

# Modelo de Tópicos e a Prefeitura de SP

// Topic  
Model //

//  
Resultados //

// Workflow //

// Futuro //

Guilherme G. Nicolau /  
Ironhack DA Student  
21/05/2021

# O que é Topic Model?

- Textual Natural Language Processing
- não-supervisionado;
- não-estruturado;
- Nos retorna tópicos no texto;

// O que  
é um  
tópico? //

// Como é  
um  
tópico? //

# // O que é um tópico? //

- 1) A principal característica dos modelos de tópicos é sua capacidade de realizar uma redução dimensional do espaço definido pelo modelo *bag-of-words* de forma a capturar estruturas semânticas presentes no espaço. Além disso temos que o novo espaço, dito espaço de tópicos, é um modelo probabilístico para a ocorrência de palavras nos documentos; ou seja, busca-se a covariância entre as palavras em um documento e a relação entre os documentos (Corpus). Vamos além da contagem de palavras.
- 2) Temos menos bias do que o 'tageamento' de um sujeito e, portanto, melhor classificação que um humano; é ótimo, por exemplo, para informational retrieval, ou seja, encontrar um conteúdo textual específico em um grande conjunto de dados;

"It is important to note from the start that the similarity estimates derived by LSA are not simple contiguity frequencies, co-occurrence counts, or correlations in usage, but depend on a powerful mathematical analysis that is capable of correctly inferring much deeper relations (thus the phrase "Latent Semantic"), and as a consequence are often much better predictors of human meaning-based judgments and performance than are the surface level contingencies that have long been rejected (or, as Burgess and Lund, 1996 and this volume, show, unfairly maligned) by linguists as the basis of language phenomena"(LANDAUER; FOLTZ; LAHAM, 1998)

# // O que é um tópico? //

- 1) A principal característica dos modelos de tópicos é sua capacidade de realizar uma redução dimensional do espaço definido pelo modelo *bag-of-words* de forma a capturar estruturas semânticas presentes no espaço. Além disso temos que o novo espaço, dito espaço de tópicos, é um modelo probabilístico para a ocorrência de palavras nos documentos; ou seja, busca-se a covariância entre as palavras em um documento e a relação entre os documentos (Corpus). Vamos além da contagem de palavras.
- 2) Temos menos bias do que o 'tageamento' de um sujeito e, portanto, melhor classificação que um humano; é ótimo, por exemplo, para informational retrieval, ou seja, encontrar um conteúdo textual específico em um grande conjunto de dados;

"It is important to note from the start that the similarity estimates derived by LSA are not simple contiguity frequencies, co-occurrence counts, or correlations in usage, but depend on a powerful mathematical analysis that is capable of correctly inferring much deeper relations (thus the phrase "Latent Semantic"), and as a consequence are often much better predictors of human meaning-based judgments and performance than are the surface level contingencies that have long been rejected (or, as Burgess and Lund, 1996 and this volume, show, unfairly maligned) by linguists as the basis of language phenomena"(LANDAUER; FOLTZ; LAHAM, 1998)

# O que é Topic Model?

- Textual Natural Language Processing
- não-supervisionado;
- não-estruturado;
- Nos retorna tópicos no texto;

// O que  
é um  
tópico? //

// Como é  
um  
tópico? //

# // Como é um tópico? //

## TÓPICOS DE CULTURA

(17, '0.228\*"anos" + 0.120\*"show" + 0.097\*"projeto" + 0.080\*"apresentacoes" + 0.076\*"musicas" + 0.074\*"banda" + 0.025\*"acervo" + 0.022\*"largo" + 0.018\*"pontos" + 0.018\*"novo")

(4, '0.164\*"sobre" + 0.112\*""o" + 0.092\*"biblioteca" + 0.075\*"diretor" + 0.062\*"exposicao" + 0.054\*"obras" + 0.051\*"reune" + 0.042\*"andrade" + 0.042\*"mario" + 0.037\*"primeiro")

(10, '0.181\*"ate" + 0.126\*"danca" + 0.092\*"inscricoes" + 0.069\*"artistas" + 0.068\*"musical" + 0.066\*"programa" + 0.052\*"janeiro" + 0.049\*"recebe" + 0.035\*"atividades" + 0.023\*"grupos")

# // Como é um tópico? //

## TÓPICOS DE CULTURA

(17, '0.228\*"anos" + 0.120\*"show" + 0.097\*"projeto" + 0.080\*"apresentacoes" + 0.076\*"musicas" + 0.074\*"banda" + 0.025\*"acervo" + 0.022\*"largo" + 0.018\*"pontos" + 0.018\*"novo")

(4, '0.164\*"sobre" + 0.112\*""o" + 0.092\*"biblioteca" + 0.075\*"diretor" + 0.062\*"exposicao" + 0.054\*"obras" + 0.051\*"reune" + 0.042\*"andrade" + 0.042\*"mario" + 0.037\*"primeiro")

(10, '0.181\*"ate" + 0.126\*"danca" + 0.092\*"inscricoes" + 0.069\*"artistas" + 0.068\*"musical" + 0.066\*"programa" + 0.052\*"janeiro" + 0.049\*"recebe" + 0.035\*"atividades" + 0.023\*"grupos")

# O que é Topic Model?

- Textual Natural Language Processing
- não-supervisionado;
- não-estruturado;
- Nos retorna tópicos no texto;

// O que  
é um  
tópico? //

// Como é  
um  
tópico? //

# Modelo de Tópicos e a Prefeitura de SP

// Topic  
Model //

//  
Resultados //

// Workflow //

// Futuro //

Guilherme G. Nicolau /  
Ironhack DA Student  
21/05/2021

# // Workflow //

#1

Coletamos 200 mil notícias institucionais dos sites das secretarias da Prefeitura de SP;

#2

Automatizado através de uma função, incorporamos tudo em um pandas DataFrame, cada linha com duas colunas: data e texto;

#3

Sincronizamos com um banco de dados PostgreSQL;

#4

Automatizado através de uma função, limpamos o texto da tabela (tiramos url, maiusculas, acentos, numeros, datas, letras unicas, simbolos, stopwords etc)

#5

Fizemos lematização - o que é? (processo lento de deep learning - nem tudo apresentado aqui incorporou a lematização)

#6

Criamos uma função para automatizar o modelo LDA para cada recorte de análise;

#7

Não incorporamos hiperparâmetros, pq? Mostrar para o cliente como o algoritmo funciona com os mesmos parâmetros em diferentes conjuntos de dados;

#8

Visualizações (automatizamos o processo em funções para retornar uma visualização para cada secretaria por gestão)

# // Workflow //

#1

Coletamos 200 mil notícias institucionais dos sites das secretarias da Prefeitura de SP;

#2

Automatizado através de uma função, incorporamos tudo em um pandas DataFrame, cada linha com duas colunas: data e texto;

#3

Sincronizamos com um banco de dados PostgreSQL;

#4

Automatizado através de uma função, limpamos o texto da tabela (tiramos url, maiusculas, acentos, numeros, datas, letras unicas, simbolos, stopwords etc)

#5

Fizemos lematização - o que é? (processo lento de deep learning - nem tudo apresentado aqui incorporou a lematização)

#6

Criamos uma função para automatizar o modelo LDA para cada recorte de análise;

#7

Não incorporamos hiperparâmetros, pq? Mostrar para o cliente como o algoritmo funciona com os mesmos parâmetros em diferentes conjuntos de dados;

#8

Visualizações (automatizamos o processo em funções para retornar uma visualização para cada secretaria por gestão)

# // Visualizações //

#1

WordClouds

#2

pyLDAvis;

#3

Retorno de documentos mais importantes por tópico  
no corpus e de cruzamento entre tópicos;

#4

Cruzamento de assuntos criados por nós com tópicos  
no corpus por gestão;

# Modelo de Tópicos e a Prefeitura de SP

// Topic  
Model //

//  
Resultados //

// Workflow //

// Futuro //

Guilherme G. Nicolau /  
Ironhack DA Student  
21/05/2021

# // Visualizações //

comparação gestão Haddad (2013-2016) e Dória-Covas (2017-2020)

#1

WordClouds

#2

pyLDAvis;

#3

Retorno de documentos mais importantes por tópico  
no corpus e de cruzamento entre tópicos;

#4

Cruzamento de assuntos criados por nós com tópicos  
no corpus por gestão;

# // Visualizações //

comparação gestão Haddad (2013-2016) e Dória-Covas (2017-2020)

#1

WordClouds

#2

pyLDAvis;

#3

Retorno de documentos mais importantes por tópico  
no corpus e de cruzamento entre tópicos;

#4

Cruzamento de assuntos criados por nós com tópicos  
no corpus por gestão;

# // Visualizações //

#1

WordClouds

## ONE BIG TABLE

CORPUS  
(2003-2020)

comparação gestão Haddad (2013-2016) e Dória-Covas (2017-2020)



obs: Generalização

# // Visualizações //

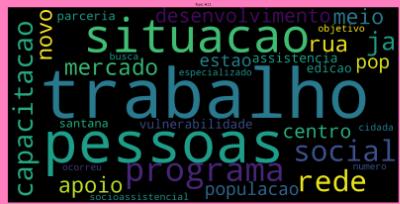
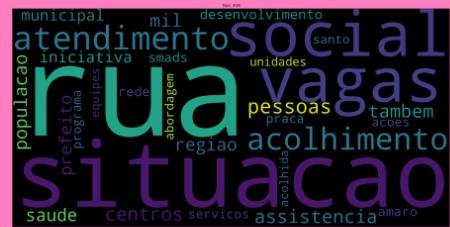
comparação gestão Haddad (2013-2016) e Dória-Covas (2017-2020)

