Similaridade de textos normativos: um ensaio sobre as leis orçamentárias[1]

Alexandre Manir Figueiredo Sarquis

Conselheiro-Substituto do TCE-SP. Aluno do Doutorado do programa de Pós-Graduação em Direito Financeiro da USP, sob orientação do Professor Titular Heleno Taveira Torres. Mestre em Economia pela UnB. É Professor da Escola Brasileira de Administração Pública (Ebap).

Resumo: Este ensaio parte da premissa de que é útil para a ciência do direito aproximar essa técnica de cognição. Dessa forma, o interesse é da ciência jurídica em relação à ciência computacional: a primeira precisa se conformar aos resultados disponíveis na segunda. A abordagem é deliberadamente matemática, aos auspícios de sistemas de computador, aproximando-se do que tem se chamado de "inteligência artificial". Neste estudo foi analisado um acervo conhecido de documentos a fim de averiguar se a análise que emerge dessas técnicas é capaz de identificar fatos estilizados esperados e se as surpresas encontradas se correlacionavam a elementos conhecidos do plano jurídico.

Palavras-chave: Diretrizes orçamentarias. Similaridade. Direito. Nuvem de palavras.

Introdução

Há uma sedimentada doutrina científica em torno do cálculo numérico da similaridade de documentos. Tal apuração é de importância para as buscas pela internet, demanda que precipitou a mencionada teorização. Dessa forma, uma fonte textual "buscada" é comparada computacionalmente com uma relação de textos "buscados", apontando-se, ao fim e ao cabo, um índice de similaridade ou de proximidade que é empregado para ordenar os resultados. *Mutatis mutandis*, a mesma técnica pode ser empregada a quaisquer textos – entendidos como sequências ordenadas de vocábulos – inclusive para textos legais.

Tal possibilidade, no entanto, ainda não tem sido explorada. Pretendemos investigar, neste breve ensaio, o cálculo de similaridade textual conhecido como "distância de cosseno", definido, com apoio nos conceitos da dissertação de mestrado *Agrupamento e categorização de documentos jurídicos* (FURQUIM, 2011).

Similaridade de textos

A similaridade de textos, como advento da moderna ciência da computação, coloca uma oportunidade e um desafio ao tradicional silogismo da lei que tem prevalecido na propedêutica e na prática jurídica. A busca de uma igualdade exata que tem prevalecido na análise dos textos legais, em que mesmo vocábulos considerados pelo vernáculo como sinônimos, uma vez alternados em um dispositivo convencional – seja contratual, seja legal –, dão ensejo a analises que culminam em diferentes implicações ao direito.

É um panorama em que o jargão não importa tanto pela sua carga semântica, mas pela multiplicidade de vênias interpretativas que pode ensejar, assim como de seus desdobramentos.

[1] Este trabalho foi desenvolvido no âmbito da pesquisa desenvolvida pelo Professor Marcos Augusto Perez.

A modernidade tem se pautado, entretanto, por uma cognição expedita e paramétrica dos conteúdos disponíveis, em que o núcleo semântico dos textos recobra a culminância no valor da mensagem transmitida.

O presente estudo parte da premissa de que é útil para a ciência do direito aproximar essa técnica de cognição. Dessa forma, o interesse é da ciência jurídica em relação à ciência computacional: a primeira precisa se conformar aos resultados disponíveis na segunda. A abordagem será deliberadamente matemática, aos auspícios de sistemas de computador, aproximando-se do que tem se chamado de "inteligência artificial".

Ao informar expressões a um sistema de busca informatizado, esse, via de regra, descartará palavras curtas e usuais e, de outra mão, reforçará a ênfase de palavras menos frequentemente buscadas. Possivelmente a busca pela expressão "definição de Savonarolas" recupere apenas textos que contenham o maior número de referências ao monograma "savonarola", descartando o plural assim como os demais vocábulos do comando de busca, por presumir sua irrelevância.

Outra diferença reside no fato de que a similaridade passa a constituir uma escala em que o "exatamente igual" e o "exatamente diverso" dão lugar a um contínuo de proximidades. Ordenados segundo essa medida, os textos ficam apresentados como mais similares ou menos similares, devendo assim ser ponderadas as consequências de tal similaridade.

Tal método apresenta vantagens progressivas com a extensão do texto legal a analisar, uma vez que, na pequena escala, pode produzir resultados inconvenientes e incorretos. Como é exemplo, o texto "Não é permitido o homicídio" pode ser classificado como intensamente similar ao texto "É permitido o homicídio", enquanto a mais despretensiosa análise humana de pronto revela tratar-se de conteúdos semânticos absolutamente opostos.

Constitui, portanto, uma estatística que deve ser aproveitada na medida de sua utilidade, sem a pretensão de cumprir o rigor matemático do silogismo da lei. A igualdade entre o caso de dois documentos ou dois dispositivos legais ou ainda a subsunção de um fato determinado a uma específica norma melhor ficam legadas às tradicionais técnicas de interpretação e integração da norma.

