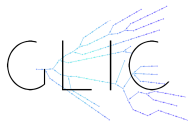


Perspectivas na Análise de Textos Não-Estruturados

Bruno Ferrari Guide



bruno.fguide@gmail.com

22 de Novembro de 2017

Índice

- 1 Introdução - A importância dos dados
- 2 Dados Estruturados vs. Dados não-Estruturados
- 3 Língua e Estrutura
- 4 Alguns métodos
- 5 Alguns problemas
- 6 Estudo de caso 1: B.Os
- 7 Estudo de caso 2: Projeto Cipoal

Introdução

- Textos são conjuntos de dados linguísticos.
- Textos não-estruturados significam dados não estruturados.
- Por que os dados são tão centrais para a Linguística Computacional?
- Por que chamamos esse tipo de dado de não-estruturado?

Importância dos dados para a Linguística Computacional

- Da wikipedia: *Computational linguistics is an interdisciplinary field concerned with the statistical or rule-based modeling of natural language from a computational perspective, as well as the study of appropriate computational approaches to linguistic questions.*

Modelos Baseados em Regras

- Modelos baseados em regras são construídos a partir de um conjunto de instruções que descrevem um determinado fenômeno.
- Exemplo: Plural do Português → 's' no fim da palavra caso termine em vogal, 'es' caso termine em consoante.
- Caso não funcione, se cria uma exceção que deve ser tratada com regras diferentes.

Modelos Probabilísticos

- Modelo probabilístico é baseado na observação de uma amostra de dados que representem um determinado processo. A partir dessa amostra o modelo tentará capturar o comportamento do processo. O modo como isso será feito define o tipo de modelo.
- **Exemplo** - Observação: A partir do estudo de 100 palavras no singular e sua forma plural, temos 90% que fazem o plural com a adição de 's' no fim da palavra, 5% com a adição de 'es' e 3% como em 'pão' - 'pães' e 2% como em 'lápis' - 'lápis'.
- Modelo: lista o conjunto de sílabas finais da amostra e associa a cada um deles a probabilidade dele seguir um dos 4 padrões observados.

Modelos Baseados em Regras vs. Probabilísticos

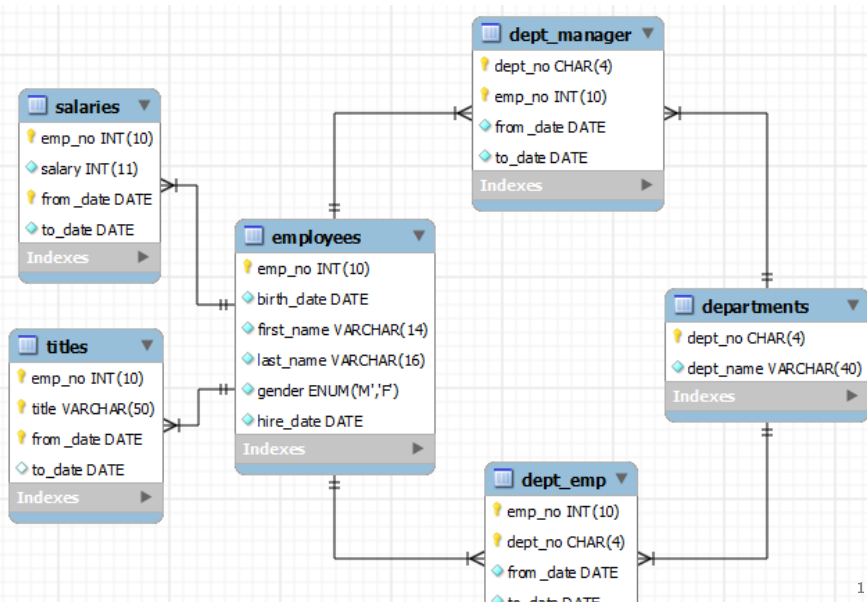
- Modelos baseados em regras ainda existem, mas no mundo da NLP, desde os anos 80, os principais modelos computacionais para linguagem natural são probabilísticos.
- Além disso, a soma de dois fatores, explosão de dados disponíveis e grande poder de processamento acessível, mostram que é muito improvável que esse cenário se reverta nos próximos anos.

Dados Estruturados e Dados não-Estruturados

Dados Estruturados e Não-Estruturados

- Dados estruturados são aqueles que tem um formato especificado, em que as informações relevantes estão dispostas de modo organizado. É possível fazer inferências sobre eles.
- Os dados não-estruturados dependem de pré-processamento para que as informações relevantes sejam extraídas.