#1

## WordClouds

SMADS

Dória-Covas  
(2017-2020)



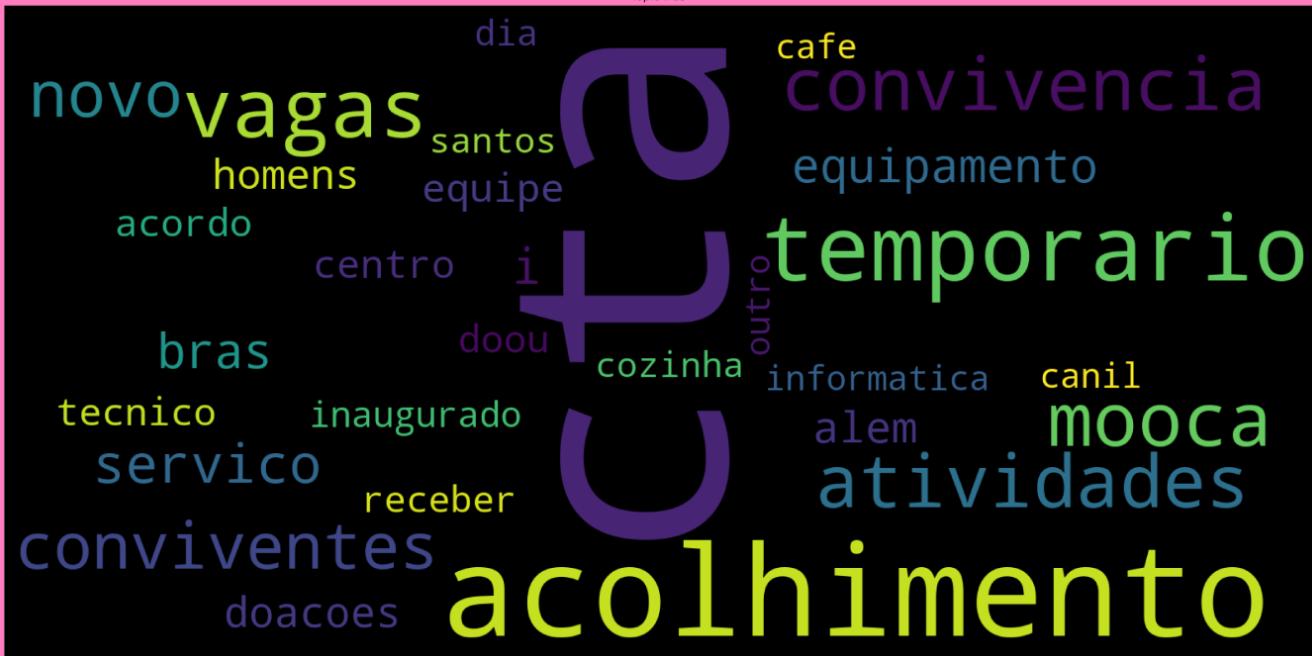
obs: Trabalho e Mercado

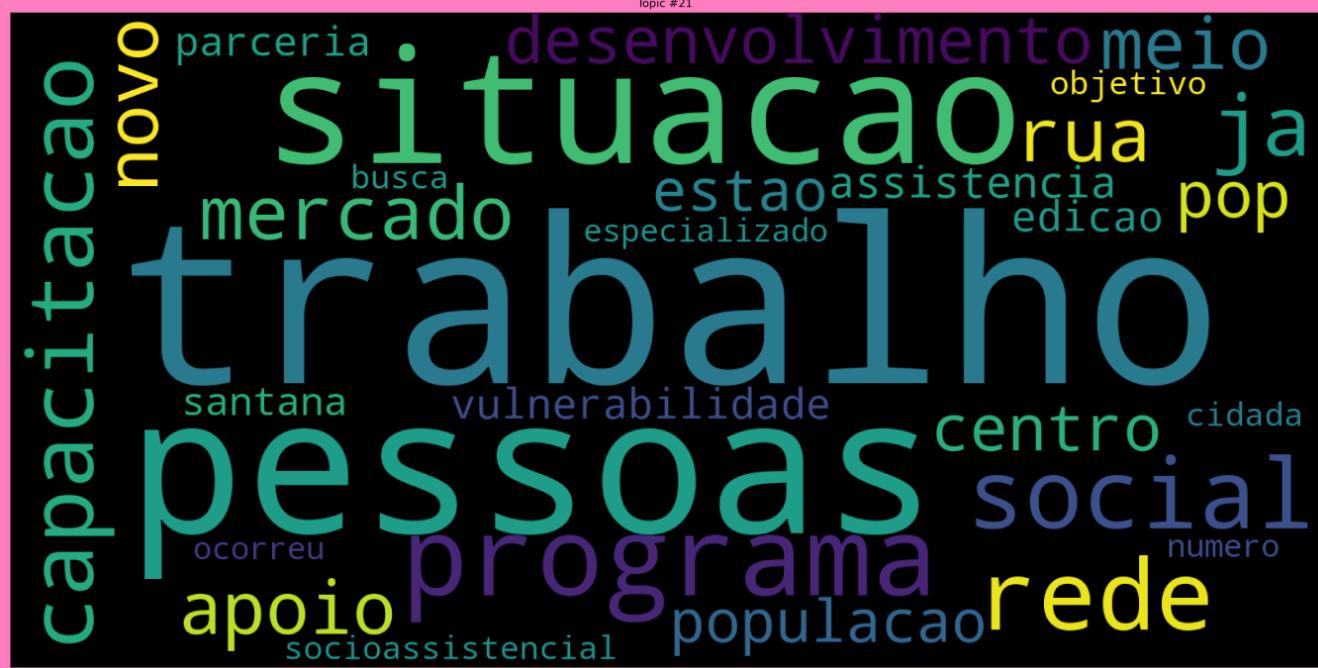
Topic #28

topic #28

municipal mil desenvolvimento  
atendimentos social  
iniciativa smads santo  
populacao unidades  
prefeito rede pessoas tambem  
programa equipes praca acoes  
rua abordagem regiao acolhimento  
situacao saude centros servicos assistencia amaro

Topic #13





## Topic #5

Topic #8

# // Visualizações //

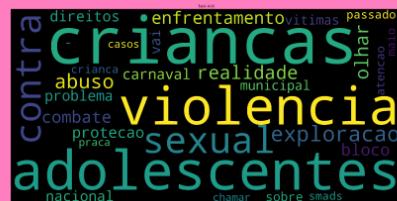
comparação gestão Haddad (2013-2016) e Dória-Covas (2017-2020)

#1

WordClouds

SMADS

Haddad  
(2013-2016)



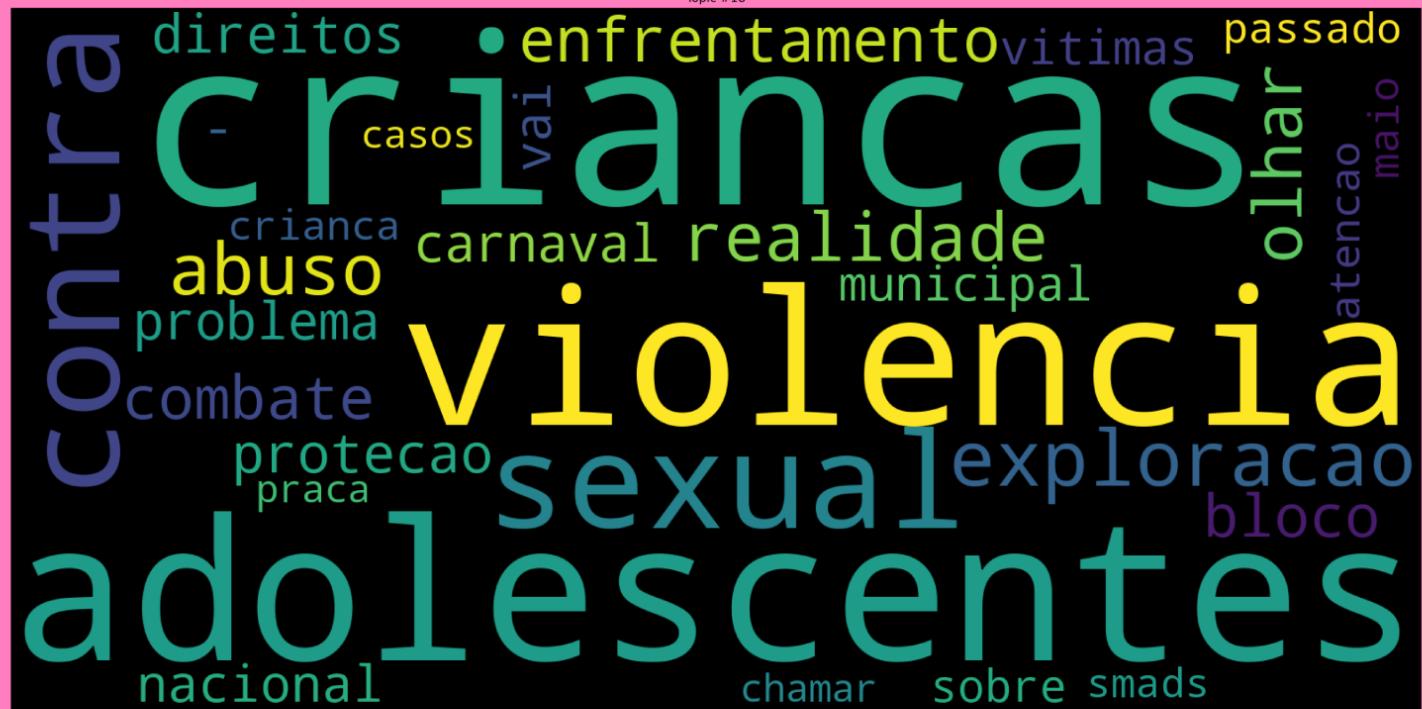
obs: Direitos Humanos

Topic #0

desenvolvimento cristina  
Smads temer  
principais municipal  
gentel luciana tambem  
ser cordeiro  
municipais apresentacao  
so jovens  
gestao adjunta publico  
contou alem  
pode alem  
todas possuir medidas ultima  
assistencia

