

UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

**Estudo de Ferramentas Text-to-SQL para
Aplicação a Dados de Saúde Pública**

Antônio Carlos Santos de Lima

PROPOSTA DE TEMA
MAC 499— TRABALHO DE
FORMATURA SUPERVISIONADO

Supervisora: Prof^a. Dr^a. Kelly Rosa Braghetto

São Paulo
8 de Maio de 2022

Sumário

1	Introdução	1
1.1	Contexto e motivação	1
1.2	Problema computacional	2
2	Proposta	3
2.1	Objetivos	3
2.2	Metodologia	3
3	Cronograma	4
	Referências	5

Capítulo 1

Introdução

1.1 Contexto e motivação

Realizar consultas em um banco de dados usando SQL exige conhecimento prévio tanto da estrutura do banco quanto da linguagem de consulta, o que torna a tarefa árdua para usuários não técnicos.

Text-to-SQL é uma forma de converter uma sentença em linguagem natural para um comando SQL. Por exemplo, supondo que, em um banco de dados, exista a tabela "Autor" e um campo "nome", então para a sentença

Retorne todos os autores cujo nome é "Antônio"

o comando SQL correspondente deve ser

```
SELECT * FROM autor WHERE nome = "Antônio"
```

Text-to-SQL envolve técnicas e modelos das áreas de Processamento de Linguagem Natural e Aprendizado de Máquina, em especial modelos do tipo Seq2Seq, arquiteturas da forma codificador-decodificador, aprendizagem por reforço e mecanismo de atenção, muitas vezes combinados entre si numa arquitetura complexa.

Dentre os trabalhos relacionados com ferramentas text-to-SQL, há aquele realizado por [Li e JAGADISH, 2014](#), em que é proposta uma Interface de Linguagem Natural para Banco de Dados Relacionais (NaLIR, em Inglês); e o trabalho de [A. F. NOVELLO, 2021](#), no qual é apresentada a ferramenta Glamorise, para processar sentenças em linguagem natural com agregação, que o NaLIR geralmente não converte em comando SQL de forma satisfatória.

Em uma sentença com agregação, uma operação é feita com vários valores e o resultado é o retorno de um único valor, conforme explicado em [COUNT \(Transact-SQL\) 2021](#). Por exemplo, utilizando a mesma tabela "Autor" do exemplo acima e supondo que existam 10 registros nessa tabela, o comando SQL

```
SELECT COUNT(*) FROM autor
```

contém a operação COUNT, que retorna a quantidade de registros na tabela, que, nesse caso, é 10.

Na ferramenta NaLIR(LI e JAGADISH, 2014), uma sentença em linguagem natural é dada como entrada e, através de determinados processos, é gerado um comando SQL, conforme a sentença dada como entrada. No sistema Glamorise(A. F. NOVELLO, 2021), parte do processamento feito em LI e JAGADISH, 2014 é utilizada, para lidar com possíveis agregações e retornar um comando SQL mais satisfatório.

1.2 Problema computacional

No projeto de A. F. NOVELLO, 2021, o NaLIR e o Glamorise são utilizados com sentenças em Inglês, para consultar bancos de dados em Inglês. Neste trabalho de conclusão de curso, o problema considerado é a conversão de sentenças em língua portuguesa para comandos SQL, utilizando um banco de dados em Português.

Capítulo 2

Proposta

2.1 Objetivos

Visando facilitar o acesso aos dados de Saúde Pública para gestores públicos e profissionais da área de saúde, os objetivos deste trabalho são:

- Analisar as ferramentas e técnicas de software para Text-to-SQL desenvolvidas por [LI e JAGADISH, 2014](#) e [A. F. NOVELLO, 2021](#)
- Adaptar as propostas de [LI e JAGADISH, 2014](#) e [A. F. NOVELLO, 2021](#) para a utilização com sentenças em língua portuguesa
- Utilizar a versão adaptada das propostas para consulta a dados de Saúde Pública

2.2 Metodologia

Será realizada uma abordagem baseada em revisão de literatura e análise de ferramentas Text-to-SQL desenvolvidas por [LI e JAGADISH, 2014](#) e [A. F. NOVELLO, 2021](#). Nessa análise, os aspectos considerados serão a configuração de cada ferramenta e o desempenho de cada ferramenta quanto ao resultado das consultas SQL. Para a adaptação dos projetos, será necessário utilizar Spacy([EXPLOSION, 2022](#)) e NLTK([PROJECT, 2022](#)).

