Text as data

desafios e oportunidades para as ciências sociais

Davi Moreira

Universidade de São Paulo

www.davimoreira.com
davi.moreira@gmail.com

26 de abril de 2017

Análise automatizada de texto

- Oportunidades
- Aplicação
- Desafios

Oportunidades

- Redução substantiva de custos
- Analise de grandes bases de conteúdo
- Classificação de novos documentos

Oportunidades

The Manifesto Project provides the scientific community with parties' policy positions derived from a content analysis of parties' electoral manifestos. It covers over 1000 parties from 1945 until today in over 50 countries on five continents.

CAP COMPARATIVE AGENDAS PROJECT
COMPARING POLICIES WORLDWIDE

CAP enables scholars, students, policy-makers and the media to investigate trends in policy-making across time and between countries. It classifies policy activities into a single, universal and consistent coding scheme. CAP monitors policy processes by tracking the actions that governments take in response to the challenges they face.

Oportunidades

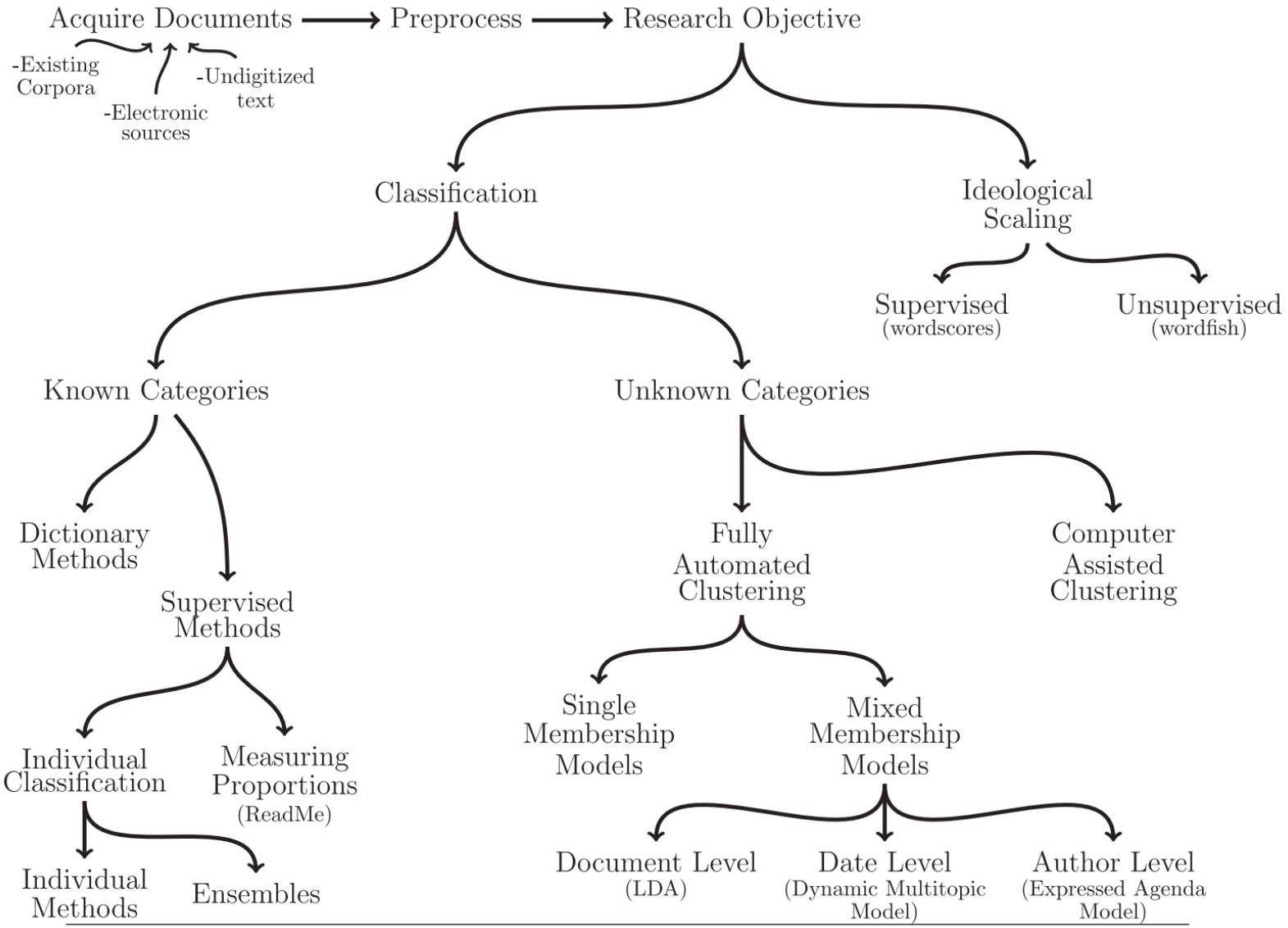
Trecho de fala do Deputado Federal Glauber Braga (PSOL-RJ) durante seu voto no processo de impeachment da então Presidenta da República Dilma Rousseff.

