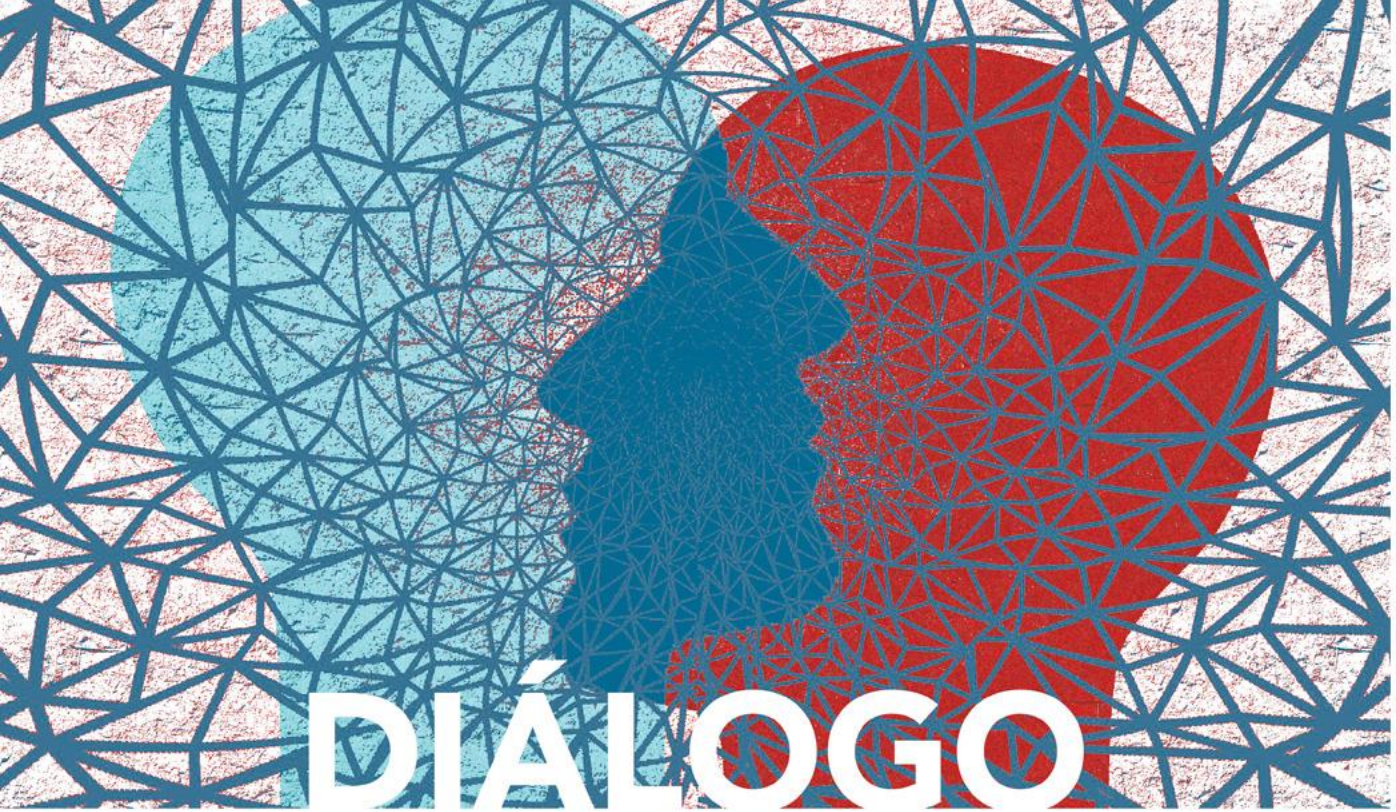


RESULTADOS DA 3ª EDIÇÃO DO PROGRAMA DE PESQUISA

Classificação automatizada de produtos da Nota Fiscal Eletrônica de Compras Públicas

Augusto Fonseca (CEFET/RJ)

Bruno Melo (TCE-RJ)

Eduardo Bezerra (CEFET/RJ)

Leonardo Lima (UFPR)

Wellington Amaral (TCE-RJ)

ECG ESCOLA DE CONTAS E GESTÃO DO TCE-RJ

Tribunal de Contas Estado do Rio de Janeiro

Sumário

- Introdução
- Fundamentos
- Metodologia
- Conclusões

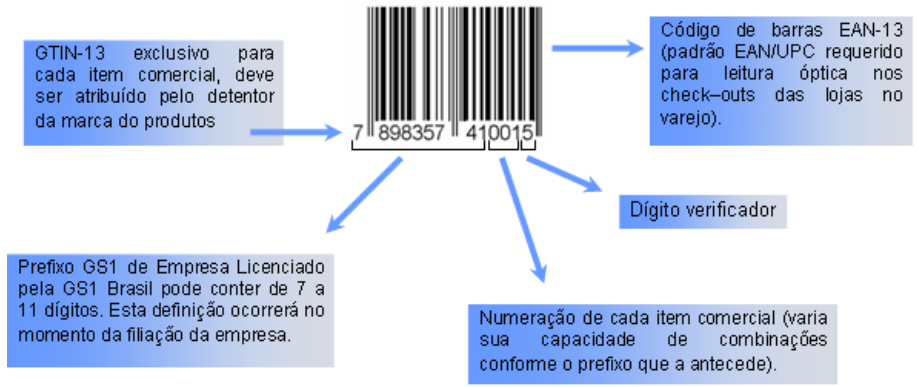
Introdução

Nota Fiscal Eletrônica (NFE)

- NFE
  - Implantada em 2006
  - Desde 2010 seu tornou obrigatória nas transações comerciais destinadas a órgãos públicos.
- Tal volume de dados gerado diariamente pode ser alvo para análises ou até mesmo servir de base para geração de modelos preditivos.
  - Exemplo: obter o preço médio de venda de um determinado produto (ou serviço), informação esta que pode ser utilizada para identificar discrepâncias nas aquisições.

EAN

- O identificador EAN (atualmente conhecido como GTIN) é um padrão criado e administrado pela GS1 Brasil.
- Em sua forma mais comum, possui 13 dígitos.
- Esse identificador estabelece a singularidade de um produto contido em algum item de uma NFE.



https://blog.bling.com.br/gtin-tres-etapas/

Problema

- Obstáculos
  - Não é raro encontrar descrições distintas em NFEs distintas que referenciam um mesmo produto comercial.
    - Posto que cada descrição é registrada como texto livre.
  - Também não é raro encontrar itens de NFEs em que não foi informado o EAN!
- Associar tais descrições ao produto (ou serviço) ao qual se referem não é uma tarefa trivial.