Topic #7

familia assistencia parelheiros indigenas aldeias cadunico unico  
familias federal programas cras social governo sociais inseridas  
beneficiarias bolsa renda reciclazaro acoes insercao nacioanal cadastro paif  
beneficio outros conhecer alexandre gomes insercao



Topic #20

The word cloud contains the following words and their approximate bounding boxes:

- baixas (center, large)
- operacao (center, large)
- temperaturas (center, large)
- mil (top left, medium)
- jantar (top left, medium)
- banho (bottom left, medium)
- carater (top center, medium)
- abordagem (middle left, medium)
- vagas (bottom left, small)
- abrigos (bottom left, small)
- comprometido (top center, small)
- especializados (top center, small)
- civil (middle center, small)
- situacao (middle center, small)
- encaminhamento (middle center, small)
- discutidos (bottom center, small)
- central (bottom center, small)
- maio (top right, medium)
- atendimento (top right, small)
- william (top right, small)
- desprotecoes (middle right, small)
- centros (middle right, small)
- acolhida (bottom right, small)
- oc (bottom right, small)
- reciclageis (bottom right, small)



# // Visualizações //

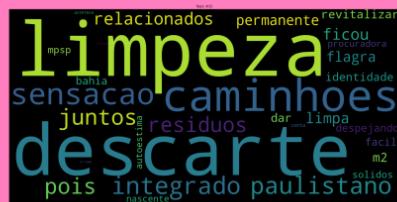
comparação gestão Haddad (2013-2016) e Dória-Covas (2017-2020)

#1

WordClouds

SMSU

Dória-Covas  
(2017-2020)



obs: Zeladoria

secretarios  
fevereiro  
cartas  
representara  
ugolini  
prefeito  
arcogiz  
secretario  
municipais  
ipiranga  
unibanco  
risco"  
praca  
operacao  
empresas  
av  
linda  
presidentes  
publicas  
andar  
secretariado  
republica  
rua  
abril  
igarape  
pratos

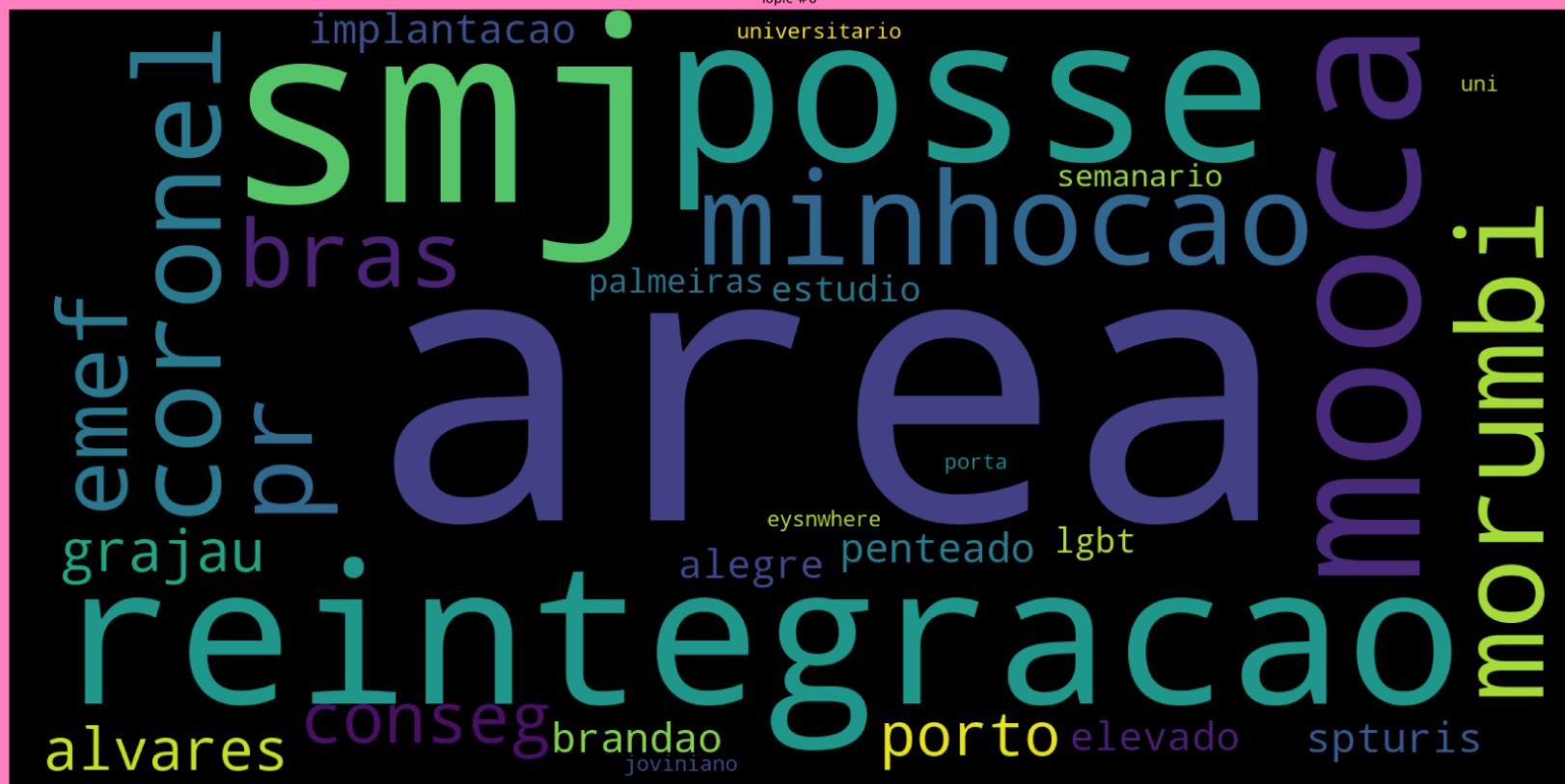
The word cloud features the following words:

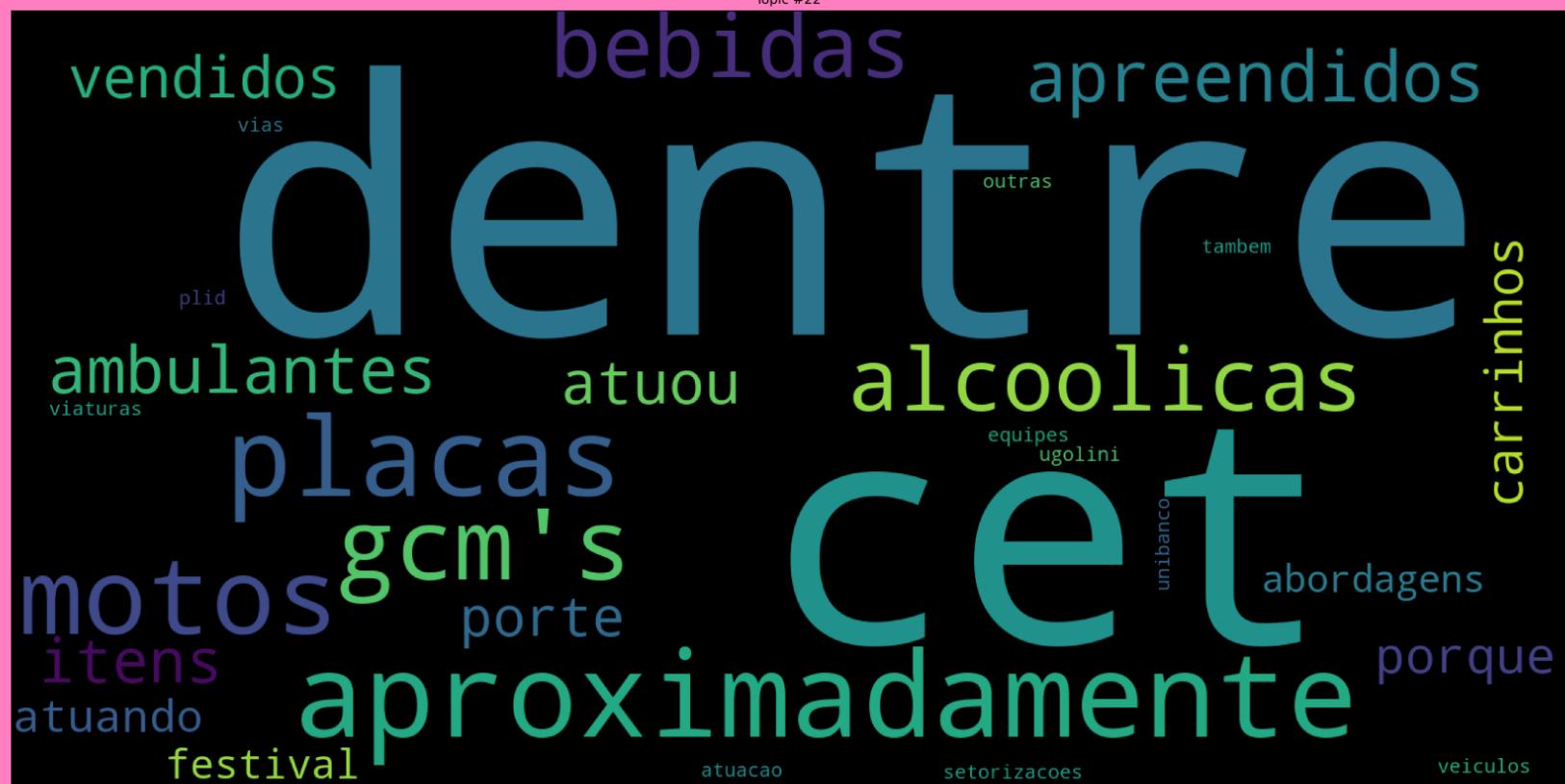
- limpeza** (large, green)
- descarte** (large, blue)
- relacionados** (yellow)
- permanente** (yellow)
- revitalizar** (green)
- ficou** (purple)
- procuradora** (purple)
- flagra** (purple)
- identidade** (teal)
- sensacao** (blue)
- caminhões** (blue)
- juntos** (yellow)
- residuos** (purple)
- dar** (blue)
- limpa** (blue)
- despejando** (blue)
- facil** (teal)
- m2** (teal)
- solidos** (teal)
- pois** (blue)
- integrado** (blue)
- paulistano** (teal)
- linda** (purple)
- bahia** (blue)
- autoestima** (blue)
- crime** (blue)
- acontece** (green)
- mpsp** (blue)
- nascente** (teal)

Topic #14

videomonitoramento flagraram  
homem喷  
familia prendeu  
floriza infrator  
conduzida latas  
"lancamento"  
spray sede  
immediatamente prende instaurado  
telefone deic  
encontradas laterais  
autorizou policial  
decisao candido  
pichando desempenho  
operadas

tiradentes  
arma joao  
acordo municipes  
linda  
detidos  
predio  
central  
policial  
atraves  
apos mundial  
conduzidas  
distrito aguas  
janeiro.  
zeladoria  
los  
flagrante  
consolacao  
regional novo  
atletico  
gcmf  
praca  
restaurante  
jabaquara





# // Visualizações //

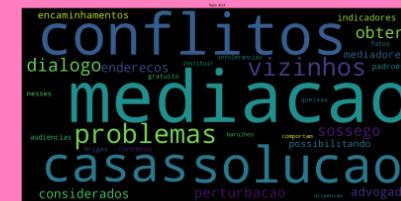
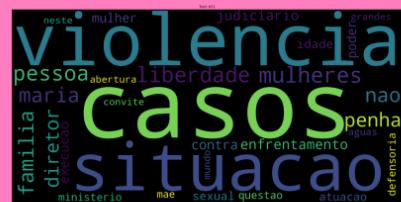
comparação gestão Haddad (2013-2016) e Dória-Covas (2017-2020)

#1

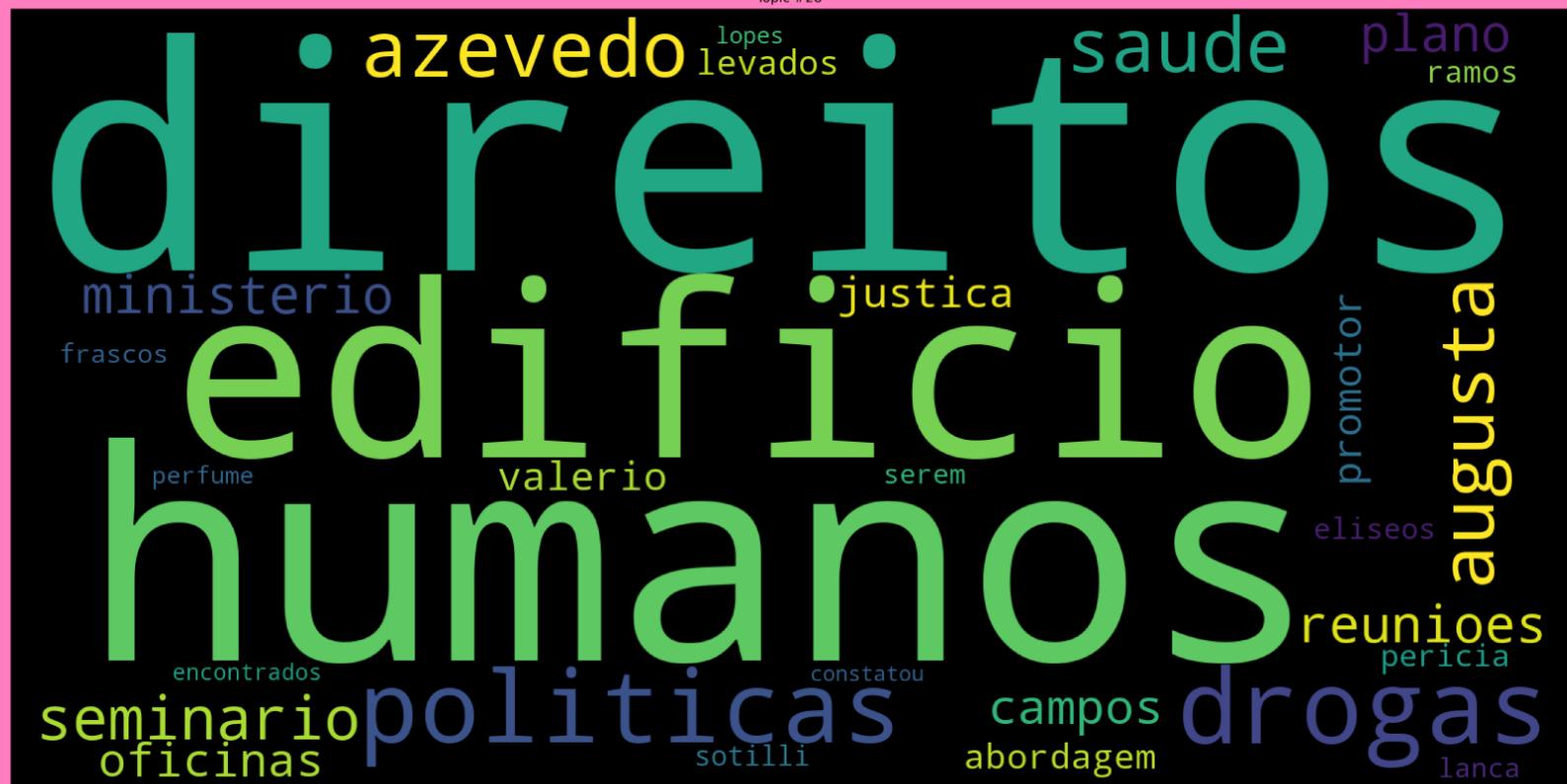
WordClouds

SMSU

Haddad  
(2013-2016)