Oportunidades

Retórica Parlamentar

Projeto experimental desenvolvido no primeiro Hackathon da Câmara dos Deputados (2013) por Davi Moreira, Manoel Galdino e Luis Carli. Posteriormente incubado pelo Laboratório Hacker Câmara dos Deputados.

Text as data methods

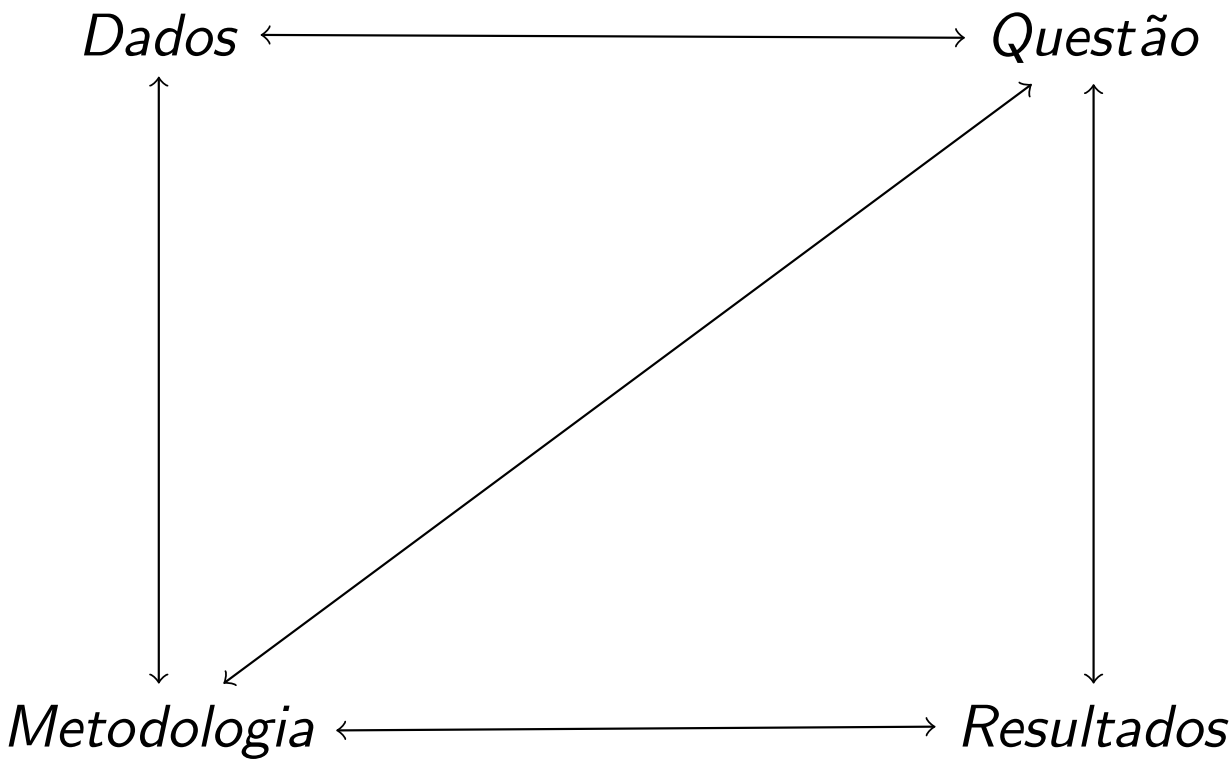


Grimmer, Justin e Brandon M. Stewart. 2013. *Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts*. Political Analysis