Problema

- Como determinar que os três itens abaixo fazem menção ao mesmo produto?

	unidade	quantidade	valor_unitario	descricao	ean
328519	UN	1.0	4.69	PARACETAMOL 750 TE.20C	7896112149705
330149	CX	6.0	2.33	PARACETAMOL TEUTO GEN 750MG C/20 CPR	7896112149705
331799	Cx	7.0	21.25	Tylenol 750mg - Cx c/ 20 comprimidos	N/I

Problema

- Como determinar que os três itens abaixo fazem menção ao mesmo produto?

	unidade	quantidade	valor_unitario	descricao	ean
328519	UN	1.0	4.69	PARACETAMOL 750 TE.20C	7896112149705
330149	CX	6.0	2.33	PARACETAMOL TEUTO GEN 750MG C/20 CPR	7896112149705
331799	Cx	7.0	21.25	Tylenol 750mg - Cx c/ 20 comprimidos	N/I

Problema

- Como determinar que os três itens abaixo fazem menção ao mesmo produto?

	unidade	quantidade	valor_unitario	descricao	ean
328519	UN	1.0	4.69	PARACETAMOL 750 TE.20C	7896112149705
330149	CX	6.0	2.33	PARACETAMOL TEUTO GEN 750MG C/20 CPR	7896112149705
331799	Cx	7.0	21.25	Tylenol 750mg - Cx c/ 20 comprimidos	N/I

Problema

- Como determinar que os três itens abaixo fazem menção ao mesmo produto?

	unidade	quantidade	valor_unitario	descricao	ean
328519	UN	1.0	4.69	PARACETAMOL 750 TE 20C	7896112149705
330149	CX	6.0	2.33	PARACETAMOL TEUTO GEN 750MG C/20 CPR	7896112149705
331799	Cx	7.0	21.25	Tylenol 750mg - Cx c/ 20 comprimidos	N/I

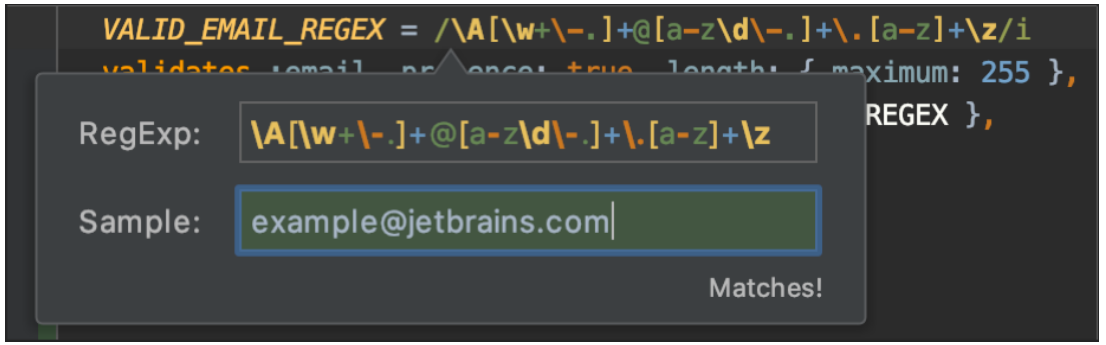
Objetivo

- Este projeto de pesquisa buscou investigar a aplicação de métodos computacionais para associar, de forma automática, uma descrição de item de NFE ao produto comercial a qual se refere.

Fundamentos

Expressões Regulares

- A busca e manipulação em texto com ERs ocorre por meio de padrões.
- Quando um padrão é encontrado no texto, diz-se que ocorreu o *match*.
- Procedimento geral:
  - Entrada: um texto e um padrão
  - Saída: trechos que casam com o padrão e/ou as posições no texto de entrada em que o padrão foi encontrado.



https://www.jetbrains.com/help/ruby/regular-expressions.html

Expressões Regulares

```
^[a-z0-9.]+@[a-z0-9]+\.[a-z]+\.([a-z]+)?$
```

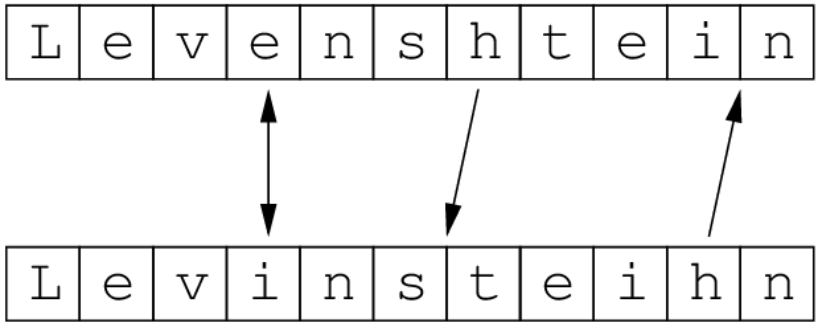
```
augusto.jose@gmail.com          true
augusto.jose@gmail.com.br       true
augusto.jose@gmail.com.br.br    false
augusto.jose@gmail.             false
augusto.jose@gmailcom           false
augusto.josegmail.com           false
@gmail.com                      false
```