obs: Desarmamento e DH



Topic #2

integrada postos crime experiencias antes posto meses gestao ano desarmamento

campanha base

comite

arma

nao

paz

instituto

casa

justica

pesquisa

fogo

politica

instalacoes

informacao

municoes

gestao

meses

ano

voce

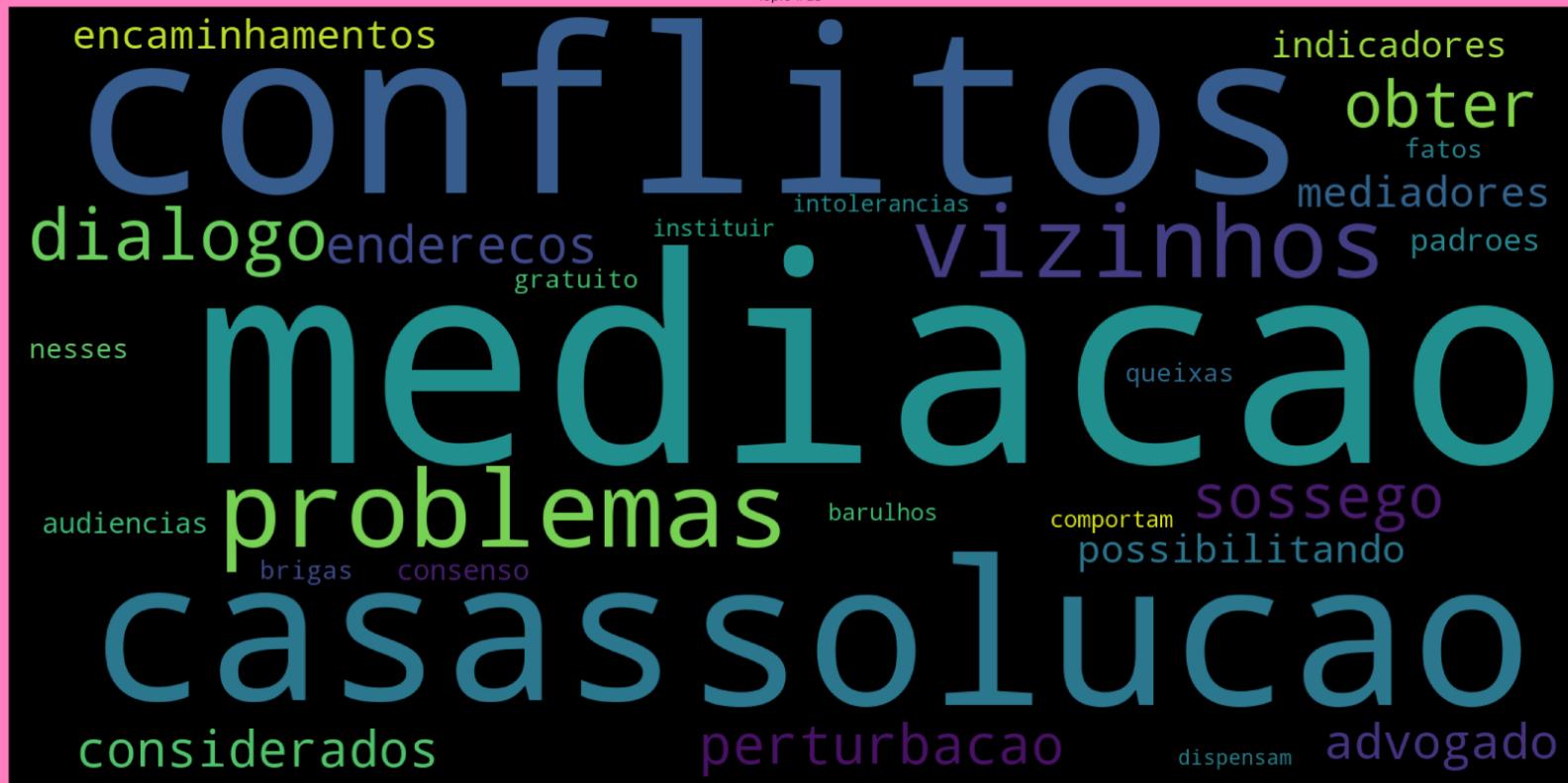
ter

qualquer

desarmamento

Violencia  
casos  
situacao

sul social novo  
ir manha santos  
inspetoria nova parelheiros  
criancas guarda apresentacao defesa  
ambiental  
programa regional vila regiao disse  
danos animais zona haddad.  
populacao gcm outubro protecao  
criancas guarda apresentacao defesa  
ambiental  
programa regional vila regiao disse  
danos animais zona haddad.  
primeira meio



Topic #10



The word cloud is centered around the word 'subprefeitura' in large green font. Other prominent words include 'comercio' (blue), 'regiao' (purple), 'metropolitana' (purple), 'guarda' (yellow), 'irregular' (green), 'avenida' (purple), 'zona' (yellow), 'irregulares' (teal), 'civil' (light green), 'civil' (light green), 'programa' (orange), 'rua' (light green), 'mil' (light green), 'civil' (light green), 'guarda' (yellow), 'bairro' (orange), 'ambulante' (purple), 'paulista' (purple), 'encaminhados' (orange), 'gcm' (green), 'palmares' (purple), 'guarda' (yellow), 'bairro' (orange), 'apreende' (orange), 'items' (orange), 'espaco' (purple), 'publico' (purple), 'viaturas' (purple), 'irregular' (green), 'controle' (blue), 'fiscalizacao' (blue), 'accao' (blue), and 'produtos' (green).

azevedo lopes levados saude plano  
ministerio justica ramos  
rascos promotor  
perfume eliseos  
valerio reunioes  
serem pericia  
encontrados Augusta  
eminario oficinas sotilli constatou  
politicas campos abordagem  
drogas lanca

# // Visualizações //

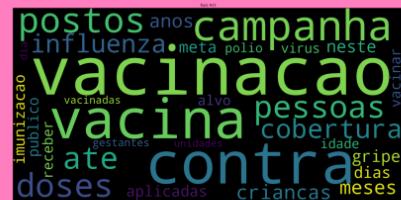
comparação gestão Haddad (2013-2016) e Dória-Covas (2017-2020)

#1

WordClouds

Saúde

## GERAL (2003-2020)



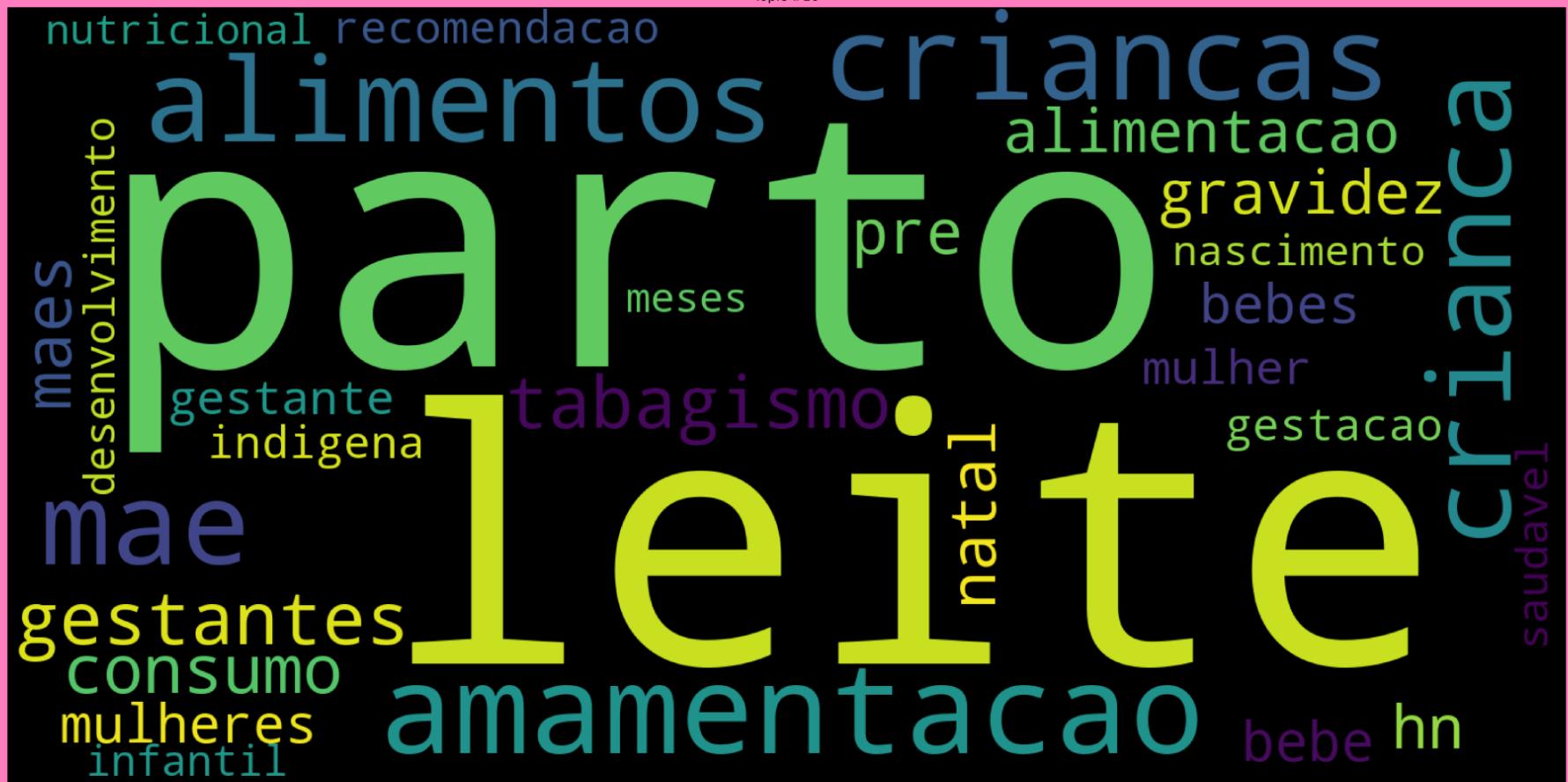
obs: Independente de governo: prevenção, idoso, vacina, gestantes, hiv e carnaval, dengue

A word cloud centered around the theme of women's health, featuring large words like "mulheres", "cancer", and "transmissíveis" surrounded by smaller related terms.