Saco de palavras

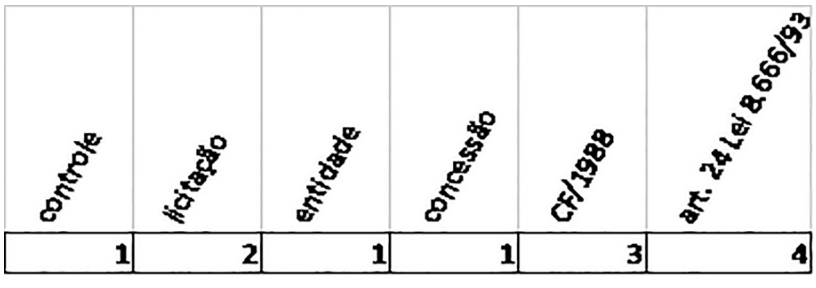
Está no centro da técnica desenvolvida, portanto, uma série de simplificações textuais que são aplicadas tanto ao original que será comparado quanto ao conjunto de eventuais fontes paradigmas para comparação. Tais técnicas buscam unicamente a comparabilidade sob os paradigmas de que a ordem dos termos é irrelevante, bem como que o peso das palavras deve ser reconhecido em suas relativas raridades. Por esse motivo:

a) números são descartados;
b) há um dicionário de palavras que são ignoradas (*stop words*), a exemplo de artigos, preposições e pronomes;
c) declinações de tempos verbais podem ser reunidas em um mesmo monograma ("lematização");
d) palavras de conteúdo semântico próximo podem ser reunidas em um mesmo anagrama, por exemplo, "grande" e "grandemente" (*stemming*);
e) as frequências de ocorrência de uma palavra dentro de um documento podem ser corrigidas pela raridade da ocorrência da palavra nos demais documentos do grupo de comparação (*if-idf*);
f) a ordem em que os monogramas ocorrem é ignorada e cada monograma é simplesmente contado.

As palavras consideradas relevantes – doravante intituladas “monogramas” ou “lematas” – são reunidas para contagem, destruindo a estrutura gramatical original. Essa técnica é conhecida como *bag of words* ou “saco de palavras”, pois remete à desorganização de algo que é reunido em um único recipiente. As classes gramaticais assim como as classes sintáticas são descartadas, de forma que a mensagem original não pode ser recuperada a partir da contagem de monogramas.

O objetivo dessa fase inicial é decompor cada documento em um vetor de monogramas que podem ser totalizados de acordo com o número de ocorrências em cada documento, assim constituindo uma assinatura numérica desse documento, passível de cálculos que podem revelar a sua similaridade.

Figura 1 – Resultado do processo inicial de contagem de monogramas



Fonte: Elaboração própria.

Após tal simplificação, que, como mencionado, é alcançada após o descarte de monogramas considerados irrelevantes, da equiparação de monogramas que significam apenas a flexão uns dos outros (“lematização”) e da equiparação de monogramas que importam carga semântica significativamente parecida (*stemming*), faz-se necessária uma última correção.

Trata-se do inverso da frequência. Independentemente da quantidade de vezes que um monograma surge em um documento, tal aparição ganha importância se ele é raro, quando se considera o universo de documentos que é pesquisado. A quantidades relativas de ocorrência do monograma no documento , , deve ser corrigido por um peso de importância daquele específico monograma:

$$N_{i,j} = n_{i,j}.w_i = n_{i,j}.\log\left(\frac{ND}{ND_i}\right)$$

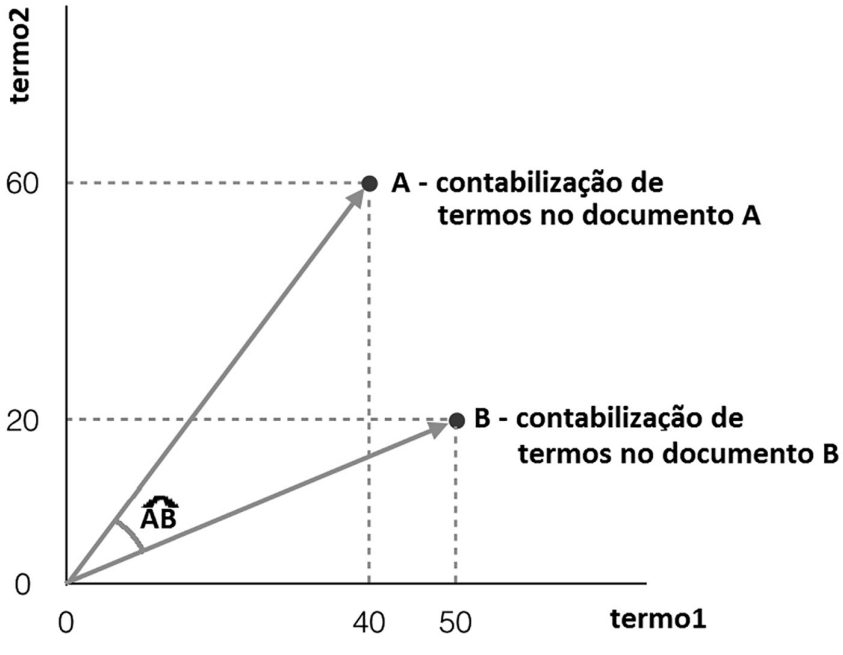
Na fórmula acima, representa o número total de documentos e o número de documentos em que aparece o monograma . Enfim calculados os pesos de cada monograma, cada documento pode ser transformado em sequências de números. Se, por um lado, perdem a estrutura e a organização que os tornaria compreensíveis ao intérprete humano, ficam especialmente legíveis pela máquina.