Distância Levenshtein

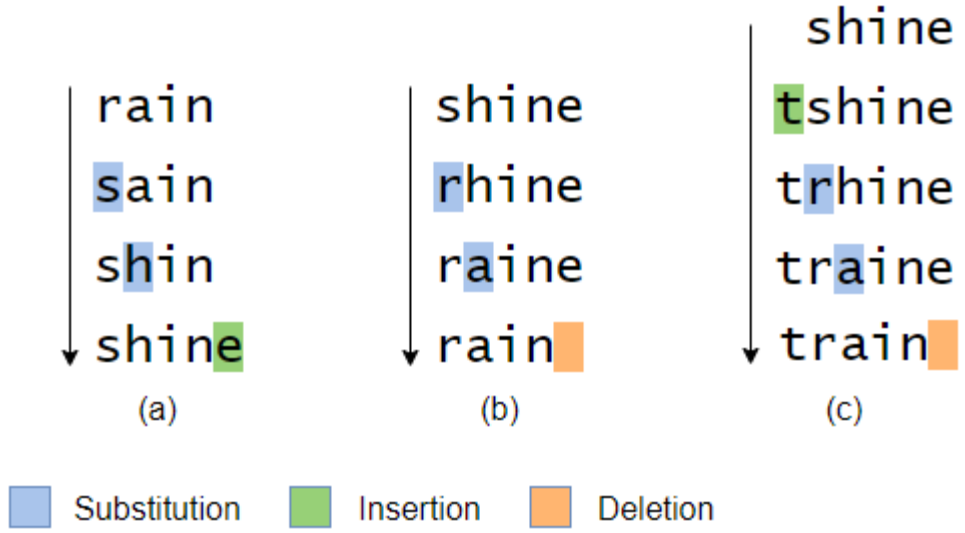
- A Distância Levenshtein, também conhecida como distância de edição, é uma métrica de similaridade calculada sobre dois fragmentos de texto.
- O algoritmo utilizado para o cálculo desta distância foi desenvolvido pelo matemático Vladimir Levenshtein em 1965.
- Sua aplicação possibilita mensurar o quanto dois fragmentos de texto são similares.

A distância Levenshtein entre duas *strings* é o número mínimo de edições de um único caractere (inserções, exclusões ou substituições) necessárias para transformar uma strings na outra.

Distância Levenshtein



https://www.researchgate.net/



https://devopedia.org/levenshtein-distance

Metodologia

Bases de dados

- NFe_PUB
- ANV_MED

Recebemos de (razão social do emitente) os produtos e/ou serviços constantes da Nota Fiscal Eletrônica ao lado.

DATA DE RECEBIMENTO	IDENTIFICAÇÃO E ASSINATURA DO RECEBEDOR	NF-e Nº 000.055.278 SÉRIE: 1

Laboratórios Xavier S/A.
Raua Pedro Castrol, 777
Distrito Industrial
Jaú/SP
13.123-456
19-1234-1234

DANFE
DOCUMENTO AUXILIAR DA NOTA FISCAL ELETRÔNICA
0 - ENTRADA
1 - SAÍDA
1
Nº 000.055.278
SÉRIE 1
Página 1 de 1

CHAVE DE ACESSO
xxx xxxx xxxx xxxx xxxx xxxx xxxx xxxx xxxx xxxx xxxx

PROTOCOLO DE AUTORIZAÇÃO DE USO
XXXXXXXXXXXXXX - XX/XX/XXX

NATUREZA DA OPERAÇÃO
Remessa de amostra grátis

INSCRIÇÃO ESTADUAL	INSCR. ESTADUAL DO SUBST. TRIBUT.	CNPJ
123.123.123.123		12.123.123/0001-90

DESTINATÁRIO / REMETENTE

NOME / RAZÃO SOCIAL			CNPJ / CPF	DATA DA EMISSÃO
Clínica Sampaio Ltda.			12.123.123/0001-12	15/06/20X1
ENDEREÇO	**BAIRRO / DISTRITO**		**CEP**	**DT SAÍDA/ENTRADA**
Avenida Pedro Albuquerque, nº 500	Centro		11.111-111	15/06/20X1
MUNICÍPIO	**FONE / FAX**	**UF**	**INSCRIÇÃO ESTADUAL**	**HORA DA SAÍDA**
Jaú	19-3333-3333	SP	1.111.111.111.111	09:09:05

FATURA / DUPLICADA
OUTROS

CALCULO DO IMPOSTO

B. DE CÁLCULO DO ICMS	VALOR DO ICMS	BASE DE CÁLCULO ICMS ST	VALOR DO ICMS SUBSTITUIÇÃO	VALOR TOTAL DOS PRODUTO
-	-			100,00

VALOR DO FRETE	VALOR DO SEGURO	DESCONTO	OUTRAS DESPESAS ACESSÓRIAS	VALOR TOTAL DO IPI	VALOR TOTAL DA NF
					100,00

TRANSPORTADOR / VOLUMES TRANSPORTADOS

NOME / RAZÃO SOCIAL	FRETE POR CONTA 9 - SEM FRETE	CÓDIGO	PLACA DO VEÍCULO	UF	CNPJ / CPF
ENDEREÇO	**MUNICÍPIO**			**UF**	**INSCRIÇÃO ESTADUAL**

QUANTIDADE	ESPÉCIE	MARCA	NÚMERO	PESO BRUTO	PESO LÍQUIDO

DADOS DOS PRODUTOS / SERVIÇOS

CÓD PROD.	DESCRIÇÃO DOS PROD./SERVIÇOS	NCM/SH	CST	CFOP	UN	QUANT.	VL. UNITÁRIO	VALOR TOTAL	V. CÁLC. ICMS	VALOR ICMS	VALOR IPI	ALÍQUOTAS ICMS	ALÍQUOTAS IPI
	Amostra grátis de remédio para pressão alta		040	5911	UN	1	100,00	100,00	-	-	-	-	-

DADOS ADICIONAIS

INFORMAÇÕES COMPLEMENTARES	RESERVADO AO FISCO
Produtos que seguem para distribuição gratuita, a título de amostra grátis. Isento do ICMS, conf. art. 3º do Anexo I do RICMS/SP. Isento do IPI, conf. inciso III, alínea "c" do artigo 54 do RIPI/2010.	

Base de dados NFe_PUB

- Uma base de dados utilizada, que batizamos de NFe_PUB, é oriunda de registros de aquisições por órgãos públicos brasileiros por meio de notas fiscais eletrônicas.
- Como restrição de escopo e para simplificar o problema, consideramos apenas entradas no conjunto de dados associadas a aquisições de medicamentos.

Base de dados NFe_PUB

	unidade	quantidade	valor_unitario	descricao	ean
328519	UN	1.0	4.69	PARACETAMOL 750 TE.20C	7896112149705
330149	CX	6.0	2.33	PARACETAMOL TEUTO GEN 750MG C/20 CPR	7896112149705
331799	Cx	7.0	21.25	Tylenol 750mg - Cx c/ 20 comprimidos	N/I

Total de 4.861.392 registros, apenas 24% com EAN registrado.