Exemplo de dados Estruturados

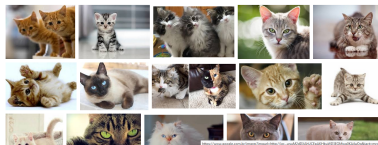


Exemplo de dados não-estruturados

guyana.gov

Felizmente, como descrevemos acima, agora é possível entender e descrever sistemas econômicos como sendo sistemas complexos a exemplo de jardins. E agora é racional afirmar que sistemas econômicos não são meramente similares a ecossistemas; eles são ecossistemas, dirigidos pelos mesmos tipos de forças evolucionárias de ecossistemas. A obra *The Origin of Wealth* (A Origem da Riqueza) de Erick Beinhocker traz a pesquisa mais lúcida disponível dessa nova economia de complexidade.

A estória que Beinhocker conta é simples, e não diferente da estória que Darwin conta. Em uma economia, assim como em qualquer ecossistema, a inovação é resultado de pressões competitivas e evolucionárias. Dentro de qualquer dado ambiente competitivo – ou o que é chamado de “paisagem saudável” – os indivíduos e grupos cooperam para competir, para achar soluções para problemas e estratégias de cooperação se espalham e multiplicam-se. Por toda parte, pequenas vantagens iniciais são amplificadas e



Dados Estruturados vs. Dados não-Estruturados

Dados estruturados	Dados não-estruturados
Menor quantidade disponível Informações organizadas e prontas para serem processadas	Imensa disponibilidade Exige pré-processamento para extração de informações relevantes

Processamento de Linguagem Natural

- Para construir qualquer aplicação usando dados linguísticos, é necessário estruturar esses dados.
- Há esforços para conseguir estruturar os mais diversos aspectos da língua para criar modelos que consigam fazer inferências sobre dados linguísticos.

Língua e estrutura

A língua não tem estrutura?

- Importante: Não há relação entre a estrutura interna do fenômeno que gera um dado e o fato deste dado estar estruturado ou não.
- No entanto, a estrutura interna do fenômeno pode ser bastante útil para organizar e padronizar os dados observados.

Como essa estrutura é útil para o processamento?

- Exemplo: previsão da próxima palavra dada a palavra anterior.
- o pai do ???
- É muito difícil fazer a previsão da próxima palavra, no entanto é bem possível que a gente faça previsões sobre algumas características da palavra que falta, como sua categoria morfossintática, gênero, número.

Porém...

- A estrutura interna da língua é um tipo de diferente do desejado para os chamados dados estruturados. É muito mais complexa, o que torna a relação entre a estrutura interna e os dados estruturados algo não trivial.

Porém...

- A estrutura interna da língua é um tipo de diferente do desejado para os chamados dados estruturados. É muito mais complexa, o que torna a relação entre a estrutura interna e os dados estruturados algo não trivial.
- Ambiguidades estruturais são comuns na língua e um pesadelo na estruturação de dados.

Porém...

- A estrutura interna da língua é um tipo de diferente do desejado para os chamados dados estruturados. É muito mais complexa, o que torna a relação entre a estrutura interna e os dados estruturados algo não trivial.
- Ambiguidades estruturais são comuns na língua e um pesadelo na estruturação de dados.
- Devido a essa grande complexidade da estrutura linguística, é bastante recorrente no campo da Lx. Computacional modelar a língua desconsiderando a estrutura problemática.

Porém...