O mecanismo matemático pelo qual dois documentos, transformados em vetores de números, podem ser comparados, e , resultando um índice de similaridade que se situa no intervalo entre e , sendo o primeiro extremo o menos similar e o outro extremo o mais similar, é chamado distância do cosseno.

$$Similaridade = \cos\left(\widehat{AB}\right) = \frac{A.B}{A.B} = \frac{\sum_{i=1}^{T} A_i.B_i}{\sqrt{\sum_{i=1}^{T} A_i^2}.\sqrt{\sum_{i=1}^{T} B_i^2}}$$

Ou, ainda, em termos gráficos, supondo apenas dois termos, para favorecer a visualização, na forma abaixo, que permite compreender que a "distância de cosseno" representa a tendência dos documentos a "apontarem" em uma direção. A similaridade é a colinearidade, uma vez que essa medida será tão menor (sugerindo documentos mais similares) quanto menor for o ângulo indicado.

Figura 2 – Representação gráfica da similaridade de cosseno



Fonte: Elaboração própria.

Empregos da similaridade

No que segue, tomaremos como verdadeiras premissas graves sobre os documentos analisados que, apesar de imperfeitas, são abstrações que buscam aproximações sucessivas da realidade. Como se mencionou antes, tais aproximações são utilizadas na medida que demonstrarem o quão são úteis.

À pergunta "por que não empregamos simplesmente a análise humana, uma vez que não padece de tais simplificações?", responde-se que em atestando a validade da análise que resulta da análise computacional, com o mesmo esforço gasto na análise de um ou dois documentos, pode-se analisar centenas ou até mesmo milhares de documentos. Nesse sentido, aceita-se que sob assunções simplificadoras, chegue-se a conclusões mais limitadas.

O presente ensaio constitui uma inquirição acerca da utilidade dessas abstrações. É razoável, portanto, que se estabeleça desde logo a magnitude da simplificação de que tratamos. Para sermos específicos, presumimos que:

a) os termos descartados não são determinantes para a compreensão da ideia central do documento analisado;
b) a metodologia de contagem de termos, ponderada pela frequência de cada termo no universo de documentos, aproxima de maneira razoável o conteúdo semântico de cada documento;
c) a similaridade entre dois documentos calculada pelo cosseno formado entre seus vetores qualificados pela frequência relativa de termos é aproximação razoável da similaridade do conteúdo semântico dos dois.

Para estimar a aptidão da técnica em surtir resultados verossímeis e relevantes, devem-se conduzir estudos controlados, ou seja, executar a técnica para um conjunto conhecido de documentos e com um conjunto conhecido de fatos estilizados que se espera observar, de forma a que se possa cotejar os resultados obtidos pela análise humana com aqueles obtidos pela análise computadorizada.

De plano há que se distinguir entre três grandes orientações da análise, uma vez que o resultado essencial é alternativo, ou seja, ou há similaridade ou não há similaridade.

Em uma primeira abordagem, há poucos documentos – ou mesmo um único – inicial e um conjunto muito grande de documentos com os quais ele é comparado. Essa técnica pode ser chamada de "busca", pois pretende identificar candidatos. A segunda abordagem é a antípoda da primeira. Tem-se uma quantidade muito grande de documentos sendo analisados e uma quantidade pequena de documentos com os quais aqueles são comparados. Essa técnica pode ser chamada de "classificação", pois cada documento paradigma pode ser compreendido como um documento típico ou representante ideal de uma categoria.

Essas duas técnicas são as mais usuais. A primeira pode servir, no âmbito jurídico, para identificar julgamentos relevantes para a tese abordada em uma petição inicial. Ao invés de se promover uma busca por termos, compara-se a petição em mãos. A segunda técnica pode servir, no âmbito jurídico, na elaboração automática de ementas ou propor sistematização de jurisprudência.

Há, entretanto, uma terceira abordagem, em que as quantidades de documentos originais e de documentos comparados é muito próxima ou mesmo idêntica. Possivelmente trata-se dos mesmos documentos, que são comparados reciprocamente. Esta é a abordagem adotada no estudo que pode ser subdividida em dois recortes documentais. Ou analisam-se documentos que remetem a um mesmo objeto jurídico, observado em momentos diversos e sucessivos no tempo, ou analisam-se documentos que remetem a diversos objetos jurídicos da mesma natureza, observados no mesmo momento do tempo.

Leis de diretrizes orçamentárias como objeto de estudo

Nos dedicaremos à análise de peças orçamentárias e, em nosso recorte, selecionamos o mesmo ente – a União –, analisando os documentos referentes aos exercícios de 2005 até 2019. Poderíamos ter procedido diversamente, selecionando um mesmo exercício financeiro para comparar leis orçamentárias de todos os municípios de um mesmo estado da Federação, ou mesmo de todos as unidades federativas.

Esses documentos, introduzidos no direito financeiro pela Constituição Federal de 1988, fazem a conexão do plano de longo prazo com a tradicional Lei Orçamentária Anual. Seu conteúdo foi adensado com a Lei de Responsabilidade Fiscal, que abordou o tema no art. 4º.

