

Geocodificação e validação de bases de dados de notícias

Arthur Domingues
Bruno Monteiro

Universidade Federal de Ouro Preto

7 de novembro de 2019

- 1 Introdução
- 2 Problema
- 3 Objetivos
- 4 Desenvolvimento
- 5 Resultados
- 6 Considerações Finais
- 7 Referências

- Crescimento na disponibilidade de informações textuais
 - Dispositivos que podem se conectar a internet
 - Mídias sociais
- Oportunidade de utilizar estes dados para a obtenção de informações geográficas
- Adicionando uma outra dimensão no processo de análise de dados
 - do que
 - quando
 - ***onde***

determinado texto se refere [Gritta et al., 2018]

- *Geographic Scope Resolution*, ou GSR, é um problema que objetiva a determinação do escopo geográfico de textos e documentos
- Para isso, é necessário a identificação e desambiguação de topônimos.
- A solução do GSR é dividida em etapas:
 - *Geoparsing*
 - *Reference Resolution*
 - *Grounding References*

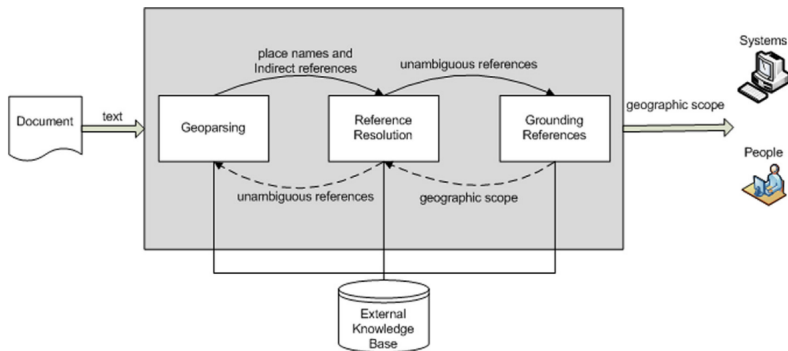


Figura: Etapas do GSR.

Fonte: [Monteiro et al., 2016]

- Alguns exemplos de aplicações do GSR
 - Indexação e rankeamento geográfico em motores de busca [Monteiro et al., 2016]
 - Sumarização de textos históricos [Rupp et al., 2013]
 - Extração de informação: tempo real e ao longo do tempo [Middleton et al., 2014, Alex et al., 2016]

- De acordo com [Monteiro et al., 2016], um dos problemas para o avanço nas soluções é a utilização de algoritmos e bases de dados próprias nas abordagens propostas
 - *Geoparsing*: ontologia e regex, heurísticas de linguagem
 - *Reference Resolution*: alg. Geométricos, aprendizagem de máquina
 - *Grounding References*: um único escopo, estrutura de dados (árvore, grafos)
- Comparação e cooperação fica comprometida
- Um dos desafios da área é a falta de bases de dados geocodificadas gratuitas [Gritta et al., 2018]
 - Base de dados com 210065 palavras, sendo 17821 palavras únicas, custo : \$1000,00 [Doran et al., 2005]

- Construir uma aplicação WEB para geração e validação de bases de dados geocodificadas
- As bases de dados terão notícias como objeto de pesquisa
- A validação dos dados será realizada manualmente pelos usuários
- A validação dos dados fornecidos pelos usuários será obtida utilizando o Alfa de Cronbach[Cronbach, 1951]

- Proposto por Lee J. Cronbach em 1951 Cronbach[Cronbach, 1951]
- Uma das ferramentas estatísticas mais importantes em pesquisas que envolvem testes e validação [Leontitsis and Pagge, 2007]
- É a média das correlações entre os itens que fazem parte de um determinado estudo
- É uma propriedade inerente do padrão de resposta da população estudada
- Não uma característica da escala por si só; ou seja, o valor de alfa sofre mudanças segundo a população na qual se aplica a escala [Streiner, 2003]
- Valores aceitáveis entre 0.7 e 0.9

- O coeficiente alpha pode ser calculado a partir da seguinte equação

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

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 ;
- k é o número de itens no questionário ($k > 1$)
- n é a quantidade de respostas no questionário ($n > 1$)

- Cliente/Servidor
 - Tecnologias Cliente
 - 1 HTML 5
 - 2 CSS 3 (Bootstrap)
 - 3 JavaScript (Jquery)
 - Tecnologias Servidor
 - 1 Python 2.7 (Flask)
 - 2 Jina 2
 - 3 Firebase
 - 4 Git
 - 5 Heroku

- A Aplicação foi dividida em três partes
 - Pré-Processamento
 - Desenvolvimento da aplicação
 - Processamento dos dados