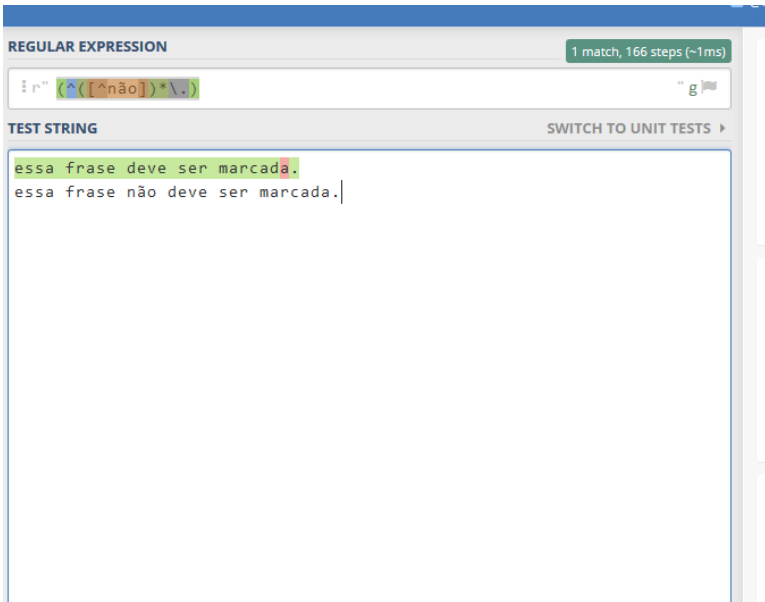
- A estrutura interna da língua é um tipo de diferente do desejado para os chamados dados estruturados. É muito mais complexa, o que torna a relação entre a estrutura interna e os dados estruturados algo não trivial.
- Ambiguidades estruturais são comuns na língua e um pesadelo na estruturação de dados.
- Devido a essa grande complexidade da estrutura linguística, é bastante recorrente no campo da Lx. Computacional modelar a língua desconsiderando a estrutura problemática.
- O simples reconhecimento de padrões *superficiais* nos dados linguísticos já é algum tipo de análise linguística.

Alguns métodos para estruturar dados linguísticos

Expressões Regulares - Regex

- **O que é:** Método para reconhecer sequências de caracteres. Permite que sejam encontrados tipos de caracteres específicos, repetições.
- **Utilidade:** Extremamente útil para extrair palavras-chave, capturar pequenas variações de expressões. Ferramenta poderosa.
- Quanto da estrutura linguística é representada nas Regex?

Expressões Regulares - Exemplo



The screenshot shows a web-based regular expression testing tool. At the top, there's a blue header. Below it, the 'REGULAR EXPRESSION' section contains the regex `r"^(^[^nãõ]*)*\."` in a text input field. To the right of the input, a green badge indicates '1 match, 166 steps (~1ms)'. Below the input field is a 'TEST STRING' section with a 'SWITCH TO UNIT TESTS' button. The test string area contains two lines of text: 'essa frase deve ser marcada.' and 'essa frase não deve ser marcada.'. The first line is highlighted in green, and the period at the end is highlighted in red, indicating a successful match. At the bottom right, there are navigation icons and a page number '23 / 53'.

REGULAR EXPRESSION 1 match, 166 steps (~1ms)

`r"^(^[^nãõ]*)*\."` " g

TEST STRING SWITCH TO UNIT TESTS ▶

essa frase deve ser marcada.
essa frase não deve ser marcada.

23 / 53

Expressões Regulares - Exemplo

REGULAR EXPRESSION1 match, 191 steps (~0ms)

`(^)+([^\nãõ])*\.`g

TEST STRINGSWITCH TO UNIT TESTS (2) ▶

fp fjwpfjprfl mef çwmfçwfm wfpmpmfwf. |
essa frase não deve ser marcada.

Lematização

- **O que é:** Traduzir as formas morfológicas diversas de uma palavra para sua versão dicionarizada (lema).
- **Utilidade:** Reduz a riqueza de formas da língua enquanto mantém algumas informações, preserva alguma parte do significado.

Lematização - Exemplo

MODO INDICATIVO						
	PRESENTE	PERFEITO	IMPERFEITO	MAIS-QUE-PERFEITO	FUTURO DO PRESENTE	FUTURO DO PRETERITO
EU	amo	amei	amava	amara	amarei	amaria
TU	amas	amaste	amavas	amaras	amarás	amarias
ELE	ama	amou	amava	amara	amará	amaria
NÓS	amamos	amamos	amávamos	amáramos	amaremos	amaríamos
VÓS	amais	amastes	amáveis	amáreis	amareis	amaríeis
ELES	amam	amaram	amavam	amaram	amarão	amariam

MODO SUBJUNTIVO			
	PRESENTE	IMPERFEITO	FUTURO
EU	ame	amasse	amar
TU	ames	amasses	amares
ELE	ame	amasse	amar
NÓS	amemos	amássemos	amarmos
VÓS	ameis	amásseis	amardes
ELES			

FORMAS NOMINAIS			
INFINITIVO		GERÚNDIO	PARTICÍPIO
<i>para eu</i>	amar	amando	amado
<i>para tu</i>	amares		
<i>para ele</i>	amar		
<i>para nós</i>	amarmos		
<i>para vós</i>	amardes		