Além disso, será feita uma caracterização das bases de dados de Saúde Pública e uma caracterização das consultas de interesse para profissionais de saúde. Para isso, serão utilizadas os dados disponíveis em [SEADE, 2020](#). Em seguida, serão feitos experimentos com sentenças em língua portuguesa, a fim de verificar a qualidade dos comandos SQL gerados, através de determinadas métricas.

Capítulo 3

Cronograma

Tarefas	Mês1	Mês2	Mês3	Mês4	Mês5	Mês6	Mês7	Mês8	Mês9	Mês10
Estudar as ferramentas e técnicas Text-to-SQL dos projetos	X	X	X							
Adaptar o NaLIR para o Português			X	X						
Adaptar o Glamorise para o Português				X	X					
Realizar experimentos com os projetos adaptados, utilizando dados relacionados à saúde pública brasileira					X	X	X			
Análise do resultado dos experimentos							X	X		
Escrever a monografia					X	X	X	X	X	
Correção de monografia, caso necessário										X

Observação: com a ciência das professoras Nina Sumiko Tomita Hirata e Kelly Rosa Braguetto(orientadora deste trabalho), este projeto iniciou de fato em outubro de 2021, com previsão de término no fim de junho de 2022.

Referências

- [AFFOLTER *et al.* 2019] Katrin AFFOLTER, Kurt STOCKINGER e Abraham BERNSTEIN. “A comparative survey of recent natural language interfaces for databases”. Em: *The VLDB Journal* 28.5 (out. de 2019), pgs. 793–819. ISSN: 1066-8888. DOI: [10.1007/s00778-019-00567-8](https://doi.org/10.1007/s00778-019-00567-8). URL: <https://doi.org/10.1007/s00778-019-00567-8>.
- [COUNT (Transact-SQL) 2021] COUNT (Transact-SQL). 2021. URL: <https://docs.microsoft.com/pt-br/sql/t-sql/functions/count-transact-sql?view=sql-server-ver15> (acesso em 12/05/2022) (citado na pg. 1).
- [EXPLOSION 2022] EXPLOSION. *Industrial-Strength Natural Language Processing*. 2022. URL: <https://spacy.io/> (acesso em 07/05/2022) (citado na pg. 3).
- [KIM *et al.* 2020] Hyeonji KIM, Byeong-Hoon So, Wook-Shin HAN e Hongrae LEE. “Natural language to sql: where are we today?” Em: *Proc. VLDB Endow.* 13.10 (jun. de 2020), pgs. 1737–1750. ISSN: 2150-8097. DOI: [10.14778/3401960.3401970](https://doi.org/10.14778/3401960.3401970). URL: <https://doi.org/10.14778/3401960.3401970>.
- [LI e JAGADISH 2014] Fei LI e H. V. JAGADISH. “Constructing an interactive natural language interface for relational databases”. Em: *Proc. VLDB Endow.* 8.1 (set. de 2014), pgs. 73–84. ISSN: 2150-8097. DOI: [10.14778/2735461.2735468](https://doi.org/10.14778/2735461.2735468). URL: <https://doi.org/10.14778/2735461.2735468> (citado nas pgs. 1–3).
- [A. NOVELLO e CASANOVA 2020] Alexandre NOVELLO e Marco Antonio CASANOVA. “A novel solution for the aggregation problem in natural language interface to databases (nlidb)”. Em: *Anais do XXXV Simpósio Brasileiro de Bancos de Dados*. Evento Online: SBC, 2020, pgs. 217–222. DOI: [10.5753/sbbd.2020.13644](https://sol.sbc.org.br/index.php/sbbd/article/view/13644). URL: <https://sol.sbc.org.br/index.php/sbbd/article/view/13644>.
- [A. F. NOVELLO 2021] Alexandre Ferreira NOVELLO. “A Novel Solution to Empower Natural Language Interfaces to Databases (NLIDB) to Handle Aggregations”. Diss. de mest. Rio de Janeiro, Brasil: Pontifícia Universidade Católica do Rio de Janeiro, abr. de 2021 (citado nas pgs. 1–3).
- [PROJECT 2022] NLTK PROJECT. *Natural Language Toolkit*. 2022. URL: <https://www.nltk.org/> (acesso em 07/05/2022) (citado na pg. 3).

REFERÊNCIAS

- [SEADE 2020] Fundação SEADE. *Repositório de dados sobre casos e óbitos decorrentes do COVID-19 nos municípios do Estado de São Paulo e sobre leitos e internações por Departamento Regional de Saúde*. 2020. URL: <https://github.com/seade-R/dados-covid-sp> (acesso em 07/05/2022) (citado na pg. 3).