> Art. 4º A lei de diretrizes orçamentárias atenderá o disposto no §2º do art. 165 da Constituição e:
>
> I - disporá também sobre:
>
> a) equilíbrio entre receitas e despesas;
>
> b) critérios e forma de limitação de empenho, a ser efetivada nas hipóteses previstas na alínea b do inciso II deste artigo, no art. 9º e no inciso II do §1º do art. 31; [...]
>
> e) normas relativas ao controle de custos e à avaliação dos resultados dos programas financiados com recursos dos orçamentos;
>
> f) demais condições e exigências para transferências de recursos a entidades públicas e privadas;
>
> §1º Integrará o projeto de lei de diretrizes orçamentárias Anexo de Metas Fiscais, em que serão estabelecidas metas anuais, em valores correntes e constantes, relativas a receitas, despesas, resultados nominal e primário e montante da dívida pública, para o exercício a que se referirem e para os dois seguintes.

§2º O Anexo conterá, ainda:

I - avaliação do cumprimento das metas relativas ao ano anterior;

II - demonstrativo das metas anuais, instruído com memória e metodologia de cálculo que justifiquem os resultados pretendidos, comparando-as com as fixadas nos três exercícios anteriores, e evidenciando a consistência delas com as premissas e os objetivos da política econômica nacional;

III - evolução do patrimônio líquido, também nos últimos três exercícios, destacando a origem e a aplicação dos recursos obtidos com a alienação de ativos;

IV - avaliação da situação financeira e atuarial:

a) dos regimes geral de previdência social e próprio dos servidores públicos e do Fundo de Amparo ao Trabalhador;

b) dos demais fundos públicos e programas estatais de natureza atuarial;

V - demonstrativo da estimativa e compensação da renúncia de receita e da margem de expansão das despesas obrigatórias de caráter continuado.

§3º A lei de diretrizes orçamentárias conterá Anexo de Riscos Fiscais, onde serão avaliados os passivos contingentes e outros riscos capazes de afetar as contas públicas, informando as providências a serem tomadas, caso se concretizem.

§4º A mensagem que encaminhar o projeto da União apresentará, em anexo específico, os objetivos das políticas monetária, creditícia e cambial, bem como os parâmetros e as projeções para seus principais agregados e variáveis, e ainda as metas de inflação, para o exercício subseqüente. (LC nº 101/00)

Seguindo a própria sistematização do art. 4º, a lei periódica subdivide-se usualmente em capítulos similares, com conteúdos determinados pela norma de regência. De uma lei de diretrizes orçamentárias espera-se seções acerca de metas e prioridades, alterações da lei orçamentária, limitação do empenho, transferências para o setor privado e público, endividamento público, controle de despesa com pessoal, política creditícia das agências financeiras oficiais, alterações na legislação orçamentária, transparência e assuntos correlatos.

Os fatos estilizados que esperamos observar são:

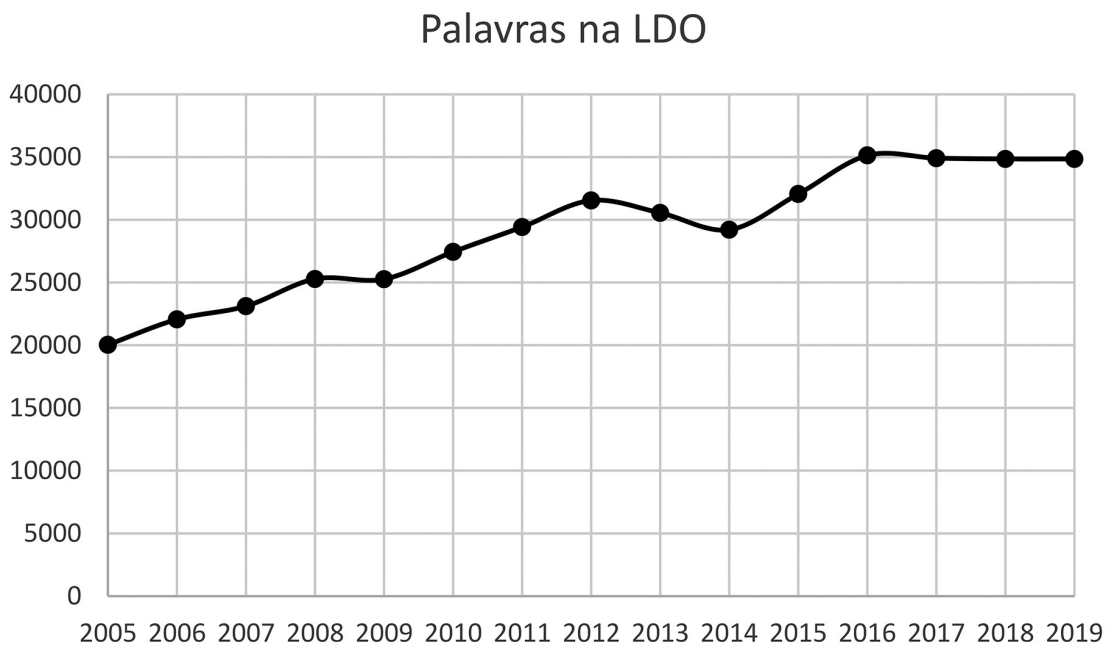
a) indução da Presidência da República, ou seja, leis oriundas de um mesmo governo tendem a ser mais similares entre si, haja vista a responsabilidade na elaboração da peça, mesmo que por intermédio de secretaria especializada;
b) indução cronológica, ou seja, leis orçamentárias mais próximas no tempo tendem a ser mais similares entre si;
c) flutuação de termos, seguindo um padrão de preocupações do direito financeiro, como obras paralisadas e contingenciamento de dotações.