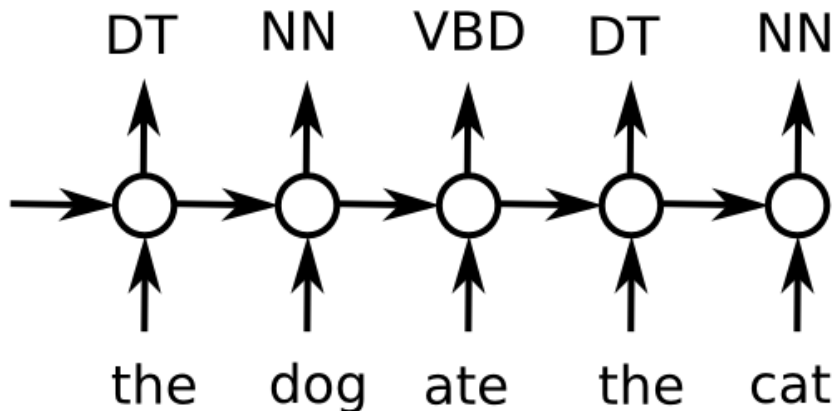
Lematização - Exemplo

amar

POS-Tagging

- **O que é:** Etiquetar a categoria morfossintática de palavras.
- **Utilidade:** Reduz MUITO a diversidade dos dados, mantém algum tipo de informação linguística que pode ser útil para algumas generalizações. Perde muita informação também.

POS-Tagging - Exemplo



Alguns problemas para estruturar dados linguísticos

"All grammars leak"

- Exceções, coisas estranhas, comportamentos opostos dentro da mesma língua.

"All grammars leak"

- Exceções, coisas estranhas, comportamentos opostos dentro da mesma língua.
- Além disso, diversos dos algoritmos que fazem as tarefas de estruturação que eu apresentei são probabilísticos e por isso o erro é algo que faz parte dessas abordagens.

"All grammars leak"

- Exceções, coisas estranhas, comportamentos opostos dentro da mesma língua.
- Além disso, diversos dos algoritmos que fazem as tarefas de estruturação que eu apresentei são probabilísticos e por isso o erro é algo que faz parte dessas abordagens.
- Portanto, essa estruturação de dados é sempre complexa e falha. Existem alternativas para lidar com isso (revisão e/ou diluição de erros).

Estado da arte em algumas áreas

- Regex - não há.
- Lematização - Inglês (93%), Português (97%).
- POS Tagging - Inglês (85.85% para itens não vistos e 96.46%), Português (85%?).

O que significam estes índices de acerto?

- O número sem a metodologia da avaliação acaba sendo meio vazio. Por exemplo, as palavras foram consideradas isoladas ou em contexto? Em inglês isso é fundamental para considerar a tarefa de pos-tagging, por exemplo.
- Se 90% das palavras estão classificadas corretamente, quer dizer que em uma sentença de 10 palavras uma delas estará errada, logo a sentença não irá ser representada do jeito correto (100% errada!).

Estudos de Caso ¹

¹Atenção: Eles são deprimentes.

Contexto do projeto

- Lei de Transparência obriga uma certa organização de segurança pública a divulgar determinados tipos de Boletins de Ocorrência.
- A divulgação é feita de forma não estruturada, apesar dos dados internamente estarem estruturados.
- O objetivo do projeto foi justamente estruturar os dados linguísticos contidos nos B.Os.

Contexto do projeto



Os dados

SECRETARIA DE ESTADO DA SEGURANÇA
PÚBLICA
POLÍCIA CIVIL DO ESTADO DE SÃO PAULO

Dependência: SETOR HOM.SEC. CARAPICUIBA

Folha: 1

Boletim No.: 3/2013

Iniciado: 30/05/2013 06:55hs e Emitido: 30/05/2013
06:55hs

Boletim de Ocorrência de Autoria Conhecida

Natureza(s):

Espécie: Título I - Pessoa (arts. 121 a 154)

Natureza: Homicídio simples (art. 121)

Consumado

Desdobramentos: Morte decorrente de intervenção policial (RES. SSP 05 - 07/01/2013)

Espécie: Título II - Patrimônio (arts. 155 a 183)

Natureza: Roubo (art. 157)

Consumado

Espécie: Título XI - Administração pública (arts. 312 a 359-H)

