre os topônimos encontrados no texto da notícia. Foi possível avaliar 20 notícias, sendo 70% delas consideradas confiáveis com base no coeficiente alfa de Cronbach.

Palavras-chaves: Bases de dados rotuladas. Questionários. Alfa de Cronbach. Sistema *web*.

Abstract

There is an increasing demand for applications that perform geographic information extraction on textual media. One of the factors driving this type of application is the ability to develop custom systems for different groups of people. There are various approaches proposing solutions to this problem and related problems. However, they use specific databases to construct algorithms. Thus, although they share similar problems, the comparison and cooperation between the proposed solutions are compromised. This work aimed to build reliable databases to facilitate the benchmark between the algorithms proposed in the literature. We used texts from a news portal to create the database. The validation of the geographic information present in each news was performed through the participation of users in the web system developed. Participation was made through questionnaires about the toponyms found in the news text. It was possible to evaluate 20 stories, being 70% considered reliable based on Cronbach's alpha coefficient.

Key-words: Labeled datasets. surveys. cronbach's alpha. web system.

Lista de ilustrações

Listagem

Lista de tabelas

Lista de abreviaturas e siglas

GSR	*Geographic Scope Resolution*
REST	*Representational state transfer*
UBS	Unidades Básicas de Saúde
UENF	Universidade Estadual do Norte Fluminense
SINAES	Sistema Nacional de Avaliação da Educação Superior
GI	Grau de Importância
GD	Grau de Desempenho

Lista de símbolos

α	Coeficiente Alfa de Cronbach
X	Matriz de respostas do questionário
σ_i^2	Variância das colunas de X
σ_τ^2	Variância da soma de cada coluna de X
k	Número de itens no questionário
n	Quantidade de respostas no questionário

Sumário

1 Introdução

A utilização de informação geográfica em aplicações vem aumentando. Um dos fatores de sustentação desse aumento é a possibilidade de desenvolvimento de aplicações personalizadas para determinados grupos de pessoas (LARSEN, 2010), ou mesmo, pela popularização de serviços como Google Maps[1] e outras ferramentas de navegação.

O problema de Resolução de Escopo Geográfico (*Geographic Scope Resolution*, ou simplesmente GSR) lida com a determinação do foco geográfico de textos e documentos. Para isso, é necessário a identificação e desambiguação dos nomes que representam lugares geográficos, ou seja, dos topônimos. A identificação dessas informações deve levar em consideração o contexto em que os topônimos estão inseridos.

Com as devidas identificações, é possível extrair o escopo geográfico, ou seja, saber qual região geográfica é abordada no texto. Em seguida, esse escopo pode ser utilizado para construir aplicações e sistemas mais direcionados, como melhoria em sistemas de busca e direcionamento de conteúdo por características geográficas.

O GSR pode agregar valor a uma gama de aplicações, como por exemplo a indexação e o ranqueamento geográfico em motores de busca, possibilitando, além da tradicional busca temática (conjunto de palavras-chave), a busca geográfica (MONTEIRO; DAVIS; FONSECA, 2016). Uma aplicação direta pode ser observada em Rupp et al. (2013), em que foi feita a sumarização de textos históricos utilizando a extração de escopo geográfico.

É possível também a extração de informação: em tempo real, de dados de serviços como o Twitter[2], para descoberta de ocorrências de desastres naturais (MIDDLETON; MIDDLETON; MODAFFERI, 2014); assim como ao longo do tempo, possibilitando aferição de moradia de determinado indivíduo, por meio da recorrência de postagens em uma região específica (ALEX et al., 2016).

Com um grande número de aplicações que podem se beneficiar, um conjunto de soluções para o GSR vêm sendo desenvolvidas. Entretanto, a comparação entre estas abordagens é comprometida pois, para cada solução utilizam-se, geralmente, estruturas próprias para a validação dos algoritmos. Este trabalho propõe o desenvolvimento de um sistema para avaliar bases de dados geográficas rotuladas. Tais bases de dados podem ser utilizadas, tanto pelas diferentes soluções apresentadas, quanto as que virão a ser desenvolvidas nos respectivos processos de avaliação.

[1] https://www.google.com.br/maps
[2] https://twitter.com

1.1 O Problema de Pesquisa

Monteiro, Davis e Fonseca (2016) e Gritta et al. (2018) apresentam revisões bibliográficas sobre os trabalhos desenvolvidos para as soluções do GSR. Em ambas as publicações, os autores relatam a falta de capacidade na análise semântica dos dados, isto é, os algoritmos conseguem identificar os topônimos (análise sintática), entretanto a determinação do escopo em que a informação está inserida é, em sua grande maioria, falha.

Gritta et al. (2018) apresentam um conjunto de exemplos em que, apesar da identificação de um topônimo, a classificação não necessariamente reflete o escopo geográfico. Dentre os exemplos citados: *"We had an amazing time at Six Flags Over Texas"*. Nesse exemplo, "Texas" é identificado como o estado americano pela maioria dos algoritmos, entretanto o escopo é apenas um parque de diversões. Já na sentença "*London Road*", "*London*" é reconhecida como uma cidade, e não apenas como a estrada que leva seu nome. Além disso, os mesmos autores citam a falta de bases de dados gratuitas para realização de testes e experimentos. Doran et al. (2005), por sua vez, indicam bases de dados no valor de até $1000,00 dólares.

De acordo com (MONTEIRO; DAVIS; FONSECA, 2016), um dos problemas existentes para o avanço nas soluções é que as abordagens propostas para o GSR utilizam algoritmos próprios, com bases de dados próprias para avaliar suas soluções, de modo que a comparação entre as várias soluções presentes na literatura fica comprometida.

1.2 Objetivos

O objetivo deste trabalho é construir uma aplicação *web* que permita a criação e validação de bases de dados geocodificadas. Especificamente, este trabalho foca em notícias como objeto de pesquisa. As bases de dados poderão ser utilizadas, posteriormente, para comparar as diferentes abordagens e soluções encontradas na literatura para o GSR.

A geração das bases de dados será realizada de modo semiautomático. A parte manual será realizada pelos usuários do sistema, que analisarão o contexto da notícia e, em conjunto com seus conhecimentos prévios, determinarão a qual localidade determinado topônimo se refere. Para cada topônimo avaliado haverá uma escala associada, de modo que o usuário indique o grau de confiabilidade de sua resposta.

A validação dos dados fornecidos pelos usuários do sistema será realizada por meio de análises estatísticas, mais especificamente será utilizado a escala de classificação em conjunto com o Alfa de Cronbach (α) para a análise dos dados obtidos, permitindo filtrar as informações coletadas, e obter um maior grau de confiabilidade.

Ao final do projeto, as bases de dados, ou seja, as notícias, estarão armazenadas em

uma estrutura de arquivos JSON, em que cada topônimo terá um conjunto de atributos associados.

1.3 Organização do Trabalho

O trabalho está estruturado em cinco capítulos. O Capítulo 1 é responsável pela introdução do trabalho. No Capítulo 2 é discorrido a respeito da fundamentação teórica do trabalho. Em seguida, o Capítulo 3 detalha todo o desenvolvimento do projeto. No Capítulo 4 são apresentados os resultados obtidos no trabalho. No Capítulo 5 são apresentadas as considerações finais e as possibilidades de trabalhos futuros.

2 Revisão Bibliográfica

Este trabalho aborda diferentes áreas do conhecimento, portanto, nas próximas seções serão apresentados conceitos, tecnologias e trabalhos relacionados que auxiliarão o leitor a um melhor entendimento sobre o desenvolvimento do projeto.

2.1 Resolução de Escopo Geográfico

A Resolução de Escopo geográfico (GSR) é um problema que consiste em identificar os topônimos (referências a locais) ou referências associadas a um determinado local presentes em um documento de texto. Entretanto, a identificação dos topônimos devem levar em consideração o contexto em que os dados estão sendo apresentados.