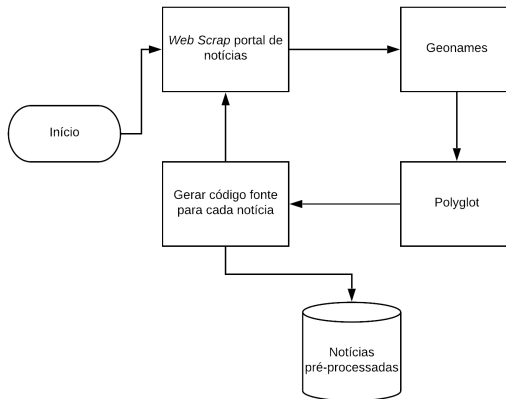


Figura: Fluxo de execução do pré-processamento

Estrutura dos dados do usuário

```
1 {  
2   "url": "..."  
3   "toponym_classifications": [  
4     {  
5       "questionary_value": ...,  
6       "toponym_geonamesId": "...",  
7       "toponym_selected": "...",  
8       "user_confiability": ...,  
9     },  
10    .  
11    .  
12    .  
13  ]  
14 }
```

Figura: Estrutura dos dados do usuário

Fonte: Próprio Autor

- Cálculo do coeficiente de cronbach (α) para cada notícia
- Classificar a notícia de acordo com seguintes critérios:
 - Número de classificações (≥ 10)¹
 - Faixa aceitável do coeficiente alfa de cronbach (0.7 a 0.9)
- Notícia será:
 - ACEITA, se atender ambos os requisitos
 - REJEITADA, caso contrário

¹Não há consenso na literatura


```
1 {
2   "url": "...",
3
4   "toponyms": [
5
6     {
7       "std_deviation": ...,
8       "toponym_geonamesId": "...",
9       "top_find_on_new": "...",
10      "mean_confiability": ...,
11      "top_selected_by_user": "..."
12    },
13    .
14    .
15    .
16
17  ],
18
19  "number_of_voters": ...,
20
21  "cronbach":...,
22
23  "title": "..."
24 }
```

Figura: Estrutura de dados da notícia Concluída

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 locais 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 de **Brasil**

Figura: Corpo da notícia

Exemplo

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

Figura: Tabela para o cálculo do coeficiente alfa de cronbach

- Calculando o coeficiente alfa de cronbach, temos:

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

$$\alpha = \frac{6}{5} \times \left[1 - \frac{15.3508}{46.0526} \right] \quad (3)$$

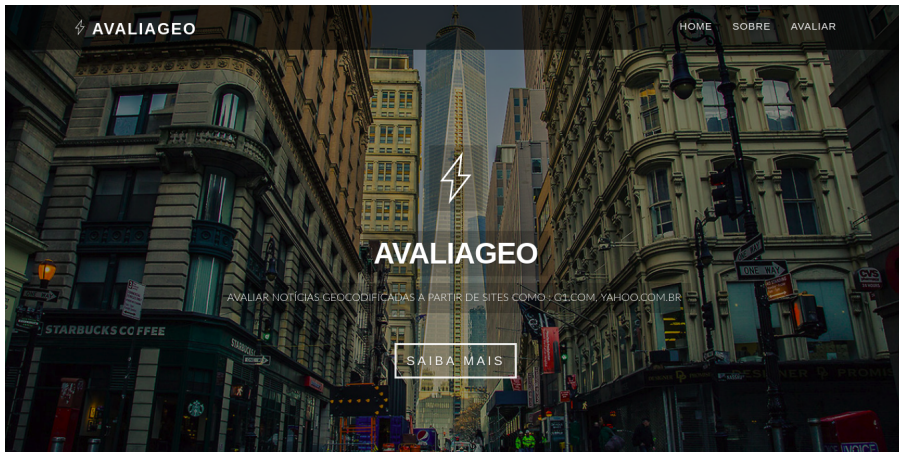
$$\alpha = 0.8 \quad (4)$$

Resultados

	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

- Resultados obtidos
 - 70 % das notícias foram classificadas como ACEITAS
 - 30 % restantes ...
 - Semelhança entre topônimos (BB, Brasil, Sicoob)
 - Localização do participante diferente da referência(DF, SP)

- Trabalhos futuros
 - Outras bibliotecas NLP
 - Permitir usuário identificar topônimo não identificado



UM POUCO SOBRE O PROJETO

UM SISTEMA QUE PERMITE AOS USUÁRIOS CLASSIFICAR REFERÊNCIAS GEOGRÁFICAS PRESENTE EM NOTÍCIAS

Olá Tudo Bem, meu nome é Arthur, sou estudante do curso de Engenharia de Computação da Universidade Federal de Ouro Preto, no campos de João Monlevade, mais especificamente no Instituto de Ciencias Exatas e Aplicadas. Este projeto está sendo desenvolvido como Trabalho de Conclusão de Curso e tem como idéia central desenvolver um sistema que servirá de auxilio para a geração e validação de bases de dados geocodificadas. Apesar do nome ser um pouco extenso, a idéia é simples, mas primeiro vou explicar do que se trata o problema.