Como dito antes, o trabalho analítico *a priori* pode identificar se a estratégia de análise digitalizada rende apontamentos relevantes ou mesmo consistentes com o que se espera.

Evidências da análise automática

A primeira análise que pode ser feita é a simples contagem de palavras no projeto. Tal investigação remete à seguinte conclusão:

Figura 3 – Quantidade de palavras na Lei de Diretrizes Orçamentárias

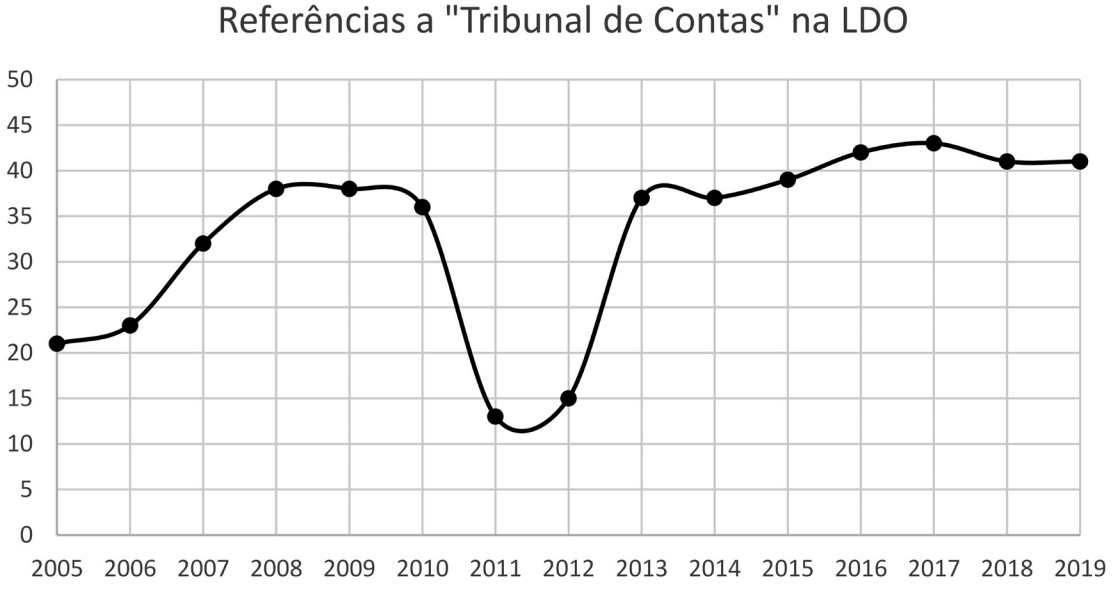


Fonte: Elaboração própria.

Durante o Governo do Presidente Luís Inácio Lula da Silva (2003 a 2011), elaboraram-se as peças até a referente ao exercício de 2012. Percebe-se aumento paulatino de disposições no documento. Possivelmente tal movimento tenha ocorrido concomitantemente com uma valorização do direito financeiro naquele período. Essa tendência se reverte nos documentos de 2013 e 2014, os dois primeiros do Governo da Presidente Dilma Rousseff, mas é retomado no documento elaborado em 2015, exercício este que teve as contas rejeitadas pelo Tribunal de Contas da União em virtude do que ficou conhecido como "pedaladas fiscais".

O Governo do Presidente Michel Temer foi responsável pela elaboração dos documentos dos exercícios de 2017, 2018 e 2019, período caracterizado por uma contagem de termos "tribunal de contas".

Figura 4 – Quantidade de referências à Corte de Contas no texto da LDO



Fonte: Elaboração própria.

Percebe-se que nos exercícios de 2011 e 2012 houve uma redução perceptível das remissões a Tribunal de Contas no texto da LDO. Possivelmente tal tenha se dado em virtude da regulamentação, ao longo da década de 2000, da competência de paralisar obras do TCU. Tal capacidade está inserta no art. 71, incs. X e XI, §§1º e 2º.

> Art. 71. O controle externo, a cargo do Congresso Nacional, será exercido com o auxílio do Tribunal de Contas da União, ao qual compete: [...]
> IX - assinar prazo para que o órgão ou entidade adote as providências necessárias ao exato cumprimento da lei, se verificada ilegalidade;
> X - sustar, se não atendido, a execução do ato impugnado, comunicando a decisão à Câmara dos Deputados e ao Senado Federal;
> XI - representar ao Poder competente sobre irregularidades ou abusos apurados.
> §1º No caso de contrato, o ato de sustação será adotado diretamente pelo Congresso Nacional, que solicitará, de imediato, ao Poder Executivo as medidas cabíveis.
> §2º Se o Congresso Nacional ou o Poder Executivo, no prazo de noventa dias, não efetivar as medidas previstas no parágrafo anterior, o Tribunal decidirá a respeito. [...]. (CF/88)