The words include:

- sangue
- precoce
- exame
- sexualmente
- mulheres
- hpv
- exposicao
- masculino
- coleta
- papanicolau
- contra
- sexocampanha
- cancer
- uterio
- anos
- bancos
- rapida
- exames
- pele
- ferias
- prevencao
- instituto
- doacao
- colo
- transmissíveis
- feminino
- realizacao
- diagnostico
- mes







# // Visualizações //

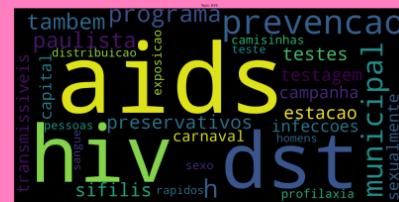
comparação gestão Haddad (2013-2016) e Dória-Covas (2017-2020)

#1

WordClouds

Saúde

Dória-Covas  
(2017-2020)



obs: Independente de governo: prevenção, idoso, vacina, gestantes, hiv e carnaval, dengue

+ Corujão

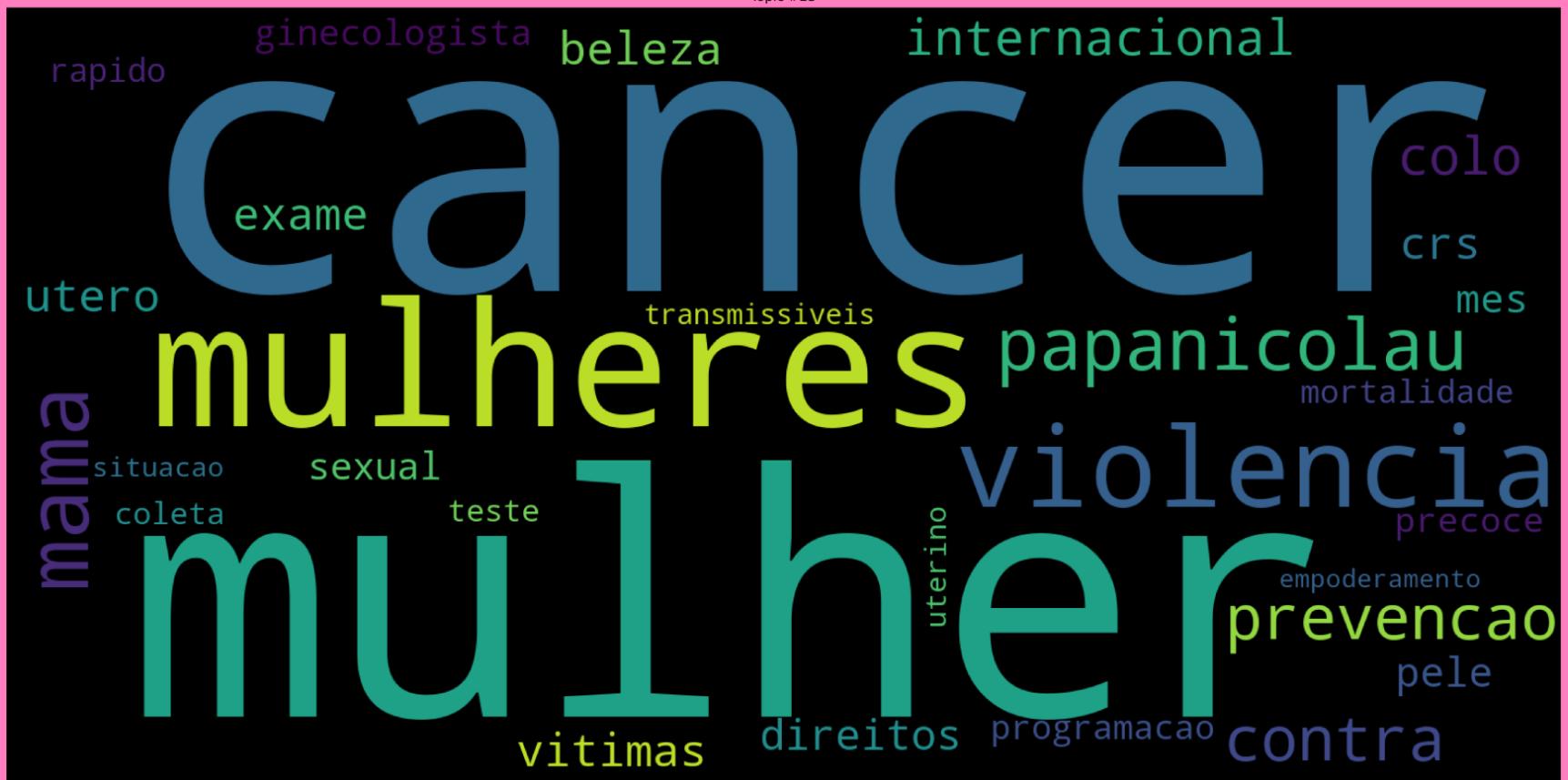
A word cloud centered around the theme of yellow fever vaccination and its related concepts. The words are color-coded by category:

- Yellow words:** febre.amarela, vacinação, ate, neste.
- Blue words:** contra, virus, casos, pessoas, doença, dia, sms, anos, dias, dose, h, capital.
- Green words:** arboviroses, doses, meses, saude, campanha, ano, cobertura, populacao.
- Purple words:** unidades, postos.
- Other words:** criancas, nao.

A word cloud centered around the words AIDS, HIV, and DST. Other words include prevention, condom, test, campaign, station, infections, men, sex, rapid, sifilis, and municipal.

The words are arranged as follows:

- Top row: tambem, programa, prevencao
- Middle row: paulista, camisinhas, teste, testes, testagem, campanha
- Second row from bottom: capital, exposicao, preservativos, estacao, infeccoes, homens
- Third row from bottom: distribuicao, carnaval, sexo, dst, profilaxia
- Bottom row: transmissiveis, sanguinas, pessoas, sifilis, rapidos, h, municipal, sexualmente



agendado tomografias • zerar milhao  
terminal libanes "corujao" • bp  
exames saude" beneficencia paciente  
hospital Sirio espera extras convenios fila  
maia variar edmundo conveniados portuguesa  
marcar "corujao"  
conveniados procedimentos agendados reavaliacao  
saudé "corujao"  
bp  
corujao  
agendamentos

A word cloud centered around the topic of dengue prevention and mosquito control. The most prominent words are 'dengue' (large purple), 'mosquito' (large green), 'aedes' (medium green), 'agentes' (large yellow), and 'controle' (medium teal). Other visible words include ' Zika' (purple), 'aedes aegypti' (blue), 'mosquitos' (green), 'criadouros' (green), 'profissionais' (yellow), 'controle' (teal), 'combate' (teal), 'agentes' (yellow), 'acao' (yellow), 'penha' (yellow), 'casa' (blue), 'sapopemba' (blue), 'prevencao' (blue), 'linhas' (green), 'vigilancia' (green), 'sudeste' (purple), 'aricanduva' (teal), 'prudente' (blue), 'visita' (green), 'ativa' (green), 'agua' (purple), 'proliferacao' (purple), 'mooca' (blue), and 'chikungunya' (purple).

# // Visualizações //

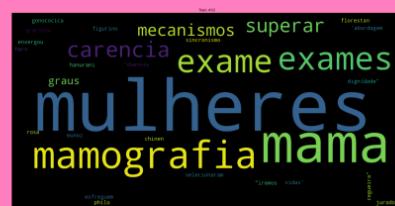
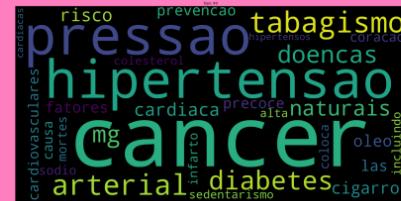
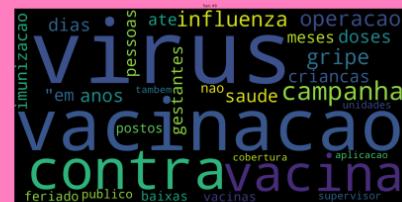
comparação gestão Haddad (2013-2016) e Dória-Covas (2017-2020)