Conjunto de dados ANV_MED

- O outro conjunto de dados usado, batizado de ANV_MED, é fornecido pela ANVISA (24.816 registros, um para cada EAN).
- Possui informação do EAN!

	produto	principio_ativo	apresentacao	ean
28412	PARACETAMOL	paracetamol	750 MG COM CT BL AL PLAS LAR X 20	7896112149705

Conjunto de dados ANV_MED

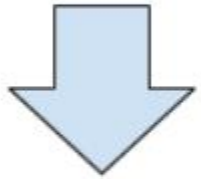
- O outro conjunto de dados usado, batizado de ANV_MED, é fornecido pela ANVISA (24.816 registros, um para cada EAN).
- Possui informação do EAN!

	produto	principio_ativo	apresentacao	ean
28412	PARACETAMOL	paracetamol	750 MG COM CT BL AL PLAS LAR X 20	7896112149705

concentração ("750 MG"),
forma farmacêutica ("COM CT BL AL PLAS LAR"),
separador (“X”)
quantidade (“20”).

Conjunto de dados ANV_MED

	produto	principio_ativo	apresentacao	ean
23858	TILENATI	PARACETAMOL	200 MG/ML SOL OR CT FR GOT PLAS OPC X 10 ML	7897848500341



	descricao	ean
0	TILENATI 200 MG/ML SOL OR CT FR GOT PLAS OPC X 10 ML	7897848500341
1	PARACETAMOL 200 MG/ML SOL OR CT FR GOT PLAS OPC X 10 ML	7897848500341

Abordagens

- Abordagem por similaridade
  - Emprego de expressões regulares e o algoritmo de Levenshtein para inferir o produto comercial baseado em um valor de similaridade entre a descrição deste na base NFe_PUB com a dos itens na base ANV_MED.
- Abordagem por Aprendizado de Máquina
  - Emprego de métodos de aprendizado de máquina para treinar um modelo de classificação capaz de inferir o produto comercial a qual determinada descrição de item de NFe pertence.

Abordagem por similaridade

- Dado um item de NFe sem EAN na base NFe_PUB, calcular a similaridade entre sua descrição e cada descrição contida na base ANV_MED.
  - Intuição: um valor de similaridade alto entre essas descrições seria um indicativo de que correspondem ao mesmo produto.

Abordagem por similaridade - passos

1. Pré-processamento do conjunto de dados NFe_PUB
   - transformação dos caracteres para maiúsculas, remoção de duplicatas, remoção de itens sem EAN registrado, etc.
2. Extração dos termos principais das descrições de itens das NFEs
   - princípio ativo ou nome comercial do medicamento, concentração, forma farmacêutica e quantidade.
3. Normatização de nova descrição
4. Cálculo da similaridade (distância de Levestein)

Abordagem por similaridade - passos

1. Pré-processamento do conjunto de dados NFe_PUB
   - transformação dos caracteres para maiúsculas, remoção de duplicatas, remoção de itens sem EAN registrado, etc.
2. Extração dos termos principais das descrições de itens das NFes
   - princípio ativo ou nome comercial do medicamento, concentração, forma farmacêutica e quantidade.
3. Normatização de nova descrição

Expressões regulares

4. Cálculo da similaridade (distância de Levestein)

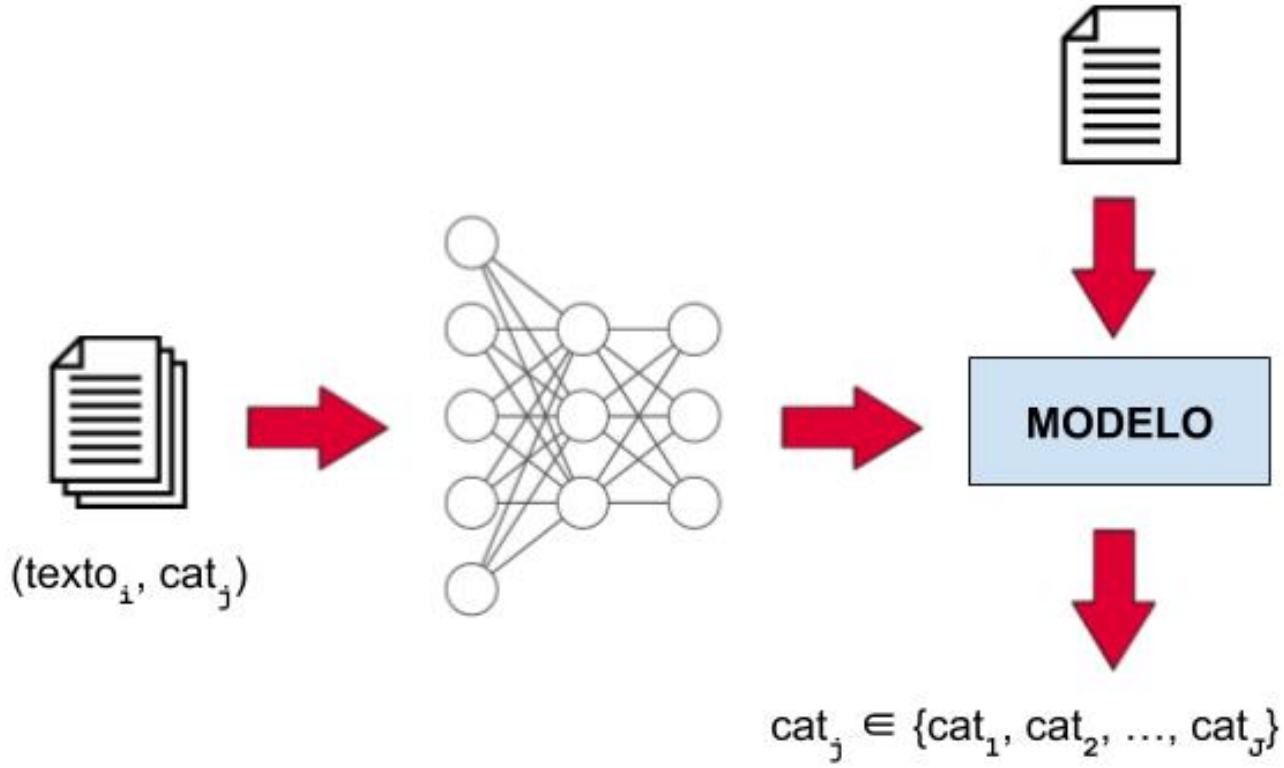
Abordagem por similaridade – passo 3

DESCRIÇÃO	PRINCÍPIO ATV	CONC	FORMA	QTD
CIPROFLOXACINO 200MG CLORI-DRATO S.FECHA	CIPROFLOXACINO	200MG	None	None
CARBOLITIUM CR 450MG C/30 CPR	CARBOLITIUM CR	450MG	COM	30
MUPIROCINA 20MG/G 15G GENERICO PRATI,DONADUZZI	MUPIROCINA	20MG/G	None	15G