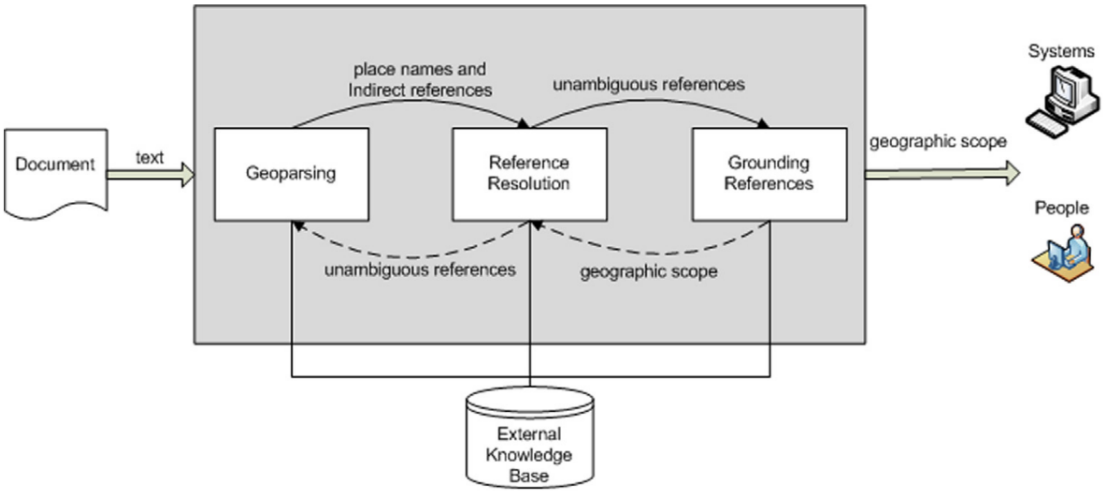
Com as devidas identificações, é possível extrair o escopo geográfico, ou seja, obter os dados que caracterizam a informação geográfica presente no texto. Em seguida, é possível utilizá-lo para a construção de diversas aplicações. A Figura 1 mostra, de forma simples, as etapas para a extração do escopo geográfico. O documento de texto serve como entrada para os algoritmos de resolução do GSR. Estes, por sua vez, determinam o escopo geográfico do documento com base nos referências à lugares presentes no documento.

Figura 1 – Etapas para extração do escopo geográfico

Fonte: Elaborado pelo autor

De acordo com (MONTEIRO; DAVIS; FONSECA, 2016), a solução do GSR pode ser dividida em três etapas: *Geoparsing*, *Reference Resolution* e *Grounding References*. A Figura 2 mostra como as etapas apresentadas se relacionam umas com as outras no processo de GSR.

Figura 2 – Etapas do GSR.



Fonte: (MONTEIRO; DAVIS; FONSECA, 2016)

A etapa de *Geoparsing* tem por objetivo a identificação das referências geográficas diretas (topônimos) ou indiretas (entidades associadas a uma localização geográfica) presentes no texto. Para auxiliar o processo, pode ser utilizado um *gazetteer* (dicionário de informação geográfica), ou outra base externa de conhecimento, como um banco de dados.

Com as referências identificadas, muitas vezes, é necessário desambiguar cada uma delas. Isso ocorre porque locais podem compartilhar um mesmo nome, e se referirem a diferentes localidades. Um exemplo deste fenômeno é a cidade de Paris (capital da França) e a cidade de Paris, no estado do Texas (Estados Unidos) (GRITTA et al., 2018). A desambiguação é realizada na etapa de *Reference Resolution*. Pode-se utilizar um grupo de heurísticas, assim como aprendizagem de máquina para desambiguar os topônimos corretamente.

Na última etapa, *Grouding References*, ocorre a determinação do escopo geográfico do texto. Com as referências identificadas e desambiguadas, o escopo de um documento pode ser determinado, seja via heurísticas, seja via algum algoritmo geométrico como, por exemplo, a menor região formada pelos topônimos presentes no texto.

Mais informações sobre cada etapa do processo do *GSR*, como as heurísticas para identificação e desambiguação de topônimos por exemplo, podem ser obtidas em (MONTEIRO; DAVIS; FONSECA, 2016) e (GRITTA et al., 2018).

O objetivo deste trabalho não é desenvolver os algoritmos em si, mas sim fazer uma ligação entre os resultados dos algoritmos e análises estatísticas a partir dos dados obtidos pelos usuários do sistema.

2.2 Análises Estatísticas

Para o desenvolvimento deste trabalho, os dados fornecidos pelos usuários foram classificados como respostas a questionário, em que o número de perguntas é igual ao número de topônimos presentes na notícia. Cada topônimo possui um conjunto de itens, ou possibilidades a serem escolhidas, juntamente com o grau de confiabilidade de sua resposta.

No trabalho, será avaliado a consistência entre as respostas fornecidas pelos usuários para cada um dos topônimos presentes em uma determinada notícia. Para realizar esta validação dos topônimos do questionário, o coeficiente Alfa de Cronbach (*Cronbach Alpha*) foi utilizado (CRONBACH, 1951). Para a confiabilidade da resposta de cada topônimo, foi utilizado uma *Discrete-ordinal Ranking Scale* (escala de classificação discreta-ordinal) e calculado as estatísticas descritivas relacionadas ao topônimo. Tanto o coeficiente Alfa de Cronbach, quanto a escala de classificação discreta-ordinal são apresentados a seguir.

2.2.1 Alfa de Cronbach

De acordo com (CORTINA, 1993) o coeficiente Alfa de Cronbach, descrito em 1951 por Lee J. Cronbach (CRONBACH, 1951) é uma das ferramentas estatísticas mais importantes em pesquisas que envolvem testes e validação. O coeficiente é um índice utilizado para medir a confiabilidade no processo de avaliação de uma escala, ou seja, para avaliar a magnitude em que os itens estão correlacionados. Em outras palavras, o Alfa de Cronbach é a média das correlações entre os itens que fazem parte de um determinado estudo.

Para se estimar o alfa, considera-se X como sendo uma matriz $n \times k$, em que: n representa a quantidade de respostas no questionário, e k é o número de itens no questionário. Este produto corresponde às respostas quantificadas de um questionário. Cada linha de X representa um sujeito e cada coluna representa uma questão. As respostas quantificadas podem estar em qualquer escala (LEONTITSIS; PAGGE, 2007).

O coeficiente alpha pode ser calculado a partir da Equação 2.1.

$$\alpha = \frac{k}{k-1}\left[\frac{\sigma_\tau^2 - \sum_{i=1}^{k}\sigma_i^2}{\sigma_\tau^2}\right] \tag{2.1}$$

reescrevendo, obtêm-se,

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\sigma_\tau^2}\right] \tag{2.2}$$

em que, σ_i^2 é a variância de cada coluna de X, σ_τ^2 é a variância da soma de cada linha de X. Os valores de k e n devem ser maiores do que 1, de forma a não haver um denominador

igual a zero no cálculos do Alfa de Cronbach;

Na Equação 2.2, k é um fator de correção. Se há consistência nas respostas quantificadas, então σ_τ^2 será relativamente grande, fazendo com que o α tenda a 1. Por outro lado, respostas randômicas farão com que σ_τ^2 seja comparável com a soma das variâncias individuais (σ_i^2), fazendo com que o α tenda a zero (LEONTITSIS; PAGGE, 2007).

2.2.1.1 Interpretação do Coeficiente Alfa de Cronbach

O coeficiente Alfa de Cronbach é uma propriedade inerente do padrão de resposta da população estudada, não uma característica da escala por si só; ou seja, o valor do Alfa sofre mudanças segundo a população na qual se aplica a escala (STREINER, 2003a).