Aplicação

$$\textit{Questão} \longrightarrow \textit{Dados} \longrightarrow \textit{Metodologia} \longrightarrow \textit{Resultados}$$

Aplicação



Aplicação

Simon Jackman

Perspectives on Political Methodology: Interview with Simon Jackman. (2013)

Justin Grimmer

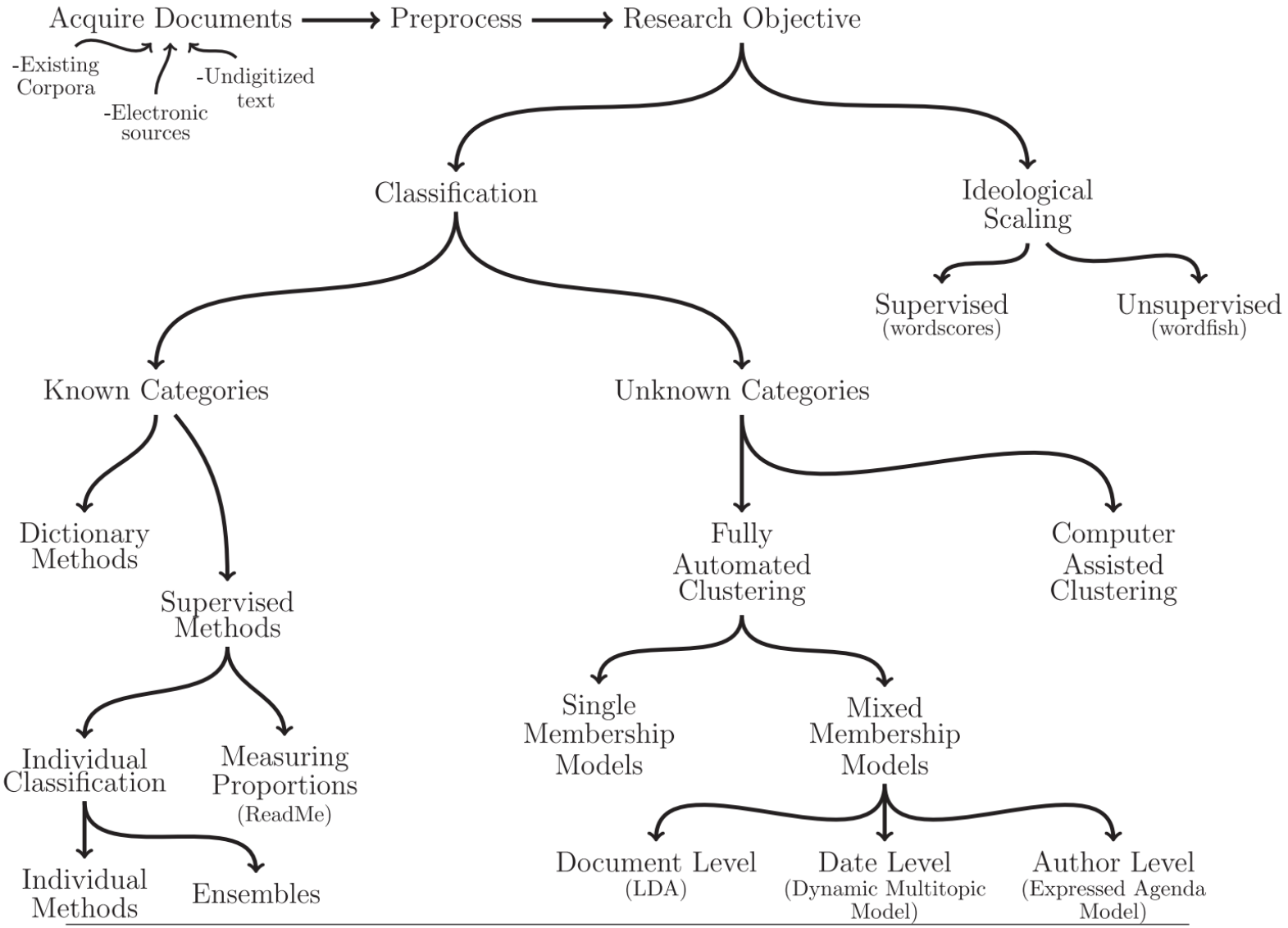
Aplicação

A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases

Justin Grimmer
Department of Government, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138
e-mail: jgrimmer@fas.harvard.edu (corresponding author)

"The statistical model and its extensions will be made available in a forthcoming free software package for the R computing language".