#1

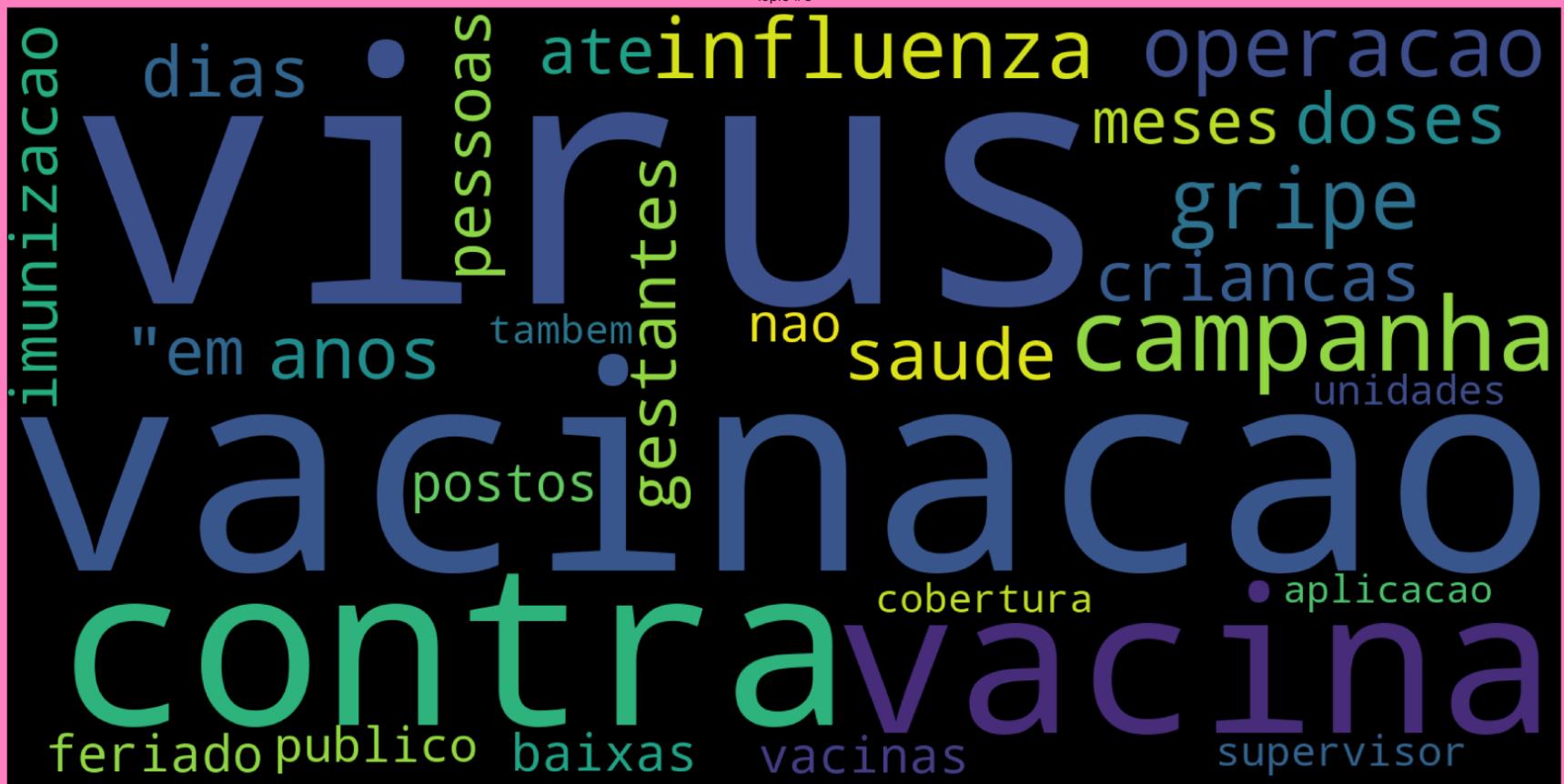
WordClouds

SMADS

## Haddad (2013-2016)



obs: Independente de governo: prevenção, idoso, vacina, gestantes, hiv e carnaval, dengue



gonococica  
gracioso  
enxergou  
haro  
figurino  
hanurani  
"doencas"  
sincronismo  
mecanismos  
superar  
carenzia  
exame exames  
dignidade"  
graus  
rosa  
munoz  
chinen  
selecionaram  
"iremos vidas"  
cegueira"  
jurado  
mulheres  
mamografia mama  
esfreguem  
philo

gestantes  
mae  
maes  
pais  
primeira  
recem  
sinasc  
criancas  
crianca  
criancabébe  
bebés  
vivos  
melhorou  
agressao  
coisa  
gestante  
sobre peso  
morcegos  
partos  
infancia  
materna  
sinas  
primeira  
recem  
gestante  
sobre peso  
morcegos  
partos  
infancia  
materna

Topic #0

cardiacas  
cardiovasculares  
causa  
sodio  
mg  
arterial  
risco  
fatores  
colesterol  
causa  
mortes  
sodio  
arterial  
prevencao  
hipertensos  
infarto  
cardiaca  
precoce  
alta  
naturais  
doencas  
coracao  
tabagismo  
coracao  
doencas  
naturais  
cigarro  
cigarro  
incluso

acao bairro centro ceu serao animais mateus populacao jardim orientacoes praca h comprovarante visita butanta havera que limpo parque vilar campor integrada regiao dias santa sera comunidade dia iii acao bairro

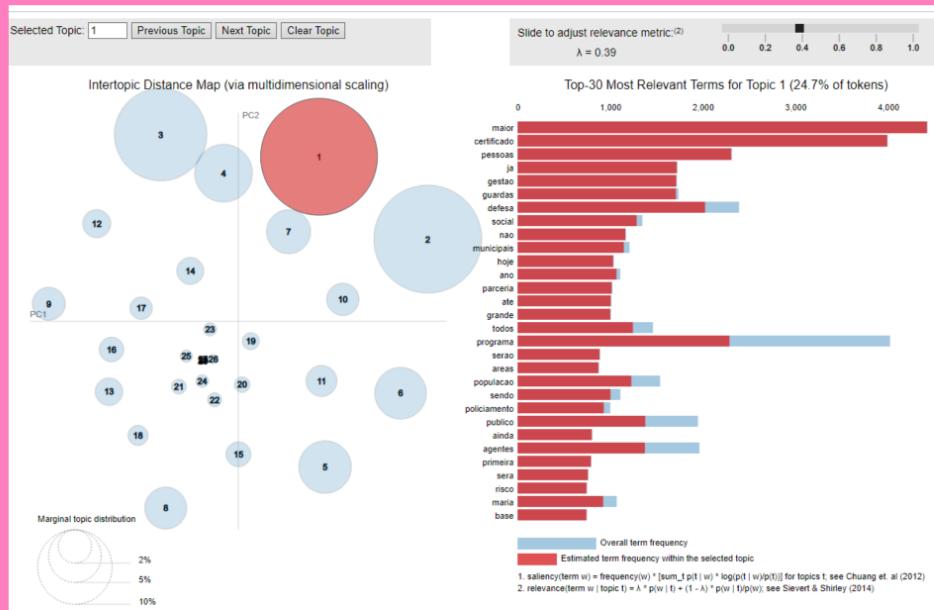
//

# Visualizações //

comparação gestão Haddad (2013-2016) e Dória-Covas (2017-2020)

#2

pyLDAvis;



"where  $\phi_{kw}$  is the probability of word w in topic k and  $\phi_{kw}/p_{kw}$  is the lift in term's probability within a topic to its marginal probability across the entire corpus (this helps discards globally frequent terms). A lower  $\lambda$  gives more importance to the second term ( $\phi_{kw}/p_{kw}$ ), which gives more importance to topic exclusivity. We can again use pyLDAvis for this. For instance, when lowering  $\lambda$  to 0.6, we can see that topic 13 ranked terms that are even more relevant to the topic of phones."

<https://towardsdatascience.com/6-tips-to-optimize-an-nlp-topic-model-for-interpretability-20742f3047e2>

file:///C:/Users/user/2.%20GIT\_PROJECTS/Desafios/Projeto%205%20-%20NLP%20Secretarias/Visualization/seguranca\_urbana.html#topic=1&lambda=0.39&term=

//

# Visualizações //

comparação gestão Haddad (2013-2016) e Dória-Covas (2017-2020)

E se buscamos o documento mais relevante cruzando assistência social e saúde?

#3

Retorno de documentos mais importantes por tópico no corpus e de cruzamento entre tópicos;

**GCM encaminha homem que tentou suicídio a atendimento social** 5/01/2010 Texto: Gláucia Arboleya A Guarda Civil Metropolitana encaminhou, na última quarta-feira (13/1) por volta das 14h30, um homem que tentou se suicidar no Viaduto do Chá para o Atendimento Médico Ambulatorial (AMA), na Sé. O homem é natural da cidade de Paulo Afonso, Bahia, e estava em situação de risco há 30 dias. A Inspetoria do Gabinete do Prefeito visualizou a vítima do lado externo da grade do viaduto e foi até o local. Após conversa com o homem, ele aceitou a sair da posição de risco do viaduto e acompanhar os guardas civis metropolitanos para a AMA. Ele estava acompanhado por um rapaz de 18 anos, também em situação vulnerável há 30 dias, natural de Val Paraíso – interior de São Paulo. Ambos foram encaminhados para atendimento médico e, posteriormente, ao atendimento social para acolhimento em albergue da região até que seja viabilizado o retorno para as suas cidades de origem. No ano passado a GCM também atendeu um homem que tentou se suicidar da passarela do DETRAN, no Ibirapuera. O encaminhamento de pessoa em situação de risco é um dos programas prioritários da GCM.

//

# Visualizações //

comparação gestão Haddad (2013-2016) e Dória-Covas (2017-2020)

E se buscamos o documento mais  
relevante cruzando segurança,  
ass. social e saúde?

GCMs mediadores de conflitos participam da comemoração dos 10 anos da atuação do MP em Justiça Terapêutica 1/02/2013 O Ministério Público do Estado de São Paulo, por meio da Promotoria de Justiça Criminal do Fórum de Santana, lançou, na manhã desta terça-feira (19/02), o vídeo "Justiça Terapêutica: é possível fazer!", um documentário de 23 minutos produzido com apoio da Procuradoria-Geral de Justiça e da Associação Paulista do Ministério Público. O lançamento aconteceu no Auditório da Ordem dos Advogados do Brasil (OAB) de Santana e contou com a presença do Procurador-Geral de Justiça, Márcio Fernando Elias Rosa, que presidiu a abertura da solenidade. Curiosos para aprender, conhecer de perto o funcionamento da Justiça Terapêutica e encontrar pontos comuns que podem ser incorporados no processo de Mediação de Conflitos da GCM, mediadores das sete unidades do Comando Operacional Norte participaram do evento. Na ocasião, puderam verificar os bons resultados dos mais de 1.300 atendimentos realizados pelo órgão nos últimos 10 anos e conhecer mais um setor que pode ser indicado àqueles que comparecem nas unidades da GCM em busca de solução para seus problemas. A Justiça Terapêutica é um modelo penal no qual o consumidor de drogas ilegais escolhe entre cumprir uma pena ou receber tratamento de saúde "Essa oportunidade foi muito importante, pois os mediadores de conflito puderam trocar experiências, criar contatos para aprofundarem o tema e formar parcerias que possam contribuir com a melhoria dos processos de mediação promovidos pela GCM", ressaltou o Comandante Operacional Norte Marcos Bazzana Delgado. Também estiveram presentes na solenidade Juízes, Promotores de Justiça, Delegados de Polícia, o Inspetor Renato Sampaio e os Guardas Civis Mediadores de Conflito Landuar Alencar Filho, da Inspetoria Regional de Santana; Renato Rodrigues de Oliveira, Inspetoria Pirituba/Jaraguá; Marta Inocêncio, Inspetoria Regional Vila Maria/Guilherme; Nerciana da Silva Santos, Inspetoria Regional Jaçanã/Tremembé; Mauricio Mendonça Villar, Inspetoria Regional Casa Verde e Vera Lúcia da Silva Vignoto, Inspetoria Regional Freguesia do Ó.