O valor mínimo considerado aceitável para o Alfa é 0,70; abaixo desse valor, a consistência interna da escala utilizada é considerada baixa. Em contrapartida, o valor máximo esperado é 0,90. Acima deste valor, pode-se considerar que há redundância ou duplicação, ou seja, vários itens estão medindo exatamente o mesmo elemento de um constructo. Valores de Alfa de Cronbach entre 0,80 e 0,90 são os desejados, segundo Streiner (2003b).

2.2.2 Escala de Classificação Discreta-Ordinal

Dentre as várias escalas presentes na literatura (STEVENS et al., 1946), a que melhor se encaixa na proposta do trabalho para a avaliação do grau de confiabilidade da resposta do usuário é a escala de classificação discreta-ordinal (*Discrete-Ordinal Ranking Scale*).

Uma escala discreta-ordinal é uma variável em que os diferentes estados são ordenados em uma sequência significativa. Sob uma escala ordinal, os sujeitos ou objetos são classificados em termos de grau ao qual eles possuem uma característica de interesse (LAWAL; LAWAL, 2003). Uma escala ordinal indica direção, além de fornecer informações nominais sobre o objeto de avaliação (MARATEB et al., 2014).

Para o projeto, a escala de classificação discreta-ordinal irá consistir em valores de 1 a 5, representados pelas porcentagens de 0%, 25%, 50%, 75%, 100% de confiabilidade na resposta do usuário ao topônimo associado.

2.3 Trabalhos Relacionados

O desenvolvimento deste trabalho tem como ideia central o desenvolvimento de um sistema *web* que possibilite aos usuários auxiliar a identificação dos topônimos presentes em textos de notícias. A notícia é tratada como um questionário. Assim, cada topônimo

identificado na notícia é considerado como uma pergunta no questionário, sendo as possíveis localizações para o topônimo identificado a resposta para a pergunta. Em seguida, é utilizado um coeficiente para medir a consistência dos dados enviados pelos usuários. Os trabalhos a seguir, apesar de não avaliarem topônimos em notícias, seguem a ideia de utilizar o coeficiente Alfa de Cronbach para medir a consistência dos dados obtidos em questionários.

Em (ALMEIDA; SANTOS; COSTA, 2010) foi utilizado o coeficiente Alfa de Cronbach para medir a confiabilidade de um questionário aplicado aos funcionários da rede de saúde pública, mais especificamente em unidades básicas de saúde, na cidade de Guaratinguetá. O objetivo era avaliar os níveis de satisfação destes funcionários nas UBS. Segundo os autores, verificando-se a confiabilidade dos resultados obtidos, é possível dar maior relevância e robustez à pesquisa feita na UBS da cidade de Guaratinguetá.

Em outro exemplo, (FREITAS; RODRIGUES, 2005) utilizou o coeficiente Alfa de Cronbach para analisar a confiabilidade de um questionário enviado a uma amostra do corpo docente da Universidade Estadual do Norte Fluminense (UENF) para sua avaliação institucional. No âmbito da avaliação educação superior, o sistema de informações educacionais gerenciado pelo INEP atualmente possui dois componentes: a Avaliação Institucional e o Sistema Nacional de Avaliação da Educação Superior (SINAES). Nesse questionário, cada docente deveria avaliar ao Grau de Importância (GI) de cada item, assim como avaliar o Grau de Desempenho (GD) da Universidade à luz dos itens.

Em (MATTHIENSEN, 2010) foi desenvolvido um exercício, o qual busca ilustrar o estabelecimento da confiabilidade de um questionário. Esse questionário foi utilizado para mensurar a qualidade de um produto ou serviço (construto) através de avaliações de percepção (indicadores). O autor também faz um levantamento nos motores de busca *Scopus*[1] e *SpringerLink*[2]. O objetivo foi identificar as áreas do conhecimento onde o coeficiente Alfa de Cronbach é utilizado com mais frequência. Os resultados mostram que o coeficiente tem um grande número de aplicações nas mais diversas áreas do conhecimento.

Os sites anteriormente citados apresentam diferenças no sistema de classificação das áreas. Logo, os resumos e os artigos encontrados apresentam-se distribuídos de forma distinta nos motores de busca, resultando em 28 áreas diferentes para o Scopus e 11 áreas diferentes para o SpringerLink.

Em ambos foi possível observar que, historicamente, a maior utilização do Coeficiente alfa de Cronbach está ligada à área de saúde (Medicina, Psicologia, Enfermagem, Ciências Comportamentais), seguido das áreas de Ciências Sociais e de Negócios e Economia. Entretanto, é digna de nota a grande abrangência do uso do Coeficiente Alfa de Cronbach nas demais áreas do conhecimento, o que mostra seu grande potencial de

[1] https://www.scopus.com/
[2] https://link.springer.com/

utilização.

3 Desenvolvimento

Neste capítulo, são detalhadas todas as etapas realizadas durante o desenvolvimento do sistema, o qual recebeu o nome de AvaliaGeo[1]. São apresentadas as atividades referentes a arquitetura da aplicação, tecnologias utilizadas, estruturas de dados em que as bases de dados serão armazenadas, assim como a interface do sistema perante o usuário.

Toda a implementação do sistema, juntamente com os dados obtidos dos usuários podem ser obtidos através da página do projeto na plataforma Github[2].

3.1 Arquitetura e Estrutura da Aplicação

Nesta etapa, foram realizadas pesquisas relativas as ferramentas e tecnologias a serem utilizadas no desenvolvimento do sistema. O principal objetivo deste estudo foi a obtenção de um maior entendimento sobre as vantagens, desvantagens e limitações de cada ferramenta.

3.1.1 Arquitetura da Aplicação

A arquitetura utilizada para o desenvolvimento do sistema é a cliente/servidor. Para a escolha das tecnologias a serem utilizadas para o desenvolvimento, foram realizados um conjunto de testes. Esses testes tiveram como objetivo identificar quais destas tecnologias melhor se encaixam no contexto do projeto. Cada uma das escolhas será apresentada a seguir, juntamente com a justificativa do porque de cada uma delas.

A implementação da parte cliente da aplicação utiliza HTML5, CSS3 e JavaScript. Em adição a estas, são utilizados: o framework jQuery[3], para simplificar a utilização de algumas funções do *JavaScript*, e o Bootstrap[4], para a manipulação das folhas de estilo. Foi utilizado o site *Distinctive Themes* [5] para a obtenção de modelos, assim como, algumas configurações para as folhas de estilo.

A escolha da linguagem do servidor foi Python. A decisão de sua escolha é justificada pela utilização de algoritmos de identificação de topônimos no sistema, que tem sua implementação na linguagem. Dessa forma, facilita o desenvolvimento com o reuso do código, evitando a reimplementação em uma outra linguagem para o servidor.

[1] http://avaliageo.herokuapp.com
[2] https://github.com/artcomp/Avaliageo
[3] https://jquery.com
[4] https://getbootstrap.com
[5] https://wordpress.org/themes/author/distinctive-themes/

Mais especificamente, foi utilizado o Flask[6], um *microframework* construído com base na linguagem de programação Python. Flask tem como característica a simplicidade no processo de desenvolvimento de aplicações de pequeno porte, fornecendo um conjunto essencial de funcionalidades e deixando a cargo do desenvolvedor incorporar novas utilidades. Para a utilização do Flask, é necessário o uso de uma outra linguagem chamada Jinga[7]. De modo geral, Jinga tem como objetivo gerenciar a troca de informações entre a parte cliente e a parte servidora da aplicação.

Uma alternativa ao Flask seria o *framework* Django[8](também desenvolvido em Python), entretanto, ele apresenta uma complexidade maior para o gerenciamento de aplicações, além de que o conjunto de recursos que já vem implementadados*(built-in)* não são necessários ao projeto.

Para o armazenamento das informações fornecidas pelos usuários, foi utilizado o Firebase Storage[9]. A decisão por este serviço se deve a simplicidade dos dados fornecidos pelos usuários. De forma que, a implementação de um banco de dados relacional, mais comumente utilizado por aplicações *Web*, não se justifica no contexto deste trabalho.