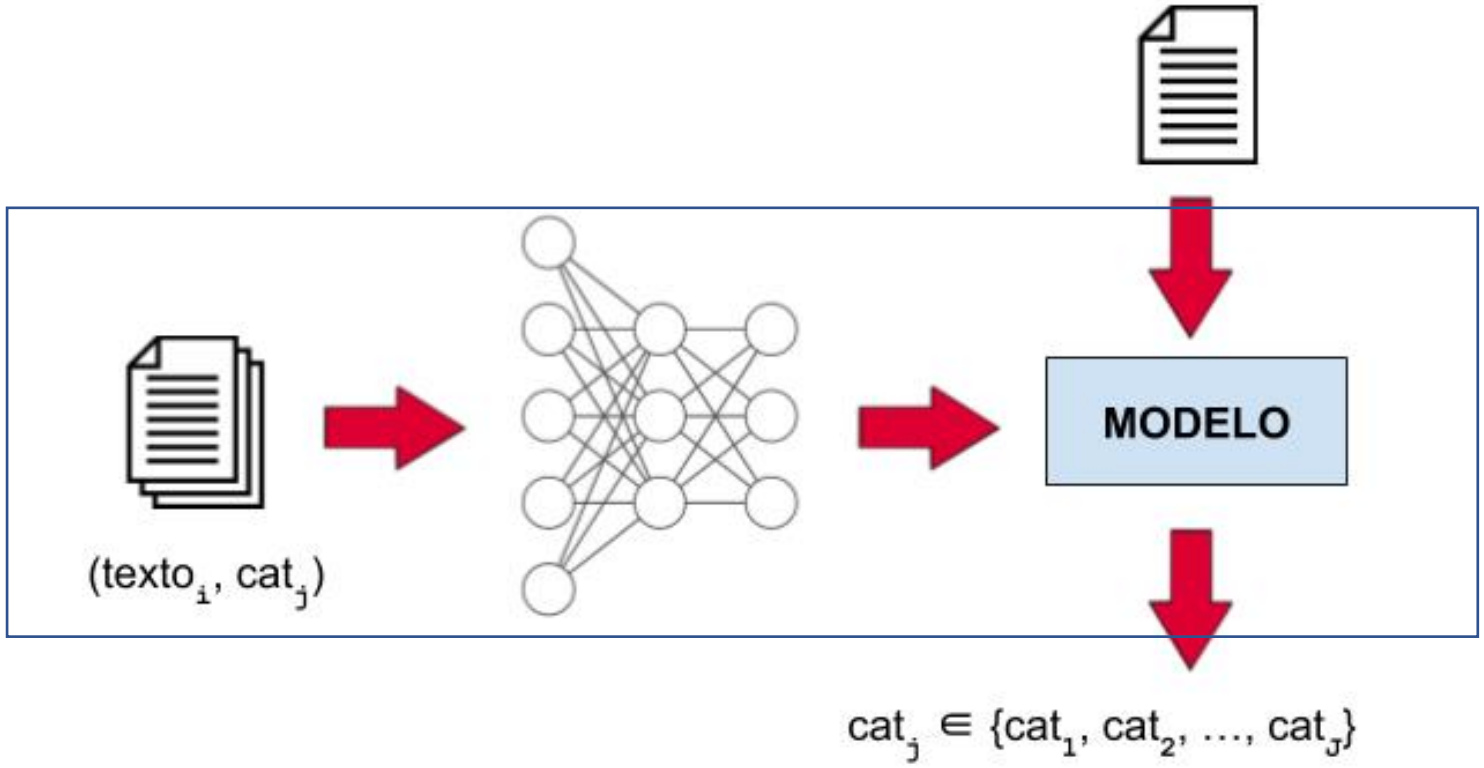
Abordagem por similaridade – passo 4

- Esse passo envolveu calcular o valor de similaridade entre descrições normatizadas e cada uma das descrições na base da ANV_MED.
- O produto associado à descrição ANVISA com maior valor de similaridade foi então considerado como o produto correspondente.
- Para avaliar a taxa de acerto desta abordagem, comparamos os EANs e calculamos o percentual de acerto.