Text as data methods



Grimmer, Justin e Brandon M. Stewart. 2013. *Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts*. Political Analysis

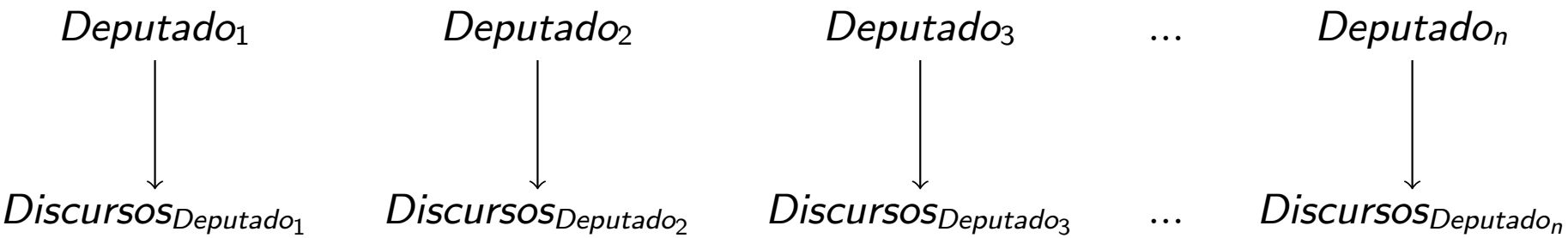
Topic Model

David M. Blei

Aplicação

[05/09/2013] "...oportunidade significativa de dar uma grande **contribuição** para o **desenvolvimento** do **esporte** nacional... projeto de **lei** que limita, entre outras medidas, a **reeleição** dos **dirigentes esportivos**. Eles terão **mandato** de 4 anos e **direito** a concorrer a apenas mais um **mandato**.
...E o mais importante: as **confederações** e **federações esportivas** só poderão receber **verbas públicas** e **isenções tributárias** se cumprirem as novas **regras**...."

Expressed Agenda Model



Aplicação

Com a palavra os nobres deputados

frequência e ênfase temática dos discursos dos parlamentares brasileiros

Aplicação

Discurso proferido no dia 10 de junho de 2013 no Grande Expediente pelo Deputado Wanderley Alves de Oliveira-PSC-RJ ("Deley"):

> *(...) neste dia de hoje, aproveitando esta oportunidade - porque todos sabemos que o Grande Expediente é feito através de sorteio e que, muitas vezes, é a chance para que possamos trazer parte daquilo que temos feito, parte de nossos pensamentos, até porque a dinâmica desta Casa muitas vezes nos dá poucas oportunidades - queremos estar falando com aquele companheiro, com aquela companheira, que muitas vezes, até por informações distorcidas, entende que trabalhamos aqui somente nas terças, quartas e, às vezes, quintas*

Aplicação

Questão de pesquisa:

A comunicação parlamentar no âmbito da Câmara dos Deputados é governada pela relação governo-oposição, assim como constatado em sua atuação no processo decisório?

Aplicação

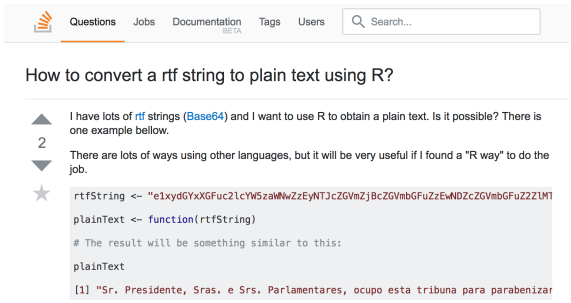
Web Scraping

Web services

Aplicação

"e1xydGYxX
GFuc2lcYW5zaWN
wZzEyNTJ cZGVmZ
jBcZGVmbGFu
ZzEwN DZcZ
GVmbGFu Z2ZlMT
A0NlxkJ..."

Formato rtfBase64



Stack OverFlow

"Sr. Presidente, Sras. e Srs. Parlamentares, ocupo esta tribuna para parabenizar a torcida paraense. O futebol paraense deu um show de civilidade neste final de semana, e a torcida bicolor do Papão da Curuzu, o Paysandu, está de parabéns..."