Para a hospedagem do sistema, foi utilizado o Heroku[10]. Os motivos da escolha foram: a simplicidade no *deploy* da aplicação, capacidade de processamento compatível com outros serviços que fornecem hospedagem a aplicações Flask sem custo, além da integração contínua com o git[11].

3.1.2 Estrutura da Aplicação

Apesar de inicialmente o projeto visar uma aplicação integrada, o sistema foi dividido em três partes: uma parte pré-processada, a aplicação propriamente dita, e uma parte para o processamento das informações fornecidas pelos usuários. Após alguns testes, foi verificado que a execução dos *scripts* para a identificação dos topônimos atrasa a resposta do sistema como um todo, já que a capacidade de processamento do servidor de hospedagem é limitada devido ao plano de serviço utilizado.

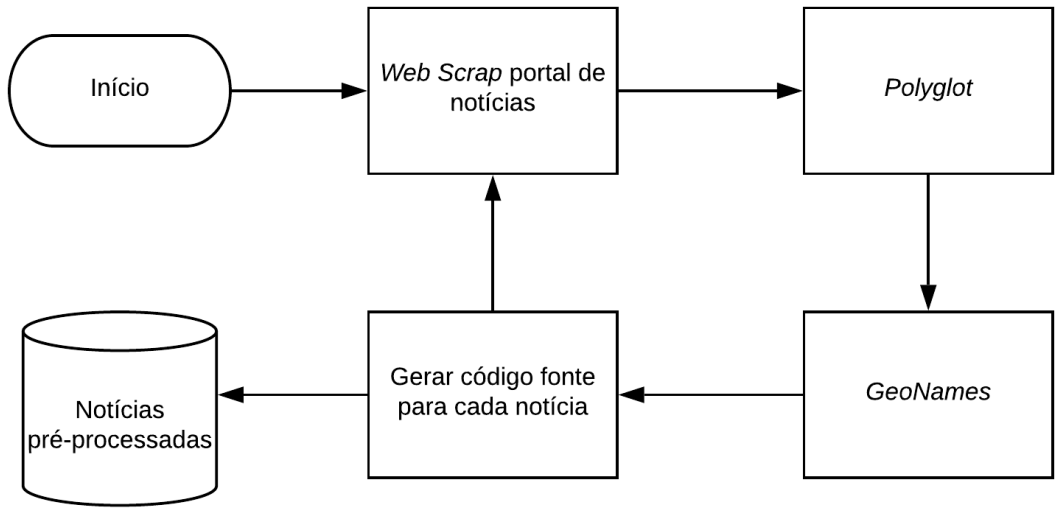
Foi assumido que, caso o tempo de resposta da aplicação seja muito grande, a probabilidade do usuário sair, ou mesmo não voltar a utilizar o sistema seria maior, já que, para cada notícia, seria necessário esperar a identificação, assim como as possíveis classificações, para cada topônimo presente. A seguir, é apresentado o desenvolvimento de cada uma das partes da aplicação.

[6] https://flask.pocoo.org
[7] www.jinja.pocoo.org/
[8] https://djangoproject.com
[9] https://firebase.google.com/
[10] https://www.heroku.com/
[11] https://git-scm.com/

3.1.2.1 Pré Processamento

Inicialmente, um conjunto de notícias são coletadas em um portal de notícias e pré-processadas para serem utilizadas na aplicação. A Figura 3 apresenta o fluxo de execução nesta etapa da aplicação.

Figura 3 – Fluxo de execução do pré-processamento



Fonte: Elaborado pelo autor

De início, é realizado um processo de *Web Scraping* (extração de todo o código fonte que é executado em uma aplicação cliente) no portal de notícias. Logo após, o conteúdo é filtrado e obtêm-se apenas o título e os parágrafos da notícia. Foi utilizado o portal do G1[12] como fonte de notícias para este trabalho.

Um detalhe importante é que, pela própria característica de um portal de notícias, estas podem sofrer alterações com o passar do tempo. Para minimizar este efeito, foi escolhido notícias com um intervalo de acontecimento maior. Com a etapa de extração concluída, o texto obtido, ou seja, os parágrafos da notícia, é passado para um algoritmo de identificação de topônimos.

Neste processo, etapa referente ao *geoparsing* no GSR, foi utilizada a biblioteca Polyglot[13] (Uma biblioteca para processamento de linguagem natural que suporta aplicativos multilínguas em massa). A escolha desta biblioteca justifica-se por suportar a identificação de topônimos em português, característica que não está presente em grande parte das soluções propostas na literatura.

[12] https://g1.globo.com
[13] https://polyglot.readthedocs.io

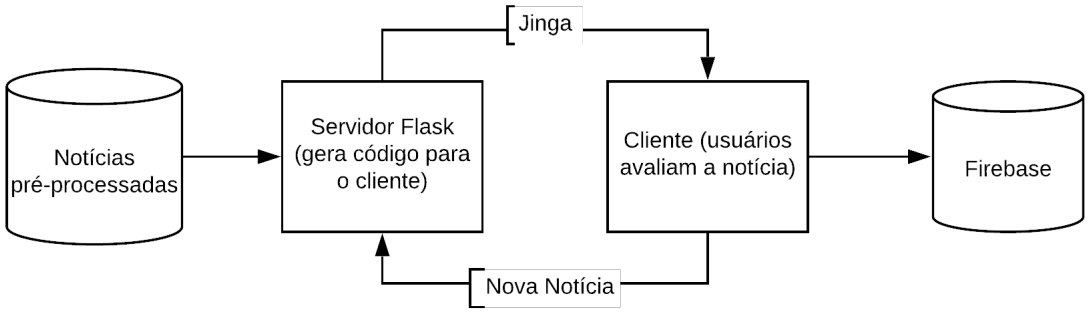
Com os topônimos identificados, foi utilizado *Web Services* baseados em REST do Geo Names[14] (Um gazetteer gratuito sob uma licença de atribuição de *creative commons*) para extrair diferentes referências geográficas associadas a cada topônimo obtido pela polyglot.

A partir dos dados dados acima, constrói-se o código fonte para a aplicação cliente, integrando à notícia informações provenientes da *Polyglot*, do *GeoNames* e da escala de classificação para a confiabiliade da resposta do usuário para topônimo.

3.1.2.2 Desenvolvimento da Aplicação

A aplicação executa o código já pré-processado e apresenta a notícia ao usuário. A Figura 4 exemplifica os passos de execução desta etapa.

Figura 4 – Fluxo de execução da aplicação



Fonte: Elaborado pelo autor

Os dados pré-processados, ou seja, as notícias com os topônimos identificados na etapa anterior, são carregados em um *template* para a exibição das notícias e renderizados para o usuário. O envio dos dados pŕe-processados pela aplicação servidora à aplicação cliente é feita através da linguagem *Jinga*, como mencionado anteriormente.

O usuário, ao abrir uma notícia no sistema, seleciona a referência geográfica que julga correta para cada um dos topônimos presentes na notícia (dentre as opções obtidas pelo Web Service do GeoNames), e avalia sua resposta com base em uma escala de classificação discreta-ordinal.

Após a classificação dos topônimos pelos usuários, estes são inseridos em uma estrutura de dados baseada no modelo apresentado em Listagem 3.1. Em seguida, os dados são enviados ao banco de dados do Firebase.

[14] https://geonames.org

Ao término da classificação de uma notícia, o usuário tem a opção de continuar avaliando mais notícias, ou, se desejar, terminar sua participação. É sugerido ao usuário relatar sua experiencia com a aplicação, apresentando possíveis melhorias ou críticas ao sistema.