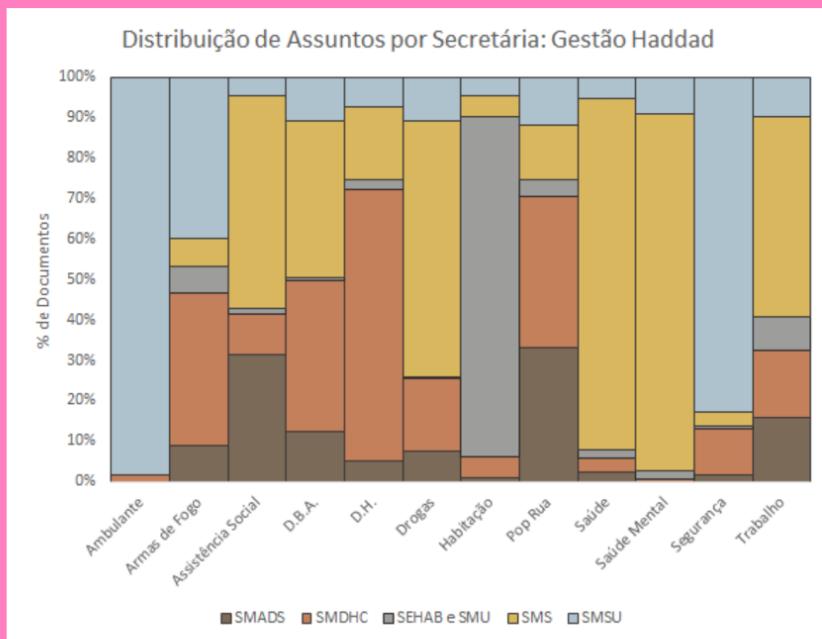
//

# Visualizações //

comparação gestão Haddad (2013-2016) e Dória-Covas (2017-2020)

#4

Cruzamento de assuntos criados por nós com tópicos no corpus por gestão;



Distribuição de Assuntos por Secretaria: Gestão Haddad

	SMADS	SMDHC	SEHAB e SMU	SMS	SMSU
<b>Ambulante</b>	0%	2%	0%	0%	98% <b>1%</b>
<b>Armas de Fogo</b>	9%	38%	7%	7%	40% <b>0%</b>
<b>Assistência Social</b>	31%	10%	2%	53%	5% <b>9%</b>
<b>D.B.A.</b>	12%	37%	1%	39%	11% <b>1%</b>
<b>D.H.</b>	5%	67%	3%	18%	7% <b>12%</b>
<b>Drogas</b>	7%	18%	0%	64%	11% <b>2%</b>
<b>Habitação</b>	1%	5%	84%	5%	5% <b>8%</b>
<b>Mananciais</b>	0%	0%	98%	2%	0% <b>0%</b>
<b>Pop Rua</b>	33%	37%	4%	14%	12% <b>1%</b>
<b>Saúde</b>	2%	3%	2%	87%	5% <b>25%</b>
<b>Saúde Mental</b>	0%	1%	2%	88%	9% <b>1%</b>
<b>Segurança</b>	2%	11%	1%	4%	83% <b>11%</b>
<b>Trabalho</b>	16%	17%	8%	50%	10% <b>15%</b>
<b>Viaduto</b>	0%	6%	1%	23%	70% <b>1%</b>
	<b>41%</b>	<b>21%</b>	<b>19%</b>	<b>11%</b>	<b>8%</b>

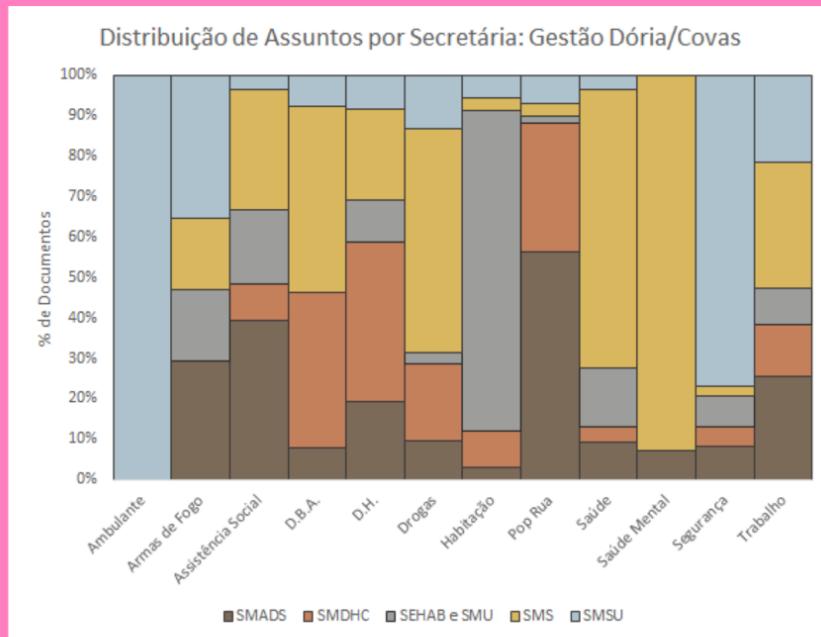
//

# Visualizações //

comparação gestão Haddad (2013-2016) e Dória-Covas (2017-2020)

#4

Cruzamento de assuntos criados por nós com tópicos no corpus por gestão;



Distribuição de Assuntos por Secretaria: Gestão Dória/Covas

	SMADS	SMDHC	SEHAB e SMU	SMS	SMSU
<b>Ambulante</b>	0%	0%	0%	0%	100%
<b>Armas de Fogo</b>	29%	0%	18%	18%	35%
<b>Assistência Social</b>	40%	9%	18%	30%	3%
<b>D.B.A.</b>	8%	38%	0%	46%	8%
<b>D.H.</b>	19%	40%	10%	22%	8%
<b>Drogas</b>	10%	19%	3%	55%	13%
<b>Habitação</b>	3%	9%	79%	3%	5%
<b>Mananciais</b>	0%	0%	100%	0%	0%
<b>Pop Rua</b>	56%	32%	1%	3%	7%
<b>Saúde</b>	9%	4%	14%	69%	3%
<b>Saúde Mental</b>	7%	0%	0%	93%	0%
<b>Segurança</b>	8%	5%	8%	2%	77%
<b>Trabalho</b>	25%	13%	9%	31%	21%
<b>Viaduto</b>	15%	0%	10%	1%	74%
	<b>30%</b>	<b>22%</b>	<b>19%</b>	<b>18%</b>	<b>11%</b>

# Modelo de Tópicos e a Prefeitura de SP

// Topic  
Model //

//  
Resultados //

// Workflow //

// Futuro //

Guilherme G. Nicolau /  
Ironhack DA Student  
21/05/2021

# // Futuro //



01      02      03      04      05      06

// ...e Deploy! //

- Automatizar (mais) nossas pipelines, transformando as funções em classes;

- ...em um comando!

# A maldição da dimensionalidade

Encontrar o melhor  
número de tópicos para  
representar o seu  
Corpus não é fácil...

**Solução:** incorporar e automatizar  
os hiperparâmetros através de  
'topic coherence';

# Informational Retrieval

- Automatizar o retorno de documentos mais importantes por tópico;

# Structured Topic Model

- Comparação entre tópicos no corpus, através de Structured Topic Model (biblioteca de R) em um heatmap;

Com isso podemos ir um nível adiante na interpretação semântica dos nossos sujeitos; por exemplo: o que significa quando no mesmo Corpus, um tópico de desenvolvimento é bastante relacionado com direitos humanos?

# NER

Gostaríamos de incorporar um reconhecimento de entidade; para descobrir quais são as pessoa, lugares e instituições mais importantes para cada gestão

Existe um ótimo modelo em redes neurais já preparado em português que tem como base o BERT do Google. Se chama BERTimbau e está incorporado na biblioteca Transformers do HuggingFace

## Sentiment Analysis

*O que significa ser mais importante? O sujeito considera isso positivo ou negativo?*

Podemos incorporar análise de sentimentos e isso é relativamente simples;

Mas pretendemos investigar outras alternativas para identificar a intenção do sujeito que não dependam de uma classificação supervisionada e prévia sobre o significado das palavras (como é a análise de sentimentos)