Plain text

Aplicação

Bag of Words

- Caixa baixa
- Remoção de números, pontuação, *stop words* e palavras selecionadas
- pacotes `R`: tm
- obtenção das raízes das palavras (*stems*)
- Algoritmo de Porter (NILC - USP) - `www.nilc.icmc.usp.br/`
- Snowball Project - `http://snowball.tartarus.org/`

Expressed Agenda Model

- Mínimo de dois documentos por autor
- Cada documento será classificado em um único tópico - Pequeno Expediente
- O *n* de tópicos existente no *corpus* deve ser definido a priori

Aplicação

Tabela: Resultado do tratamento aplicado à coleção de discursos de cada legislatura

	51	52	53	54
# inicial de discursos	19.883	41.756	36.613	35.398
# final de discursos	19.064	39.702	35.075	33.941
# inicial de oradores	526	572	573	591
# final de oradores	491	548	543	552
# inicial de palavras únicas	119.480	160.674	152.791	153.111
# Raízes únicas	4.104	3.989	3.757	3.906

Redução brusca de informação: *Trade-off* entre modelos estatísticos de classficação e a mão de obra humana para análise de conteúdo

Aplicação

Tabela: *Document Term Matrix* (DTM)

	$stem_1$	$stem_2$	...	$stem_m$
$Discurso_1$	$Freq_{stem_1}$	$Freq_{stem_2}$	...	$Freq_{stem_m}$
$Discurso_2$	$Freq_{stem_1}$	$Freq_{stem_2}$	...	$Freq_{stem_m}$
...	...	...	...	...
$Discurso_n$	$Freq_{stem_1}$	$Freq_{stem_2}$	...	$Freq_{stem_m}$

Tabela: Deputados $\times$ Discursos

	Primeiro	Último
$Deputado_1$	$discurso_{11}$	$discurso_{n2}$
$Deputado_2$	$discurso_{12}$	$discurso_{n3}$
...	...	...
$Deputado_d$	$discurso_{1d}$	$discurso_{nd}$

Aplicação

Definição do número de Tópicos

Estatisticamente, tópicos são funções densidade de probabilidade de palavras (raízes), que determinam a probabilidade de uma palavra ser usada em um discurso sobre um tema.

- *Dirichlet process prior*
- Análise qualitativa de diferentes modelos

Aplicação - Matéria Prima

Tabela: Dados descritivos da quantidade de discursos no Pequeno Expediente por Deputado Federal em cada legislatura que serão utlizados no *expressed agenda model*

	51	52	53	54
Min.	2	2	2	2
1Q.	7	15	12.5	10
Mediana	20	38.5	33	28
Média	38.83	72.45	64.59	61.49
3Q.	48.50	94	76	65.25
Máx.	783	1.081	983	1.190
# Deputados	491	548	543	552
# Discursos	19.064	39.702	35.075	33.941
# Raízes únicas	4.104	3.989	3.757	3.906
# Tópicos	27	31	31	39