3.1.2.3 Processamento dos Dados

Nesta etapa, os dados fornecidos pelos usuários são avaliados. Para cada notícia, todas as clissificações registradas são utilizadas para o cálculo do coeficiente Alfa de Cronbach, de acordo com a Equação 2.1. Assim, é possível extrair o grau de consistência entre todas as respostas fornecidas pelos usuários que avaliaram a determinada notícia.

Se, ao final desta etapa, o número de usuários para a notícia for maior ou igual a 10[15,16], e o valor do coeficiente estiver dentro do limiar estabelecido na literatura, a notícia será classificada como "Aceita" no experimento. Caso contrário, será classificada como "Rejeitada". As notícias serão armazenadas conforme o modelo apresentado em Listagem 3.2.

3.2 Estrutura dos Dados

Nesta seção é apresentado as estruturas de dados utilizadas para o armazenamento de dados na aplicação.

3.2.1 Estrutura dos Dados Após a Classificação do Usuário

Após as classificações dos topônimos presentes em uma notícia, os dados fornecidos pelo usuário são armazenados de acordo com a estrutura de dados apresentada em Listagem 3.1. Esses dados são enviados ao Firebase, e utilizados na terceira etapa da aplicação, ou seja, no processamento de todos os dados fornecidos pelos usuários para cada uma das notícias.

[15] Não há um consenso na literatura para um número ideal de usuários. Os valores são definidos de acordo com critérios adotados pelos desenvolvedores do questionário.

[16] Foi determinado, pelo método de inspeção, que um mínimo de 10 avaliações já eram suficientes para a análise de uma notícia.

Listagem 3.1 – Estrutura de dados para cada classificação do usuário

```
{
    "url": "..."
    "toponym_classifications": [
        {
            "questionary_value": ...,
            "toponym_geonamesId": "...",
            "toponym_selected": "...",
            "user_confiability": ...,
        },
        .
        .
        .
    ]
}
```

Abaixo segue a descrição de cada um dos atributos presentes em Listagem 3.1:

- "url" : Contém o link de acesso para a notícia em sua fonte.
- "toponym_classifications" : Apresenta os dados fornecidos pelos usuários durante a avaliação de uma notícia.
  - → "questionary_value" : É um valor que mapeia cada opção de topônimo presente para o usuário (*string*) em um número único. É uma representação numérica para atributos de uma escala nominal. Este número é utilizado posteriormente para o cálculo do coenficiente Alfa de Cronbach de acordo com a Equação 2.1.
  - → "toponym_geonamesId" : É o identificador do topônimo escolhido na plataforma *GeoNames*. Este valor pode ser utilizado futuramente para obter informações mais detalhadas sobre o topônimo. Para os casos em que a resposta não corresponde a um topônimo listado, o valor é definido como 0000000.
  - → "top_selected_by_user": É o topônimo escolhido pelo usuário.
  - → "user_confiability": É a confiabilidade que o usuário julga possuir para o topônimo escolhido.

3.2.2 Estrutura dos Dados para Notícias já Classificadas

Para armazenar os dados das notícias já processadas e classificadas, foram utilizados arquivos no formato JSON. Listagem 3.2 apresenta uma estrutura de dados genérica para os dados referentes a uma determinada notícia.

Listagem 3.2 – Estrutura de dados para notícias processadas e classificadas

```
{
    "url": "..."

    "toponyms": [

        {
            "std_deviation": ...,
            "toponym_geonamesId": "...",
            "top_find_on_new": "...",
            "mean_confiability": ...,
            "top_selected_by_user": "..."
        },
        .
        .
        .

    ],

    "number_of_voters": ...,

   "cronbach":...,

   "title":"..."
}
```

É apresentado, na sequência, a descrição de cada um dos atributos presentes em Listagem 3.2:

- "url" : Contém o link de acesso para a notícia em sua fonte.
- "toponyms" : Apresenta as informações extraídas, por topônimo, de todas as classificações dos usuários para a determinada notícia.
  - → "top_find_on_new" : Representa o topônimo encontrado pela biblioteca *Polyglot*.
  - → "top_selected_by_user": Representa o topônimo com o maior número de classificações(maioria simples) pelos usuários. Este topônimo é escolhido de acordo com as possíveis localizações extraídas do *GeoNames*.

→ “toponym_geonamesId“ : É o identificador do topônimo escolhido na plataforma *GeoNames*. Caso a escolha não seja um topônimo, ou este não esteja listado nas opções, é definido como 0000000.

→ “mean_confiability“: Valor médio de confiabilidade da resposta de acordo com o atributo “ *top_selected_by_user* “

→ “std_deviation“ : Representa o desvio padrão do valor médio de confiabilidade da resposta calculado no atributo anterior *(“mean_confiability“)*.

- “number_of_voters“ : Representa a quantidade de usuários que avaliaram os topônimos presentes em uma notícia.
- “cronbach“ : Representa o coeficiente Alfa de Cronbach para a notícia. O cálculo é feito utilizando a Equação 2.2.
- “title“ : Contém o título da notícia avaliada.

3.3 Interface do Sistema

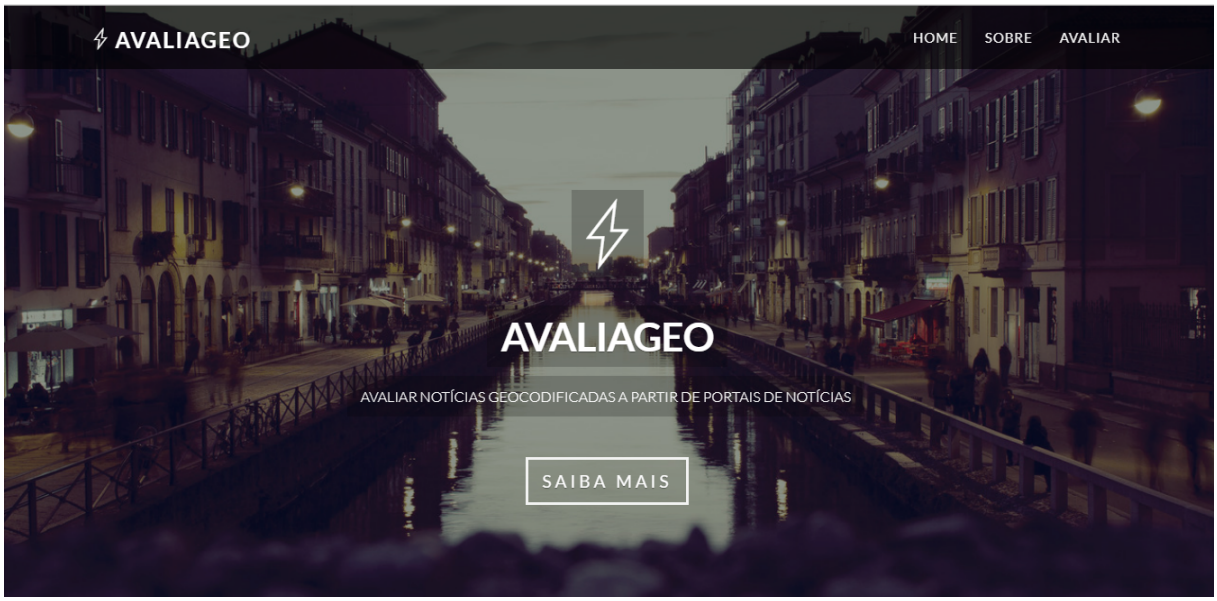
Durante a implementação do sistema, o objetivo foi desenvolver uma interface de fácil utilização, simples e objetiva, de forma a atrair a atenção do usuário. Como complemento, foi adicionado um vídeo explicativo à página inicial, demostrando a classificação de uma nóticia.

A Figura 5 apresenta a tela inicial do sistema. Nesta, é possível verificar uma descrição sucinta do objetivo do sistema. Há a possibilidade de obter mais informações logo abaixo através do botão ’Saiba Mais’. Além desse, na parte superior à direita, o botão ’Sobre’ desempenha a mesma função.