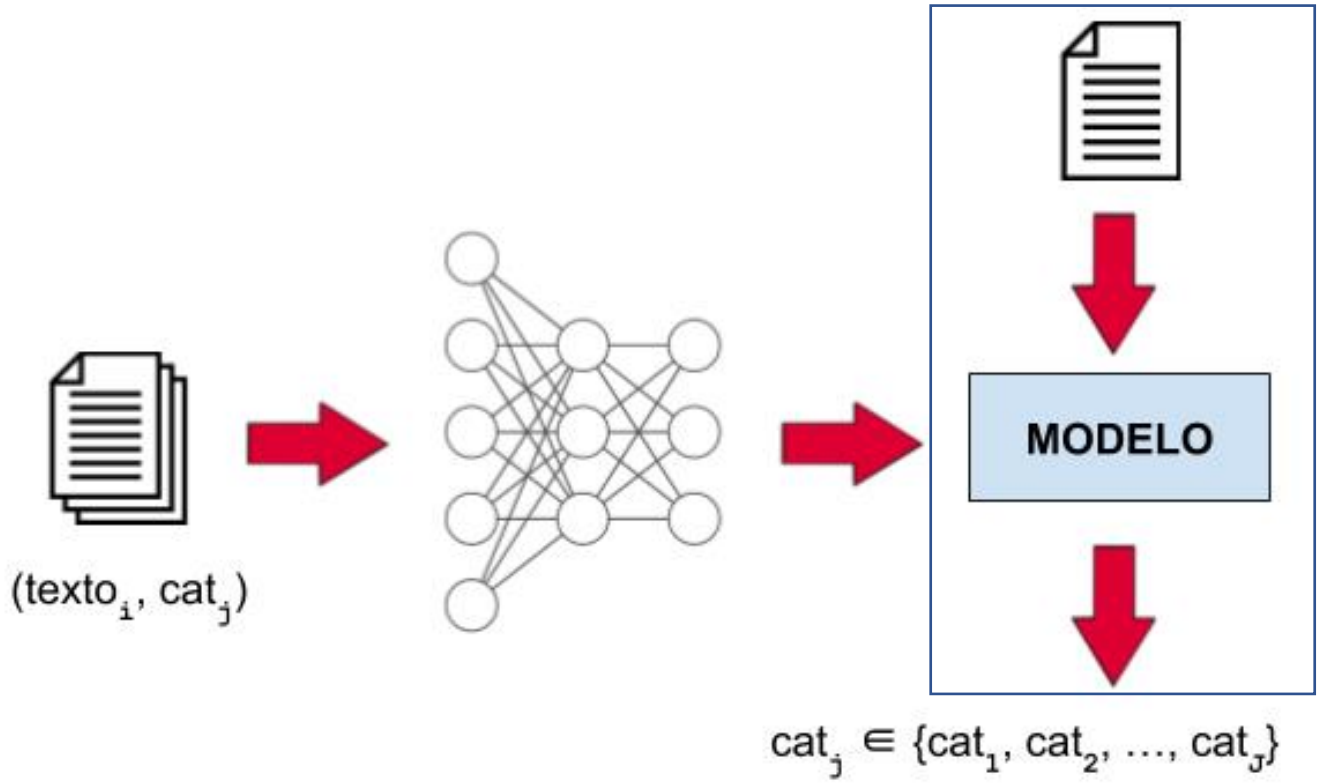
Abordagem por Aprendizado de Máquina



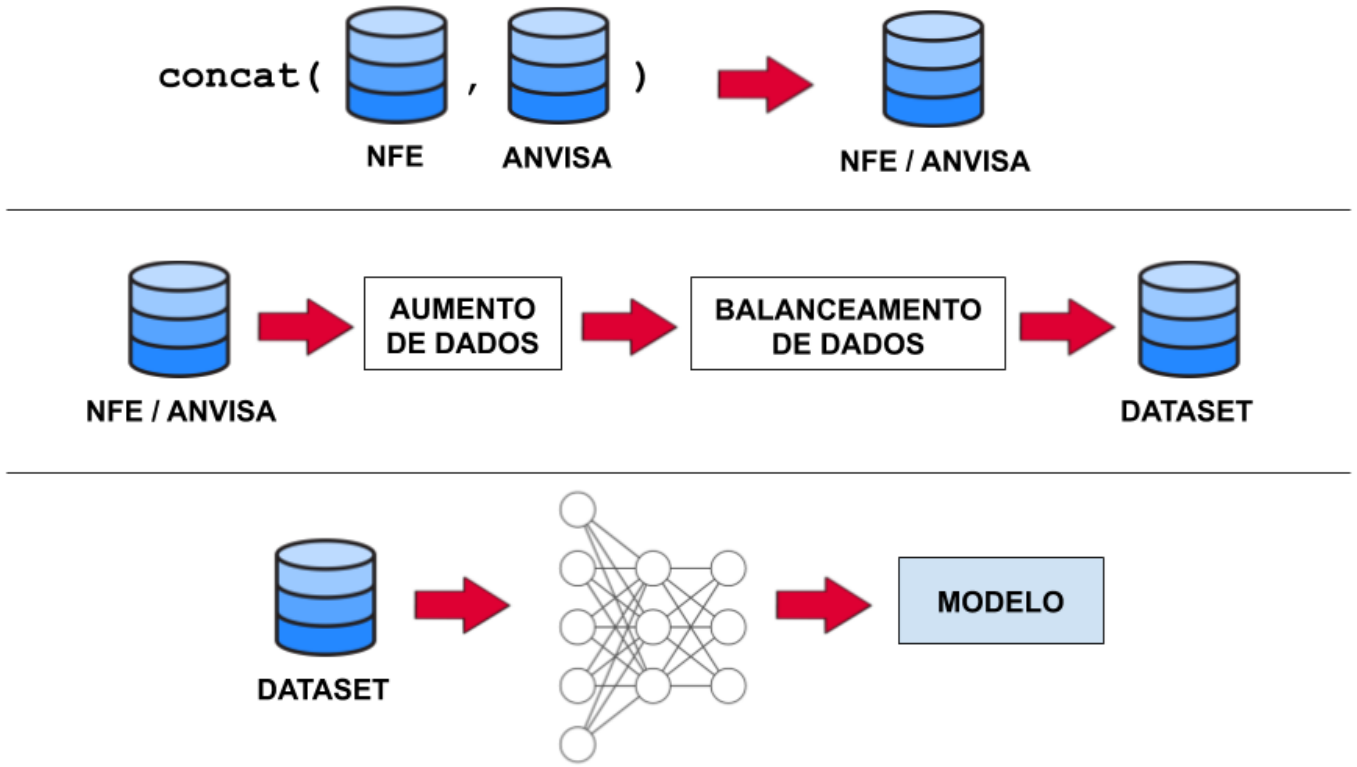
Abordagem por Aprendizado de Máquina



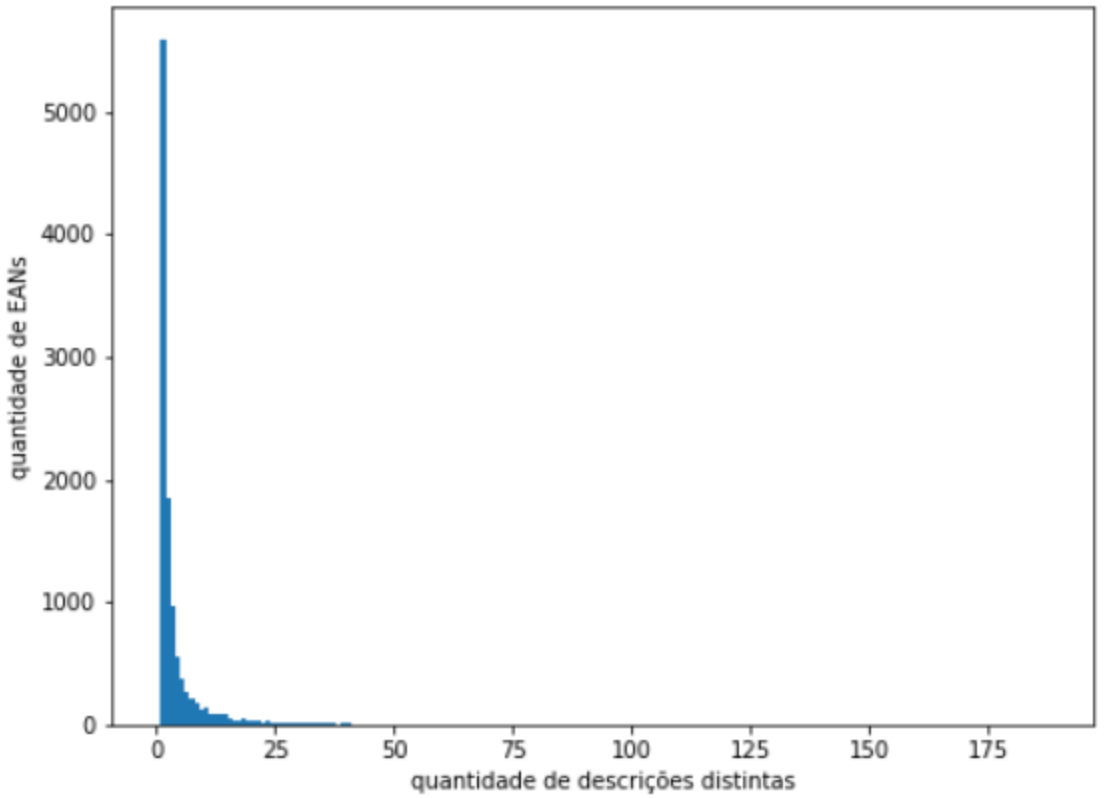
Abordagem por Aprendizado de Máquina



Abordagem por Aprendizado de Máquina