◀ ◻ ▶ ◀ ◻ ▶ ◀ ≡ ▶ ◀ ≡ ▶ ≡

Método

```
def info_vitimas(vetor_bo):
    '''recebe como input um vetor e devolve um numero n de copias
    desse vetor de acordo com o número de vitimas envolvidas. Alem
    disso, extrai e organiza as infos das vítimas.'''
    vitimas = vetor_bo[-1]
    # v_resto = vetor_bo[:-1]
    #sequencia definindo as REGEX pra processar as coisas
    rg_patr = '(?<=RG: )([0-9]*)'
    nomes_patr = '([A-Z ]+)( [\\(\\)])'
    natural_patr = '(Natural de: )([A-Z.]+)( -[A-Z]+)'
    nacionalidade = '(?<=Nacionalidade: )([A-Z.]+)'
    sex_patr = '(Masculino|Feminino)'
    nascimento_patr = '([0-9]{2}\\/[0-9]{2}\\/[0-9]{4})'
    idade_patr = '([0-9]+) (?=anos)'
    estado_civil_patr = '(?<=Estado Civil: )(Casado|Solteiro)'
    profissao_patr = '(?<=Profissão: )([A-Z]+)'
    instrucao_patr = '(?<=Instrução: )([^-]*)'
    cutis_patr = '(?<=Cutis: )([PBA][^A-Z]*)'
    natur_patr = '(?<=Naturezas Envolvidas: )(\\.+)(\\)'
```

Resultados

S	T	U	V	W	X	Y
NOME_VITIMA	RG_VITIMA	NATURALIDADE_VITIMA	NATURALUF_VITIMA	NACIONALIDADE_VITIMA	SEXO_VITIMA	DATA_NASCIMENTO
	44032760	S,PAULO	-SP	BRASILEIRA	Masculino	12/11/1984
	47514757	BACABAL	-MA	BRASILEIRA	Masculino	18/05/1989
	40179156	ARACATUBA	-SP	BRASILEIRA	Masculino	16/08/1985
	34869473	BUERAREMA	-BA	BRASILEIRA	Masculino	27/09/1979
	52973786	S,LUZ	-BA	BRASILEIRA	Masculino	22/01/1985
	33190889	nan	nan	nan	Feminino	nan
	14723787	INDIAPORA	-SP	BRASILEIRA	Masculino	13/11/1962
	17517257	BARRETOS	-SP	BRASILEIRA	Masculino	02/09/1969
	41015409	OSASCO	-SP	BRASILEIRA	Masculino	29/01/1994
	43750598	S,PAULO	-SP	BRASILEIRA	Masculino	13/05/1984
	56154279	S,PAULO	-SP	BRASILEIRA	Masculino	04/07/1997
	4377505	nan	nan	BRASILEIRA	Masculino	20/03/1949
	26196500	S,PAULO	-SP	BRASILEIRA	Masculino	14/10/1975
	33978962	DIADAMA	-SP	BRASILEIRA	Masculino	03/06/1982
	49258433	S,PAULO	-SP	BRASILEIRA	Masculino	09/01/1993
	48090399	nan	nan	BRASILEIRA	Masculino	19/02/1992
	24335562	S,PAULO	-SP	BRASILEIRA	Masculino	25/11/1981
	28563986	S,PAULO	-SP	BRASILEIRA	Masculino	23/05/1973
	43188415	GUARULHOS	-SP	BRASILEIRA	Masculino	08/10/1988
	44131675	S,PAULO	-SP	BRASILEIRA	Masculino	04/12/1991
	35176737	S,PAULO	-SP	BRASILEIRA	Masculino	24/06/1985
	16980072	NEOPOLIS	-SE	nan	Feminino	04/01/1964
	21718960	nan	nan	nan	Masculino	14/11/1972
	19722712	TAUBATE	-SP	BRASILEIRA	Masculino	21/02/1971
	42019396	S,PAULO	-SP	BRASILEIRA	Masculino	09/03/1986
	21775104	nan	nan	nan	Masculino	22/06/1980
	24407326	PAULISTA	-SP	BRASILEIRA	Masculino	07/07/1971

TABELA2013

Planilha 1 / 1

Padrão

Soma=0

10:50

Contexto do projeto Cipoal

- **Corpus:** Leis aprovadas pela câmara dos vereadores do município de São Paulo.
- **Problema:** Conjunto de dados grande, armazenado de um jeito caótico, mídias diferentes.
- Para propor *qualquer* novo projeto de lei, é necessário ver se o projeto entra em conflito com leis anteriores, ver se ele já não foi proposto, etc...