Ao clicar em qualquer dos botões descritos acima, o usuário será redirecionado, e será apresentado uma descrição mais detalhada sobre o projeto, assim como um vídeo explicativo. Esse vídeo apresenta um exemplo de avaliação de uma notícia no sistema.

Ainda na tela inicial, caso o usuário deseje ir direto para a classificação das notícias, há um botão, na parte superior à direita, chamado ’Avaliar’, em que o usuário será redirecionado para a página de avaliação. Útil caso o usuário seja reincidente no sistema e, por conseguinte, espera-se que já saiba como utilizar o sistema.

Figura 5 – Tela inicial

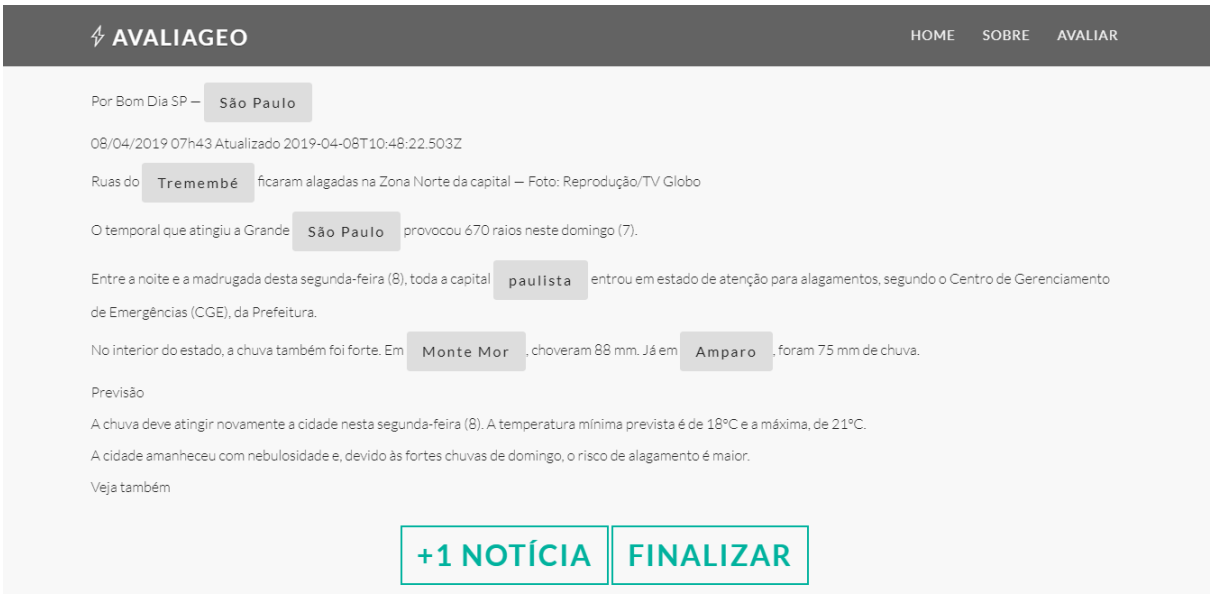


Fonte: Elaborado pelo autor

Na Figura 6, é possível verificar o corpo da notícia. Nota-se que os topônimos, identificados pela *polyglot*, estão aparecendo em botões destacados no texto. Esses botões, quando ativados, apresentaram as possibilidades de classificação para cada topônimo.

Ao fim da notícia, após a classificação de cada topônimo, há duas possibilidades de ações para o usuário: primeira, avaliar uma nova notícia, e, segunda, finalizar sua participação no sistema.

Figura 6 – Corpo na notícia



Fonte: Elaborado pelo autor

Na Figura 7, é apresentado o processo de classificação de um topônimo. O usuário já avaliou os dois primeiros topônimos no notícia, identificados pela diferença de coloração nos botões. A avaliação do terceiro topônimo está em andamento, tendo o usuário escolhido o topônimo que julga correto, e está determinando o grau de confiabilidade associado a sua escolha.

Ao termino destas duas etapas, escolher o topônimo e determinar o grau de confiabilidade, o usuário confirma suas escolhas, e prossegue para o próximo topônimo.

Figura 7 – Avaliação da confiabilidade de topônimo na notícia

Fonte: Elaborado pelo autor

Finalizada a notícia, e decidindo não realizar uma nova avaliação, o usuário é redirecionado para a tela final do sistema, demonstrada na Figura 8. Nesta, é possível a apresentação de sugestões e/ou críticas ao sistema.

Os comentários a respeito do sistema são opcionais, e, caso não deseje realizá-los, pode-se apenas finalizar a participação clicando no botão 'Finalizar'. Ao fazê-lo, retorna a página inicial do sistema.

Figura 8 – Comentários do usuário sobre o sistema

Fonte: Elaborado pelo autor

4 Resultados

Neste capítulo, são apresentados os resultados obtidos durante o desenvolvimento do trabalho. Estes foram obtidos através dos dados que os usuários do sistema forneceram no decorrer do período em que o sistema esteve disponível para uso.

4.1 Teste

Inicialmente, foi feito um pré-teste para realizar alguns ajustes na interface do sistema, assim como, determinar o número mínimo de avaliações necessárias para validar a notícia. Esse limiar foi determinado através da inspeção das respostas dos usuários. Feitos os ajustes, o sistema foi disponibilizado.

O sistema ficou disponível para uso durante 45 dias. Neste período, foi utilizado um *preset* de 20 notícias que foram pré-processadas e disponibilizadas no sistema em um banco de notícias.

A ordem em que as notícias aparecem para os usuários foi aleatória, entretanto, após uma notícia ser classificada, esta não aparece mais na sessão do usuário, até que todas as outras já tenham sido classificadas. Evitando, assim, que um usuário avalie a mesma notícia antes das demais disponíveis no sistema.

A seguir, como exemplo, são apresentadas as etapas para a classificação da Notícia 12 [1], uma das 20 notícias utilizadas no experimento. Para as demais, o procedimento foi realizado de forma semelhante. Os resultados obtidos em todas as notícias utilizadas neste experimento, incluindo a citada acima, são apresentados na Tabela 2.

4.1.1 Exemplo de base de dados avaliada

O corpo da Noticia 12, utilizada como exemplo, é transcrito na Figura 9. As palavras destacadas representam os topônimos identificados pela *polyglot*. A numeração dos topônimos utilizada na Tabela 1 segue a ordem de cronológica apresentada na notícia.

[1] https://g1.globo.com/mg/minas-gerais/noticia/2019/03/20/ipva-2019-em-mg-prazo-para-pagar-3a-parcela-termina-quarta.ghtml

Figura 9 – Corpo da Notícia 12

IPVA 2019 em MG: prazo para pagar 3a parcela termina quarta

Por G1 Minas - Belo Horizonte

Anel Rodoviário, em Belo Horizonte - Foto: Reprodução/TV Globo

O prazo do pagamento da terceira parcela do Imposto sobre a Propriedade de Veículos Automotores (IPVA) de 2019 termina nesta quarta-feira (20) para as placas de finais 9 e 0. O calendário de pagamento da última parcela começou dia 14 para as placas de finais 1 e 2.

O atraso gera multa de 0,3% ao dia. Se a inadimplência for maior que 30 dias, o acréscimo será de 20% sobre o valor do imposto devido.

Neste ano, o estado deve arrecadar R$ 5,44 bilhões com IPVA para um total de 9,7 milhões de veículos emplacados até 19 de outubro do ano passado.