Hoje a quantidade de informação disponível na Internet vem crescendo de forma exponencial, ocasionado por diversos fatores como: aumento de redes sociais, blogs, sites de noticias, etc. Logo uma pergunta natural que se faz é a seguinte: O que fazer com esta quantidade de dados??? Bom, uma das diversas aplicações possíveis é a obtenção de informação geográfica, ou seja, saber a qual lugar geográfico determinada mídia(texto, áudio, video) se refere. Com a obtenção desta informação é possível desenvolver soluções personalizadas para determinados grupo de pessoas baseados em sua localização.

Para a realização do trabalho é utilizado processamento de textos para a obtenção da referência geográfica. Neste processo, um algoritmo (ou um conjunto deles) é utilizado para tentar inferir tal informação, utilizando para isso um conjunto de parâmetros para a tomada de decisão. Entretanto os algoritmos podem apresentar falhas, podendo estas falhas ser ocasionada por : falta de parâmetros, quantidade de parâmetros não são suficientes para obter informações corretas, parâmetros utilizados podem ter pesos não ajustados para a sua aplicação na decisão tomada.

POBREZA EXTREMA CRESCE EM 25 ESTADOS BRASILEIROS, APONTA ESTUDO

Lembrando que : Local a que se refere Quão certo de sua resposta você está?

Por Tais Laporta, G1

10/10/2018 07h00 Atualizado 2018-10-10T14:01:08.343Z

O percentual de famílias que vivem em extrema pobreza aumentou em quase todos os estados do Brasil PCLI nos últimos quatro anos, em especial no Grande Região Nordeste RGN, apontou um estudo feito pela Tendências Consultoria.

A condição de extrema pobreza atinge pessoas com renda familiar per capita de até R\$ 85 por mês, segundo a medição do governo.

Na média nacional, a miséria subiu para 4,8% da população em 2017, contra 3,2% em 2014.

Nestes quatro anos, ela só não aumentou em dois dos 27 estados brasileiros, Tocantins ADM1 e Paraíba ADM1.

Adriano Pitoli, diretor da Tendências, aponta uma forte correlação entre a crise econômica e a evolução da pobreza.

pobreza extrema.

No **Maranhão ADM1** , ela chegou a 12% em 2017, o pior resultado do país.

O **Acre ADM1** , foi o estado que mais teve um aumento da pobreza extrema entre 2014 e 2017, de 5,6%.

Enquanto isso, estados do **Kwanza Sul ADM1** , e **Sudeste Asiático RGN** , estão entre os menos prejudicados pela crise, apesar da piora generalizada.

Segundo Pitoli, a maior parte dos estados da região **Grande Região Nordeste RGN** , passou por um "efeito ressaca" que levou a região a sofrer de forma mais intensa os efeitos da recessão econômica.

"O **Grande Região Nordeste RGN** , era um destaque positivo de renda e consumo nos anos anteriores à crise, com peso grande de aposentadorias, do Bolsa Família e da folha de pagamento de servidores.

Regiões mais dependentes dessa transferência de renda sofreram mais", analisa o diretor da Tendências.

Pitoli aponta que, mesmo sem cortes de benefícios e programas sociais, a redução de gastos públicos afetou os projetos de investimento do governo e pegou em cheio a região.

"As mesmas razões que deram destaque à região nos anos anteriores levaram a uma piora maior na crise".

O estudo ainda não levantou os dados de 2018, mas a expectativa, segundo Pitoli, é de uma melhora muito discreta na taxa de extrema pobreza no país, devido à lenta recuperação da economia.

NOVA NOTÍCIA**FINALIZAR**



Alex, B., Llewellyn, C., Grover, C., Oberlander, J., and Tobin, R. (2016).

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)*, pages 3936–3944. European Language Resources Association (ELRA).



Cronbach, L. J. (1951).





Coefficient alpha and the internal structure of tests.

psychometrika, 16(3):297–334.



Doran, C., Mani, I., Clancy, S., and Hitzeman, J. (2005).

Ace 2005 english spatialml annotations version 2.

-  Gritta, M., Pilehvar, M. T., Limsopatham, N., and Collier, N. (2018). What's missing in geographical parsing?
Language Resources and Evaluation, 52(2):603–623.
-  Leontitsis, A. and Pagge, J. (2007).
A simulation approach on cronbach's alpha statistical significance.
Mathematics and Computers in Simulation, 73(5):336–340.
-  Middleton, S. E., Middleton, L., and Modafferi, S. (2014).
Real-time crisis mapping of natural disasters using social media.
IEEE Intelligent Systems, 29(2):9–17.
-  Monteiro, B. R., Davis, C. A., and Fonseca, F. T. (2016).
A survey on the geographic scope of textual documents.
Computers & Geosciences, 96:23–34.



Rupp, C., Rayson, P., Baron, A., Donaldson, C., Gregory, I., Hardie, A., and Murrieta-Flores, P. (2013).

Customising geoparsing and georeferencing for historical texts.

In *Big Data, 2013 IEEE International Conference on*, pages 59–62. IEEE.



Streiner, D. L. (2003).

Being inconsistent about consistency: When coefficient alpha does and doesn't matter.

Journal of personality assessment, 80(3):217–222.

Perguntas ?