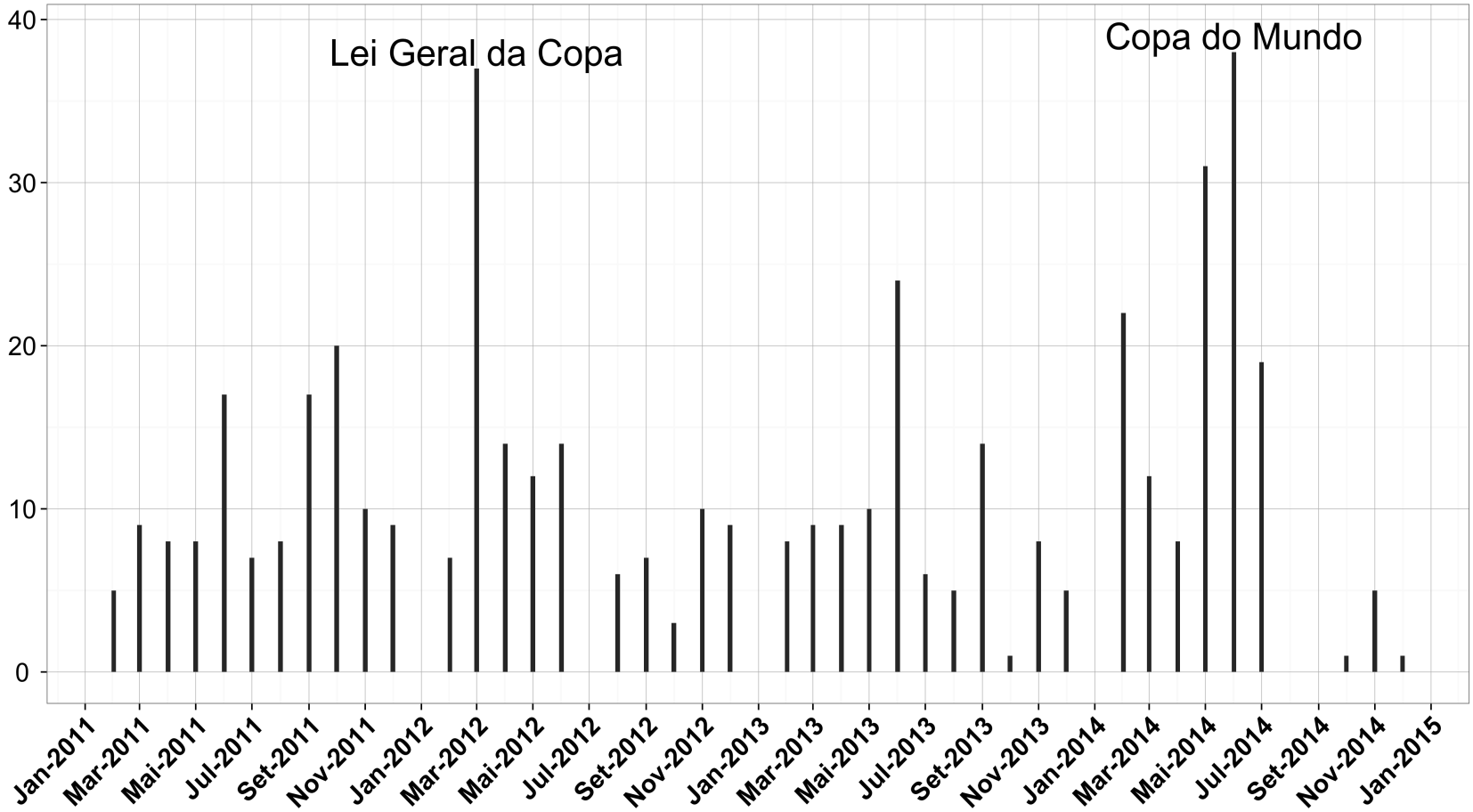
Aplicação - Validação

Tabela: Temas selecionados dos discursos proferidos na legislatura 51

Rótulo	Raízes	%
...	...	...
Sistema Político	polit, pov, candidat, trabalh, eleitoral	6.1
Trabalho	trabalh, direit, projet, empreg, lei	5.7
Amazônia e Meio ambiente	regia, desenvolv, amazon, agu, projet	4.4
Direitos Humanos e Minorias	direit, human, sociedad, pesso, negr	4.1
Economia	econom, polit, desenvolv, ano, setor	4.0
Corrupção	cpi, denunc, corrupca, senador, fat	3.8
Agropecuária	produtor, produca, agricultur, produt, milho	3.4
Educação	educaca, ensin, escol, univers, professor	3.3
...	...	...