O contribuinte pode pagar o IPVA em caixas eletrônicos ou imprimir a guia do imposto no site do Departamento de Trânsito de Minas Gerais (Detran) e fazer o pagamento nos bancos credenciados. Confira abaixo os locais disponíveis:

Banco do Brasil
Mais BB (correspondente bancário do BB)
Banco Postal (correspondente bancário BB)
Santander
Caixa Econômica Federal
Agências Lotéricas (correspondentes bancários da Caixa)
Sistema Financeiro Cooperativo do Brasil (Sicoob)
Mercantil do Brasil

Fonte: Elaborado pelo autor

A partir dos dados coletados dos usuários, foi extraído o atributo *"questionary_value"*. Este valor mapeia o nome do topônimo escolhido pelo usuário para um número específico. Esse atributo é utilizado para diferenciar as possibilidades de escolha, fazendo a representação ordinal da escala nominal utilizada. Mais especificamente, foram utilizados números inteiros no intervalo $[0..n]$, sendo n o número de opções disponíveis para o usuário para cada topônimo.

Com estes valores, é possível calcular o coeficiente Alfa de Cronbach para as classificações de todos os avaliadores da notícia. Podendo, assim, medir a consistência

entre as respostas dos usuários para a notícia. A Tabela 1 apresenta os dados utilizados para o cálculo do coeficiente Alfa de Cronbach.

Tabela 1 – Classificação dos topônimos da Notícia 12

Avaliadores	Topônimos da notícia					
	01	02	03	04	05	06
A	0	0	1	5	5	5
B	0	0	1	0	0	0
C	0	0	1	0	0	0
D	0	0	1	5	5	5
E	0	0	1	5	5	5
F	0	0	1	0	0	0
G	0	0	1	5	5	5
H	0	0	1	0	0	0
I	0	0	1	0	0	0
J	0	0	1	0	0	0
K	0	0	1	0	0	0
L	0	0	1	0	0	0
M	0	0	1	0	0	0
N	0	0	1	0	0	0
O	0	0	1	0	0	0
P	0	0	1	0	0	0
Q	0	0	1	0	0	0
R	0	0	1	5	5	5
S	0	0	1	0	0	0
T	0	0	1	0	0	0

Com os dados da Tabela 1, é possível aplicar a Equação 2.2, utilizada para a obtenção do coeficiente, e determinar o grau de consistência das respostas. Para a Notícia 12, o valor de $k = 6$, $\sum_{i=1}^{6} \sigma_i^2 = 15.3508$ e $\sigma_\tau = 46.0526$.

$$\alpha = \frac{k}{k-1} \times \left[1 - \frac{\sum_{i=1}^{k} \sigma_i^2}{\sigma_\tau^2}\right] \tag{4.1}$$

$$\alpha = \frac{6}{5} \times \left[1 - \frac{15.3508}{46.0526}\right] \tag{4.2}$$

$$\alpha = 0.8 \tag{4.3}$$

Como a notícia possui uma quantidade superior ao limiar definido de 10 avaliações, e o coeficiente alfa de cronbach está no limiar aceitável na literatura, a notícia foi classificada como “Aceita“.

Com uma análise da Tabela 1, é possível observar que os três primeiros topônimos obtiveram a mesma classificação de todos os participantes. Já para as três últimas houve divergências entre as respostas.

Verificando o corpo da notícia, é possível notar a presença do nome "Brasil" nos três últimos topônimos. O topônimo mais comumente associado a este nome, refere-se ao país "Brasil". Entretanto, no contexto da notícia, "Brasil" é apenas uma parte do nome de topônimos com nome composto, os quais: "Banco do Brasil", "Sistema Financeiro Cooperativo do Brasil (Sicoob)" e "Mercantil do Brasil".

Para determinar as estatísticas descritivas (média e desvio padrão) para a confiabilidade da resposta, foi considerado o topônimo com o maior número de classificações (maioria simples). Os dados de confiabilidade para os demais foram desconsiderados no cálculo.

Na Listagem 4.1, pode ser verificado a de base de dados referente à Notícia 12 analisada e concluída.

Listagem 4.1 – Exemplo de base de dados concluída

```
{
    "url": "https://g1.globo.com/mg/minas-gerais/noticia
            /2019/03/20/ipva-2019-em-mg-prazo-para-pagar-3a
            -parcela-termina-quarta.ghtml",

    "toponyms": [
        {
            "std_deviation": 0.0,
            "toponym_geonamesId": "3470127",
            "top_find_on_new": "Belo Horizonte",
            "mean_confiability": 5.0,
            "top_selected_by_user": "Belo Horizonte PPLA"
        },
        {
            "std_deviation": 0.0,
            "toponym_geonamesId": "3470127",
            "top_find_on_new": "Belo Horizonte",
            "mean_confiability": 5.0,
            "top_selected_by_user": "Belo Horizonte PPLA"
        },
        {
            "std_deviation": 0.0,
            "toponym_geonamesId": "3457153",
```

```
            "top_find_on_new": "Minas Gerais",
            "mean_confiability": 5.0,
            "top_selected_by_user": "Minas Gerais ADM1"
        },
        {
            "std_deviation": 0.9155,
            "toponym_geonamesId": "3469034",
            "top_find_on_new": "Brasil",
            "mean_confiability": 4.1333,
            "top_selected_by_user": "Brasil PCLI"
        },
        {
            "std_deviation": 0.7432,
            "toponym_geonamesId": "3469034",
            "top_find_on_new": "Brasil",
            "mean_confiability": 4.5333,
            "top_selected_by_user": "Brasil PCLI"
        },
        {
            "std_deviation": 0.9155,
            "toponym_geonamesId": "3469034",
            "top_find_on_new": "Brasil",
            "mean_confiability": 4.1333,
            "top_selected_by_user": "Brasil PCLI"
        }
    ],

    "number_of_voters": 20,

    "cronbach": 0.808,

    "title": "IPVA 2019 em MG: prazo para pagar 3a parcela
            termina quarta "
}
```


4.1.2 Análise dos Resultados

Os resultados obtidos com a realização do experimento para as demais notícias são apresentados na Tabela 2.

Tabela 2 – Resultados obtidos com as classificações das notícias

	Número de topônimos	Coeficiente alfa de cronbach	Número de avaliações	Status
Notícia 01	4	1.0	14	Aceita
Notícia 02	6	0.9914	3	Rejeitada
Notícia 03	8	0.9986	11	Aceita
Notícia 04	3	0.8425	20	Aceita
Notícia 05	7	0.88	9	Rejeitada
Notícia 06	8	0.9644	14	Aceita
Notícia 07	7	0.9656	6	Rejeitada
Notícia 08	12	0.9988	7	Rejeitada
Notícia 09	8	0.9882	15	Aceita
Notícia 10	6	0.9895	12	Aceita
Notícia 11	6	0.9855	12	Aceita
Notícia 12	6	0.808	20	Aceita
Notícia 13	11	0.9674	10	Aceita
Notícia 14	8	1.0	11	Aceita
Notícia 15	6	0.9164	10	Aceita
Notícia 16	6	1.0	2	Rejeitada
Notícia 17	4	0.4954	12	Rejeitada
Notícia 18	3	0.9956	16	Aceita
Notícia 19	12	0.9984	13	Aceita
Notícia 20	5	0.9803	11	Aceita

Somando o número de avaliações de cada notícia - a partir dos dados apresentados na coluna “Número de avaliações“ da Tabela 2 - é possível verificar que houve 228 avaliações, ou seja, 228 participações no trabalho. Durante o experimento, foi possível gerar 14 notícias classificadas como “Aceita“ e 6 notícias classificadas como “Rejeitada“.

As Notícias 01, 14 e 16, apresentaram um coeficiente 1.0, o valor máximo possível. Com uma análise detalhada destas notícias, é possível verificar que alguns topônimos, presentes nas Notícias 01 e 14, se repetem ao longo da notícia, além de representarem exatamente as referencias mais conhecidas associadas a estes locais.

Já a noticia 16, analisando o número de participantes na notícia, é possível verificar que a quantidade de avaliadores foi muito baixa, o que por si só rejeita a notícia. Avaliando o impacto desta quantidade baixa de avaliadores, é possível determinar que o cálculo da consistência dos dados é tendenciosa, seja para um valor muito alto (como é o caso desta notícia), ou para uma valor muito baixo.