Com a regulamentação e com a paulatina experiência aplicando-se a tais disposições, instalou-se um conflito entre o Poder Executivo e o Tribunal de Contas por conta de grandes obras que tiveram recomendação de paralisação. É de se perceber pelas manchetes publicadas pela imprensa da época:

a) "TCU recomenda paralisação de 41 obras federais, sendo 13 do PAC. Relatório será enviado para o Congresso, que dará a palavra final. Ministros reclamam de críticas feitas pelo Poder Executivo ao TCU" (TCU..., 2009).
b) "Lula contraria TCU e libera verba para obras irregulares. Tribunal recomendou paralisação de quatro projetos da Petrobras por problemas 'graves'. Pagamentos liberados pelo no Orçamento chegam a R$ 13,1 bi; presidente do TCU, Ubiratan Aguiar diz que corte 'cumpriu sua parte'" (SALOMON, 2010).
c) "Tribunal de Contas da União foi alvo de críticas de Lula. Os critérios para fiscalização de obras do Tribunal de Contas da União (TCU) foram alvo constante de ataques do ex-presidente Luiz Inácio Lula da Silva durante o seu governo. O petista criticou o trabalho dos auditores diversas vezes, principalmente quando as decisões do órgão atingiram empreendimentos do Programa de Aceleração do Crescimento (PAC)" (TRIBUNAL..., 2011).

Nuvens de palavras

Uma forma de inspecionar visualmente os documentos sob o paradigma que desconsidera o conteúdo sintático e morfológico dos signos é a "nuvem de palavras". Uma vez imputadas no sistema computacional e preparadas para a análise automática, as palavras podem ser contadas.

O tratamento a que se faz referência consiste na mudança para letras minúsculas, a exclusão de toda a pontuação, a remoção de espaços em branco e a exclusão de palavras que ocorrem com muita frequência, mas quem servem apenas de conexão a outras, não importando valor cognitivo relevante quando observadas individualmente.[2]

A técnica prontamente revela a similaridade que há entre as leis de diretrizes orçamentárias, como sugerido pela análise realizada anteriormente.

[2] Por exemplo, "de", "a", "art." ou "legal".

Figura 5 – Nuvem de palavras da LDO de 2005

Fonte: Elaboração própria.

Figura 6 – Nuvem de palavras da LDO de 2019

Fonte: Elaboração própria.

Stemming e lematização

Essa primeira abordagem mais ingênua esconde uma questão que merece ser debatida. Embora a análise seja deliberadamente simplista e estilizada, o valor morfológico das palavras pode ser utilizado uma vez para aproximar signos que remetem a valores cognitivos similares. Para se tomar um exemplo simples: "graves" e "grave" importam diferença apenas se tomado em conta o contexto em que se inserem. Ora, mas o contexto é perdido e não pode ser recuperado.

A ideia simples que se apresenta é igualar, para fins da análise, as palavras "grave" e o seu plural, "graves". Há duas soluções que se apresentam. A primeira é chamada *stemming*, do inglês *stem from*, ou seja, "ser oriundo". A segunda, mais poderosa, é chamada "lematização", que remete a "lema", uma redução lógica.

A primeira técnica consiste em reduzir palavras às suas raízes, ignorando sufixos e declinações. Segundo essa técnica, tanto "grave" quanto "graves" se transformam em "grave", como também "gravidade", "gravemente" ou "agravado".

A segunda técnica consiste em obter, a partir de um dicionário de paralelos chamados de "lematas". Segundo essa técnica, as expressões "é", "foi", "será" são todas transformadas em "ser". Aqui o insumo mais relevante é o dicionário de lematas e, a despeito de fontes de formidável respeitabilidade,[3] faz-se necessário testar tais dicionários com o conteúdo jurídico, a fim de aceitá-los.

No trabalho que segue, empregaremos o *stemming* que, por dispor de dicionários mais aceitos pacificamente na comunidade, não suscitarão dúvida quanto à interpretação. Refazendo

[3] Veja, por exemplo, o trabalho do Núcleo Interinstitucional de Linguística Computacional da Universidade de São Paulo (http://nilc.icmc.usp.br).

as “nuvens de palavras” para as LDOs dos exercícios de 2005 e de 2019, temos a representação que segue.

Figura 7 – Nuvem de palavras da LDO de 2005 com *stemming*

Fonte: Elaboração própria.

Figura 8 – Nuvem de palavras da LDO de 2019 com *stemming*

Fonte: Elaboração própria.

Embora se perceba uma proximidade grande entre os vinte termos mais frequentes nos dois documentos, seu peso relativo difere. Preocupações “federativas” parecem informar grande parte do conteúdo do documento, enquanto preocupações “sociais”, parecem ter perdido algum peso de 2005 para 2019.

Para fins de comparação, é sugestivo realizar a mesma análise para um texto legal presumivelmente muito diferente. Na Figura 7 olhamos o resultado da mesma técnica quando aplicada no Código Civil de 2002.

Figura 9 – Nuvem de palavras do Código Civil de 2002 com *stemming*

Fonte: Elaboração própria.

De fato, os conteúdos são muito diversos, enquanto os conteúdos das LDOs de 2005 e de 2019 são muito próximos, embora não idênticos. A Tabela 1 demonstra numericamente o peso relativo de cada um dos dez termos mais frequentes nos dois diplomas.