Os dados

PORTARIA SECRETARIA MUNICIPAL DOS TRANSPORTES/DSV Nº 114 DE 30 DE DEZEMBRO DE 2011

[Voltar](#) | [Imprimir](#)[DETALHES DA NORMA](#)[TEXTO CONSOLIDADO](#)

CREDENCIA 34 POLICIAIS MILITARES COMO AGENTES DE TRANSITO PARA FISCALIZACAO/AUTUACAO DE VEICULOS, NOS TERMOS DO CONVENIO DE 24/05/2006.

PORTARIA 114/11 DSV/SMT

de 29 de Dezembro de 2011

O DIRETOR DO DEPARTAMENTO DE OPERAÇÃO DO SISTEMA VIÁRIO DSV , órgão integrante do Sistema Nacional de Trânsito, nos termos do artigo 7.º, inciso III, da Lei n.º 9.503, de 23 de setembro de 1997, que institui o Código de Trânsito Brasileiro CTB, e do Decreto Municipal n.º 37.293, de 27 de janeiro de 1998, que estabelece a competência do DSV na área de circunscrição do Município, usando de suas atribuições legais, e

CONSIDERANDO o disposto nos artigos 280 e 269 do Código de Trânsito Brasileiro CTB, que dispõem sobre a atuação de infração de trânsito e adoção de medidas administrativas por agente de autoridade de trânsito, que poderá ser servidor civil ou policial militar;

CONSIDERANDO que agente da autoridade de trânsito é a pessoa credenciada pela autoridade de trânsito para o exercício das atividades de fiscalização;

Método

- Neste primeiro momento: Estruturar o corpus e classificar o projeto de lei com Expressões Regulares.

Método

- Neste primeiro momento: Estruturar o corpus e classificar o projeto de lei com Expressões Regulares.
- Para estruturar: identificar os padrões linguísticos que determinam o que é o cabeçalho, as normas de citação de outras leis e projetos, valores.

Método

- Neste primeiro momento: Estruturar o corpus e classificar o projeto de lei com Expressões Regulares.
- Para estruturar: identificar os padrões linguísticos que determinam o que é o cabeçalho, as normas de citação de outras leis e projetos, valores.
- Para classificar: primeiro, definir as classes relevantes, depois relacionar com palavras-chave e expressões multi-palavra que identifiquem as coisas a serem classificadas.

Resultados Parciais

The logo for Cipoal, featuring the word "Cipoal" in a bold, green, sans-serif font. The letters are filled with a detailed texture of green grass blades. A thick, dark green underline curves beneath the text.Comece aqui sua busca

Resultados Parciais

Dashboard

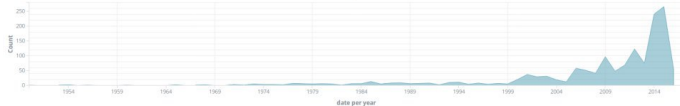
Total

1,448

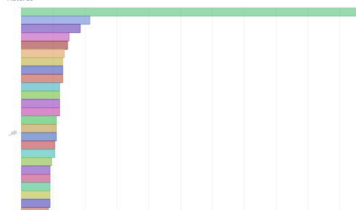
Lets Catalogadas

Linha do Tempo

Count



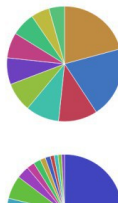
Autores



Partidos

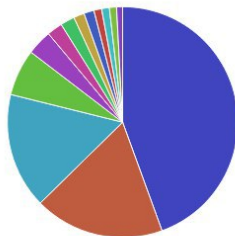


Categorias



Resultados Parciais

Categorias



- ▶ Calendário
- Outra
- Alteração de nome d...
- Habitação
- Saúde
- Orçamento
- Saúde Calendário
- Transporte
- Idosos
- Fiscal
- Habitação Calendário
- Calendário Orçamento
- Fiscal Habitação

Resultados Parciais





Let's

number.keyword: Descending ▾	brief.keyword: Descending ▾	Count
0	Lei Orgânica do Município de São Paulo	1
10007	Aprova plano de melhoramentos no 329 subdistrito - Capela do Socorro, e dá outras providências.	1
10032	Dispõe sobre a criação de um conselho municipal de preservação do patrimônio histórico, cultural e ambiental da cidade de São Paulo.	1
10072	Dispõe sobre a instalação de bancas de jornais e revistas em logradouros públicos, e dá outras providências &	1
10094	Introduz alterações no quadro nº 7b, anexo à lei nº 8001, de 24 de dezembro de 1973.	1

Contato

Muito obrigado!
bruno.fguide@gmail.com