É possível notar que, em alguns casos, os coeficientes Alfa de Cronbach apresentam valores superiores a 0.9, considerado como um limite superior. Entretanto, esses valores são esperados, já que, na própria literatura, são justificados por itens duplicados no questionário.

Interpretando esta informação no trabalho, a duplicação de itens significa um mesmo topônimo aparecendo mais de uma vez na notícia. Não obstante, não pode ser retirado, já que é informação útil para a construção dos algoritmos de GSR. Assim, na

presença de topônimos repetidos, aliados com determinada consistência na resposta dos usuários, podem levar a valores superiores a 0.9.

5 Considerações Finais

Este trabalho apresentou o desenvolvimento de uma ferramenta que possibilita a criação de bases de notícias contendo as informações geográficas presentes nos textos. A validação dos topônimos foi realizada utilizando o coeficiente Alfa de Cronbach. Esse coeficiente determina o grau de consistência entre as respostas fornecidas pelos usuários para dada notícia.

Assim, apesar de não se poder determinar se a classificação do topônimo está correta, ou seja, se reflete exatamente a intenção do autor da matéria, é possível determinar que as respostas de todos os usuários estão, de certa forma, convergindo.

Neste trabalho, foi possível classificar como "Aceita" 14 das 20 notícias disponibilizadas para serem classificadas. Isso representa 70% de todas as notícias utilizadas no experimento. Por conseguinte, podem ser utilizadas como bases para avaliar soluções, ou seja, algoritmos de GSR, disponíveis.

Das notícias rejeitadas no experimento, as que possuem um corpo de texto maior, em grande parte, não conseguiram atingir a quantidade mínima de avaliações necessárias. Isso pode significar que os usuários não desejaram terminar a validação da notícia, já que, para aceitar a classificação da notícia, todos os topônimos, assim como a confiabilidade da resposta, devem ser avaliados.

Para as notícias em que o coeficiente ficou abaixo do limite estabelecido na literatura, o valor é justificado pela presença de topônimos que possuem nomes similares a outras referências geográficas, entretanto mais populares. Este é o caso, por exemplo, do topônimo "Brasil", que mais comumente é associado ao país. Entretando, existem outras referências que levam seu nome, como: " Banco do Brasil " e " Sistema de Cooperativas de Crédito do Brasil (Sicoob) ".

Ou, ainda, os que se referem a locais distantes da localização onde a maioria dos participantes do experimento residem. Como exemplo, a Notícia 17, que se refere a localidades do Distrito Federal(DF), sendo a maioria dos participantes do experimento residentes em Minas Gerais (MG).

Em trabalhos futuros, há a possibilidade da implementação de uma funcionalidade em que os usuários possam adicionar topônimos que não foram identificados previamente pelos algoritmos de detecção. Este tipo de trabalho se mostra mais complexo, já que, como a quantidade de referências geográficas em uma notícia não é a mesma para todos os avaliadores, análises estatísticas mais elaboradas são necessárias.

É interessante ainda, incorporar ao sistema a possibilidade de carregamento do

texto a ser validado pelo usuário no sistema. Assim como utilizar outras bibliotecas, além da *Polyglot*, para a identificação dos topônimos. Permitindo, dessa forma, verificar as notícias classificadas de acordo com os diferentes algoritmos de identificação de topônimos.

Por fim, o desenvolvimento de uma maior quantidade de testes pode ser interessante. Possibilitando uma maior robustez dos resultados obtidos durante o desenvolvimento do projeto, já que a quantidade de dados será maior.

Referências

ALEX, B. et al. Homing in on twitter users: Evaluating an enhanced geoparser for user profile locations. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. [S.l.]: European Language Resources Association (ELRA), 2016. p. 3936–3944. ISBN 978-2-9517408-9-1. Citado na página 15.

ALMEIDA, D.; SANTOS, M. d.; COSTA, A. F. B. Aplicação do coeficiente alfa de cronbach nos resultados de um questionário para avaliação de desempenho da saúde pública. *XXX Encontro Nacional de Engenharia de Produção*, Associação Brasileira de Engenharia de Produção São Paulo, v. 15, p. 1–12, 2010. Citado na página 22.

CORTINA, J. M. What is coefficient alpha? an examination of theory and applications. *Journal of applied psychology*, US: American Psychological Association, v. 78, n. 1, p. 98, 1993. Citado na página 20.

CRONBACH, L. J. Coefficient alpha and the internal structure of tests. *psychometrika*, Springer, v. 16, n. 3, p. 297–334, 1951. Citado na página 20.

DORAN, C. et al. *ACE 2005 English SpatialML Annotations Version 2*. 2005. Disponível em: <https://catalog.ldc.upenn.edu/LDC2011T02>. Citado na página 16.

FREITAS, A. L.; RODRIGUES, S. A avaliação da confiabilidade de questionários: uma análise utilizando o coeficiente alfa de cronbach. In: . [S.l.: s.n.], 2005. Citado na página 22.

GRITTA, M. et al. What's missing in geographical parsing? *Language Resources and Evaluation*, Springer, v. 52, n. 2, p. 603–623, 2018. Citado 2 vezes nas páginas 16 e 19.

LARSEN, N. *Market Segmentation - A framework for determining the right target customers*. Dissertação (Bachelor's Thesis) — Aarhus School of Business, Aarhus BSS, Denmark, 5 2010. Citado na página 15.

LAWAL, B.; LAWAL, H. B. *Categorical data analysis with SAS and SPSS applications*. [S.l.]: Psychology Press, 2003. Citado na página 21.

LEONTITSIS, A.; PAGGE, J. A simulation approach on cronbach's alpha statistical significance. *Mathematics and Computers in Simulation*, Elsevier, v. 73, n. 5, p. 336–340, 2007. Citado 2 vezes nas páginas 20 e 21.

MARATEB, H. R. et al. Manipulating measurement scales in medical statistical analysis and data mining: A review of methodologies. *Journal of research in medical sciences: the official journal of Isfahan University of Medical Sciences*, Wolters Kluwer–Medknow Publications, v. 19, n. 1, p. 47, 2014. Citado na página 21.

MATTHIENSEN, A. Uso do coeficiente alfa de cronbach em avaliações por questionários. *Embrapa Roraima-Documentos (INFOTECA-E)*, Boa Vista, RR: Embrapa Roraima, 2010., 2010. Citado na página 22.

MIDDLETON, S. E.; MIDDLETON, L.; MODAFFERI, S. Real-time crisis mapping of natural disasters using social media. *IEEE Intelligent Systems*, IEEE, v. 29, n. 2, p. 9–17, 2014. Citado na página 15.

MONTEIRO, B. R.; DAVIS, C. A.; FONSECA, F. T. A survey on the geographic scope of textual documents. *Computers & Geosciences*, v. 96, p. 23–34, 2016. ISSN 0098-3004. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0098300416301972>. Citado 4 vezes nas páginas 15, 16, 18 e 19.

RUPP, C. et al. Customising geoparsing and georeferencing for historical texts. In: IEEE. *Big Data, 2013 IEEE International Conference on*. [S.l.], 2013. p. 59–62. Citado na página 15.

STEVENS, S. S. et al. On the theory of scales of measurement. Bobbs-Merrill, College Division, 1946. Citado na página 21.

STREINER, D. L. Being inconsistent about consistency: When coefficient alpha does and doesn't matter. *Journal of personality assessment*, Taylor & Francis, v. 80, n. 3, p. 217–222, 2003. Citado na página 21.

STREINER, D. L. Starting at the beginning: an introduction to coefficient alpha and internal consistency. *Journal of personality assessment*, Taylor & Francis, v. 80, n. 1, p. 99–103, 2003. Citado na página 21.