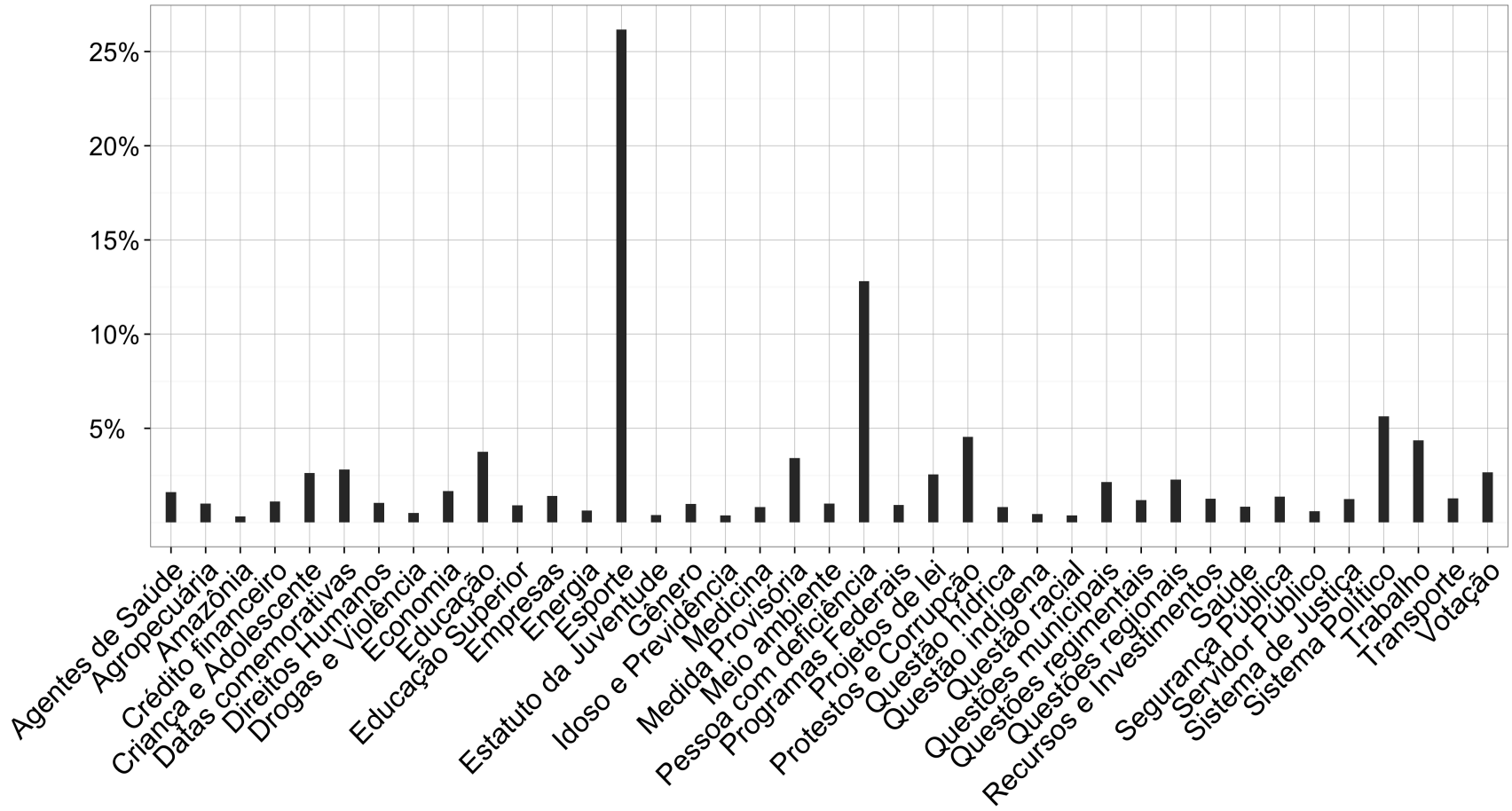
Aplicação - Validação

Figura: Pronunciamentos classificados na categoria Esporte ao longo da 54ª legislatura



Aplicação - Validação

Figura: Ênfase Temática dos Pronunciamentos realizados pelo Deputado Federal Romário PSB-RJ na 54ª legislatura



Desafios

Balanço - Agendas Econômica e Social

$$\text{Balanço}_{Deputado_i} = \hat{\text{E}}\text{nf. Social}_{Deputado_i} - \hat{\text{E}}\text{nf. Economia}_{Deputado_i} \quad (1)$$

Quanto maior o valor da variável Balanço, maior é a dedicação relativa de um $Deputado_i$ a temas da área social. Quanto menor o valor da variável Balanço, maior é a dedicação relativa de um $Deputado_i$ a temas da área econômica.