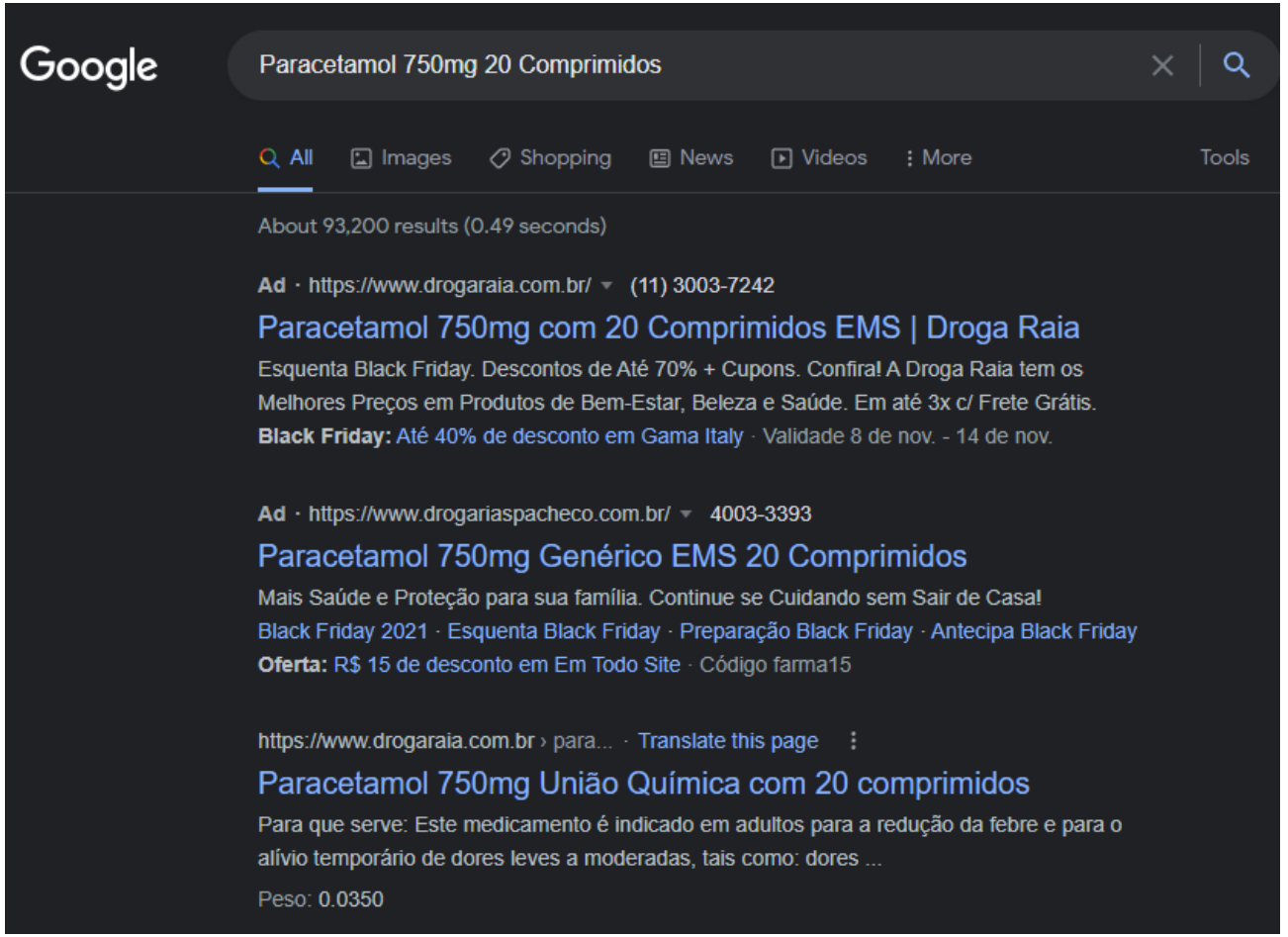


Desbalanceamento da base NFe_PUB



Técnicas para aumento de dados

- Web scrapping (coleta de descrições na Web).
  - Um algoritmo de coleta foi implementado para obter descrições adicionais de um dado medicamento.
  - Foi usado o mecanismo de busca do Google.