Tabela 1 – Participação relativa dos 10 termos mais frequentes nas LDOs de 2005 e de 2019, com lematização

LDO 2005		**LDO 2019**	
feder	24%	feder	28%
social	21%	fiscal	18%
fiscal	13%	social	15%
inclus	10%	inclus	10%
tribun	9%	tribun	9%
central	6%	data	6%
total	6%	grave	6%
grave	4%	capit	3%
legal	3%	regim	3%
base	3%	anterior	3%
	100%		100%

Fonte: Elaboração própria.

Comparação entre documentos

Na Tabela 1 pode-se verificar uma aproximação entre os documentos de 2005 e de 2019, ainda que não se tenha a igualdade entre eles. Seria possível aquilatar quão diferentes são eles entre si? Empregaremos a análise de diferença de cosseno, como discutido ao início deste trabalho.

Algumas observações são relevantes quanto à metodologia que foi empregada:

a) Os exercícios de 2005 e de 2019 foram excluídos da comparação em virtude de se tratar de leis limítrofes, que não seriam comparadas com as anteriores ou posteriores, possivelmente surtindo resultados espúrios.
b) O resultado obtido – "distância de cosseno" – por sua própria natureza, é o contrário da "similaridade". Na Tabela 2 apresentamos a similaridade.
c) Quanto mais escura a célula maior a similaridade entre os exercícios, quanto mais clara, mais diversos.
d) Empregou-se o corretor de "inverso de frequências", mencionado ao início deste trabalho.

Tabela 2 – Similaridade entre as LDOs dos diversos exercícios

	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
2007	0,93											
2008	0,86	0,91										
2009	0,90	0,94	0,91									
2010	0,88	0,91	0,88	0,95								
2011	0,80	0,83	0,79	0,85	0,85							
2012	0,69	0,72	0,70	0,74	0,74	0,86						
2013	0,88	0,89	0,85	0,91	0,91	0,85	0,76					
2014	0,83	0,85	0,80	0,86	0,86	0,79	0,70	0,90				
2015	0,81	0,83	0,79	0,85	0,84	0,77	0,69	0,89	0,97			
2016	0,80	0,82	0,78	0,83	0,83	0,76	0,68	0,87	0,95	0,97		
2017	0,80	0,81	0,77	0,83	0,82	0,75	0,67	0,86	0,94	0,96	0,97	
2018	0,80	0,81	0,77	0,82	0,82	0,75	0,66	0,86	0,91	0,92	0,93	0,93

Fonte: Elaboração própria.

Os documentos de 2006 a 2012 foram elaborados pelo Governo Lula, os documentos de 2013 a 2017 foram elaborados pelo Governo Dilma e o documento de 2018 foi elaborado pelo Governo Temer. De fato, percebe-se grande similaridade entre os documentos dos exercícios imediatamente subsequentes, mas há quebras, a mais clara delas no documento do exercício de 2012, o mais dissimilar da série, embora o documento do exercício de 2011.

Os documentos elaborados no Governo Dilma são caracterizados por grande consistência, sendo os mais similares entre si na série. Por fim, o documento produzido pelo Governo Temer promove grande alteração, sendo perceptivelmente diferente em relação aos antecessores, embora não tanto quanto o que se observou em 2011 e 2012.

Um cálculo médio revela que o exercício de 2013 teve o documento mais similar aos demais, representando algo como “um ponto médio” em termos normativos, tal como se aprecia na Tabela 3.

Tabela 3 – Similaridade média das LDOs em relação às demais

Exercício	Similaridade
2013	87%
2009	87%
2014	86%
2015	86%
2010	86%
2007	85%
2016	85%
2017	84%
2018	83%
2006	83%
2008	82%
2011	80%
2012	72%

Fonte: Elaboração própria.

Conclusões

As novas técnicas de análise automatizada impõem desafios e nos provocam a realizar experiências em busca de usos apropriados para a tecnologia. Este estudo busca analisar os efeitos da análise automatizada de palavras. Os computadores ainda não são capazes de investigar o conteúdo semântico dos textos jurídicos e, aparentemente, esse panorama ainda está distante.

Nada obstante, há largo repertório de técnicas disponíveis para comparação de textos, desde que se ignorem suas estruturas sintáticas. Essa é a forma que os buscadores de textos utilizam para identificar textos similares. Uma vez que a técnica da hermenêutica jurídica coloca muita ênfase sobre os signos normativos e sobre a cognição contextual das normas, percebe-se que há grande distância entre as duas formas de abordar a investigação de conteúdo de um texto jurídico.

Neste estudo buscamos analisar um acervo conhecido de documentos a fim de averiguar se a análise que emerge dessas técnicas é capaz de identificar fatos estilizados esperados e se as surpresas encontradas na análise se correlacionavam a elementos conhecidos do plano jurídico. Quanto aos fatos estilizados, ponderamos que deveria haver:

a) indução da presidência da república;
b) indução cronológica;
c) flutuação de termos, seguindo um padrão de preocupações do direito financeiro.

Quanto ao primeiro ponto, foi possível identificar regiões de transição entre as mudanças de administração consistentes com uma diferença entre os documentos imediatamente subsequentes. Quanto ao segundo ponto, também foi possível perceber que documentos imediatamente subsequentes são mais similares do que os mais distantes entre si no tempo. Quanto ao terceiro termo, também se apreciou mudanças em termos empregados que, na breve análise que foi possível aqui, pareceram guardar relação com as preocupações do direito financeiro.