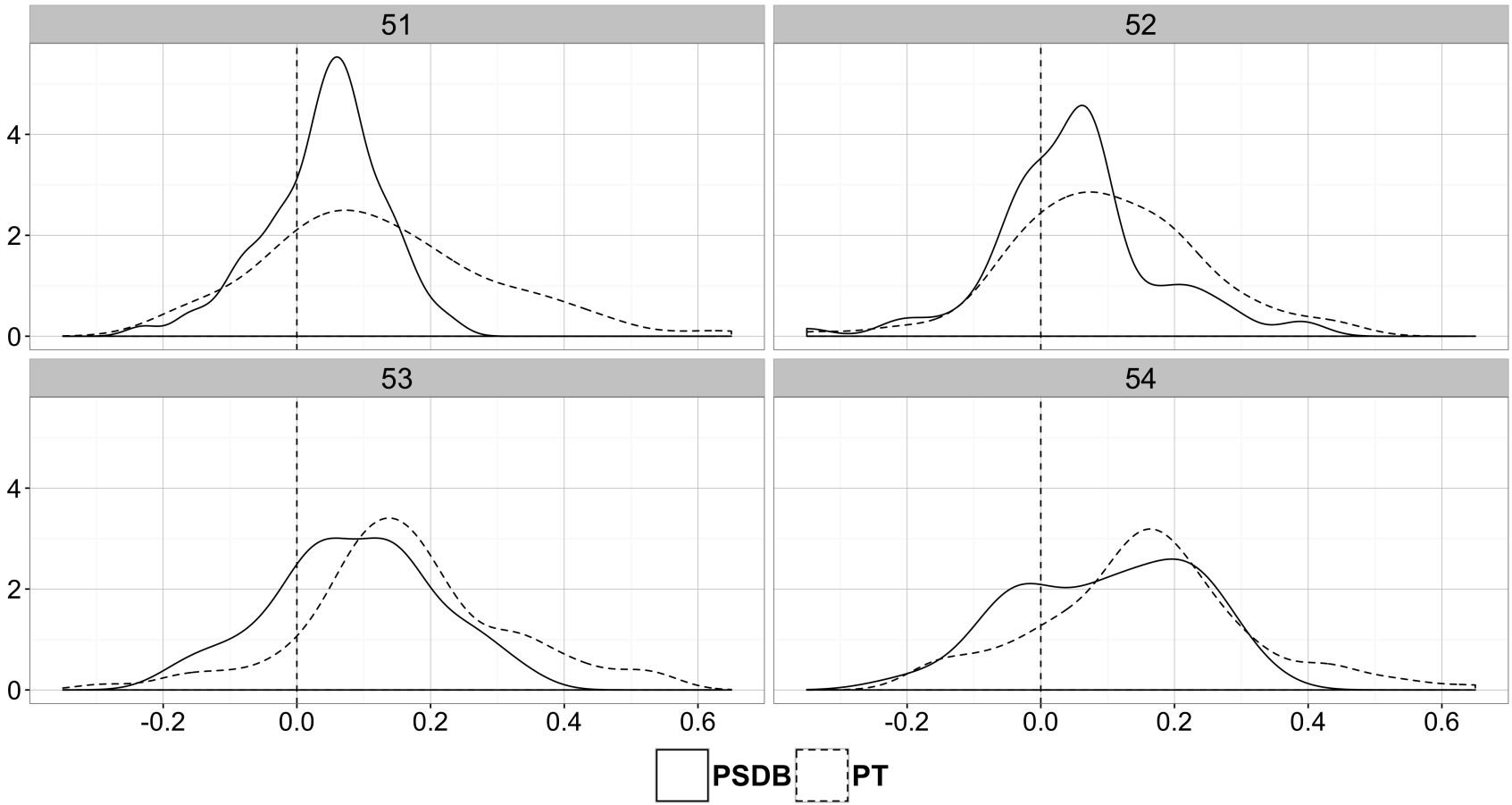
Aplicação - Validação

Tabela: Temas selecionados dos discursos proferidos na legislatura 51

Rótulo	Raízes	%
...	...	...
Sistema Político	polit, pov, candidat, trabalh, eleitoral	6.1
Trabalho	trabalh, direit, projet, empreg, lei	5.7
Amazônia e Meio ambiente	regia, desenvolv, amazon, agu, projet	4.4
Direitos Humanos e Minorias	direit, human, sociedad, pesso, negr	4.1
Economia	econom, polit, desenvolv, ano, setor	4.0
Corrupção	cpi, denunc, corrupca, senador, fat	3.8
Agropecuária	produtor, produca, agricultur, produt, milho	3.4
Educação	educaca, ensin, escol, univers, professor	3.3
...	...	...

Desafios - Balanço - Agendas Econômica e Social

Figura: Distribuição da variável Balanço dos Deputados Federais do PT e do PSDB em cada legislatura



Desafios

Irregular forms: Creation - Kupka - 1911

Obrigado