Técnicas para aumento de dados

- Derivamos novas descrições por meio de transformações.
  - regras de expressão regular foram empregadas para substituição de termos, inclusão ou remoção de espaços, permutação de palavras, dentre outros.

ORIGINAL	DERIVADO
SELOKEN 1 MG/ML SOL INJ X 5 ML	SELOKEN 1 MG ML SOL INJ X 5ML
SELOKEN 1 MG/ML SOL INJ X 5 ML	SELOKEN 1MG ML SOL INJ X 5 ML
SELOKEN 1 MG/ML SOL INJ X 5 ML	SELOKEN 1G/1000ML SOL INJ X 5 ML

Técnicas para aumento de dados

- Após os procedimentos de Web scraping e de transformações, ainda aplicamos um procedimento de *oversampling*.
- Com isso, conseguimos aumentar em 1000X a quantidade de registros para treinamento do modelo!

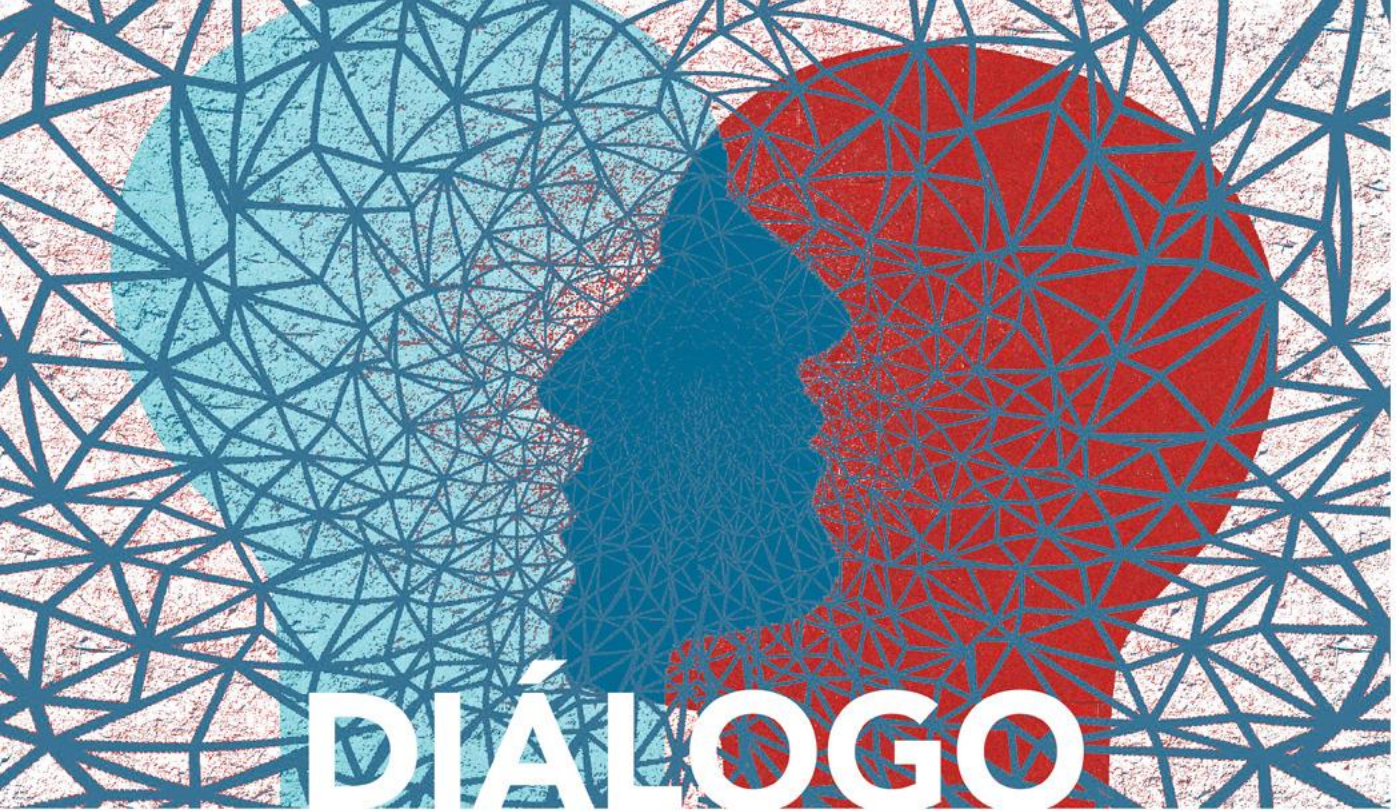
Conclusões

Conclusões

- Classificação por similaridade
  - Apesar de parecer promissora a princípio, essa abordagem resultou em uma acurácia muito baixa.
    - Apenas 16% de acerto.
  - Baixa acurácia tem causa provável na ausência de informações mais detalhadas nas descrições de itens de NFE.
  - Nível de detalhamento das apresentações dos produtos ANV impactam a classificação por similaridade.

Conclusões

- Classificação por AM
  - Alcançou o objetivo do trabalho: resultados obtidos mostraram que o método é eficaz, gerando um modelo de classificação com 0,936 pela métrica F1 Score.
  - Com as devidas adequações, a classificação por AM pode ser aplicável em outros domínios.



RESULTADOS DA 3ª EDIÇÃO DO PROGRAMA DE PESQUISA

Classificação automatizada de produtos da Nota Fiscal Eletrônica de Compras Públicas

Augusto Fonseca (CEFET/RJ)

Bruno Melo (TCE/RJ)

Eduardo Bezerra (CEFET/RJ)

Leonardo Lima (UFPR)

Wellington Amaral (TCE/RJ)

ECG ESCOLA DE CONTAS E GESTÃO DO TCE-RJ

Tribunal de Contas Estado do Rio de Janeiro

Ferramentas utilizadas

- Linguagem de programação Python
- Bibliotecas:
  - python-Levenshtein
  - Imbalanced-learn
  - fastText

Estatísticas dos conjuntos de dados

- SEFAZ_ORIGINAL: 390.341 registros (somente produtos farmacêuticos)
- SEFAZ_PREPROCESSADO: 46.028 descrições distintas com EAN
- ANVISA: 24.816 registros (um para cada EAN)
- SEFAZ_JOIN_ANVISA: 52.226