O fato inesperado foi a repentina redução de remissões ao Tribunal de Contas nos documentos dos exercícios de 2011 e 2012. Tal fato parece guardar relação com o momento que se observava na época, de crítica do Poder Executivo às paralisações de obras públicas então promovidas pelo TCU.

Esta pesquisa pode ser seguida pela sua repetição para outros documentos orçamentários ou outros recortes, ou ainda expandida para outras categorias de documentos jurídicos. Outro sentido poderia ser a comparação entre o resultado obtido por meio da "stemização" e aquele obtido pela lematização, possivelmente contribuindo para a criação de um dicionário de lematas jurídicos apropriado.

Referências

FURQUIM, Luis Otávio de Colla. *Agrupamento e categorização de documentos jurídicos*. Dissertação (Mestrado em Informática) – Faculdade de Informática, PUCRS, Porto Alegre, 2011.

FURQUIM, Luis Otávio de Colla; LIMA, Vera Lúcia Strube de. Clustering and categorization of Brazilian Portuguese legal documents. *In*: CASELI, Helena; VILLAVICENCIO, Aline; TEIXEIRA, António; PERDIGÃO, Fernando (Ed.). *Proceedings of the 10th international conference on Computational Processing of the Portuguese Language (PROPOR'12)*. Berlin: Heidelberg Springer-Verlag, 2012.

GONÇALVES, T.; QUARESMA, P. A preliminary approach to the multilabel classification problem of Portuguese juridical documents. *In*: PIRES, F. M.; ABREU, S. P. (Ed.). *EPIA 2003. LNCS (LNAI)*, Berlin, v. 2902, p. 435-444, 2003.

GREGHI, Juliana Galvani; MARTINS Ronaldo Teixeira; NUNES, Maria das Graças Volpe. DIADORIM – A Lexical Database for Brazilian Portuguese. *In*: RODRÍGUEZ Manuel G.; ARAUJO, Carmem P. S. (Ed.). *International Conference on Language Resources and Evaluation LREC 2002*. Las Palmas de Gran Canaria. Proceedings of the Third International Conference on Language Resources and Evaluation. 2002. v. IV. Disponível em: http://nilc.icmc.usp.br/nilc/download/GreghiMartinsNunes.pdf.

MUNIZ, M.; NUNES, M. *A construção de recursos linguístico-computacionais para o português do Brasil*: o projeto de Unitex-PB. Dissertação (Mestrado em Computação) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2004.

SALOMON, Marta. Lula contraria TCU e libera verba para obras irregulares. *Folha de S. Paulo*, 28 jan. 2010. Disponível em: https://www1.folha.uol.com.br/fsp/brasil/fc2801201019.htm. Acesso em: 1º ago. 2020.

TCU recomenda paralisação de 41 obras federais, sendo 13 do PAC. *Gazeta do Povo*, 29 set. 2009. Disponível em: https://www.gazetadopovo.com.br/vida-publica/tcu-recomenda-paralisacao-de-41-obras-federais-sendo-13-do-pac-bx2br56avlxebzy2g5x554mdq/. Acesso em: 30 jun. 2020.

TRIBUNAL de Contas da União foi alvo de críticas de Lula. *O Globo*, 18 abr. 2011. Disponível em: https://oglobo.globo.com/politica/tribunal-de-contas-da-uniao-foi-alvo-de-criticas-de-lula-2794723. Acesso em: 12 ago. 2020.

Informação bibliográfica deste texto, conforme a NBR 6023:2018 da Associação Brasileira de Normas Técnicas (ABNT):

SARQUIS, Alexandre Manir Figueiredo. Similaridade de textos normativos: um ensaio sobre as leis orçamentárias. *Controle Externo: Revista do Tribunal de Contas do Estado de Goiás*, Belo Horizonte, ano 2, n. 3, p. 153-166, jan./jun. 2020.

JURISPRUDÊNCIA SELECIONADA
ACÓRDÃOS NA ÍNTEGRA

JURISPRUDÊNCIA – INTEIRO TEOR

A seção de julgados e pareceres destina-se a divulgar decisões e manifestações relacionadas a temas relevantes para os tribunais de contas, para gestores, demais atores do sistema de controle brasileiro, para a administração pública e pesquisadores.

Como não poderia deixar de ser, a escolha dessa edição recaiu sobre o controle externo das ações públicas voltadas ao combate dos efeitos da pandemia da COVID19.

A primeira, do Tribunal de Contas da União, trata da concessão de auxílio emergencial de caráter pessoal, cuja materialidade para o Governo Federal foi estimada em mais de 68 (sessenta e oito) bilhões de reais. O TCU atua na fiscalização desse subsídio, e, no presente caso, identificou irregularidades na concessão quanto a beneficiários que não preenchem os requisitos legais para a percepção do benefício.

Vale mencionar que os tribunais de contas estaduais, após aderirem a um acordo de cooperação, também atuam no âmbito de suas competências para identificar, apurar e coibir referidas irregularidades em suas esferas de atuação.

Por fim, a segunda decisão apresenta o resultado da fiscalização concomitante do TCE do Rio Grande do Norte na contratação de EPI's pela Secretaria de Estado da Saúde, cuja atuação resultou em expressiva economia aos cofres estaduais.