

# 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 //

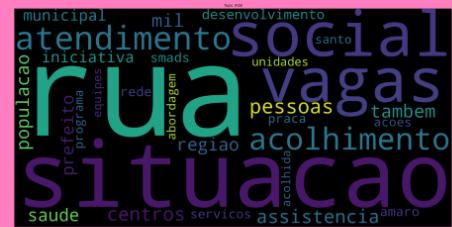
comparação gestão Haddad (2013-2016) e Dória-Covas (2017-2020)

#1

## WordClouds

SMADS

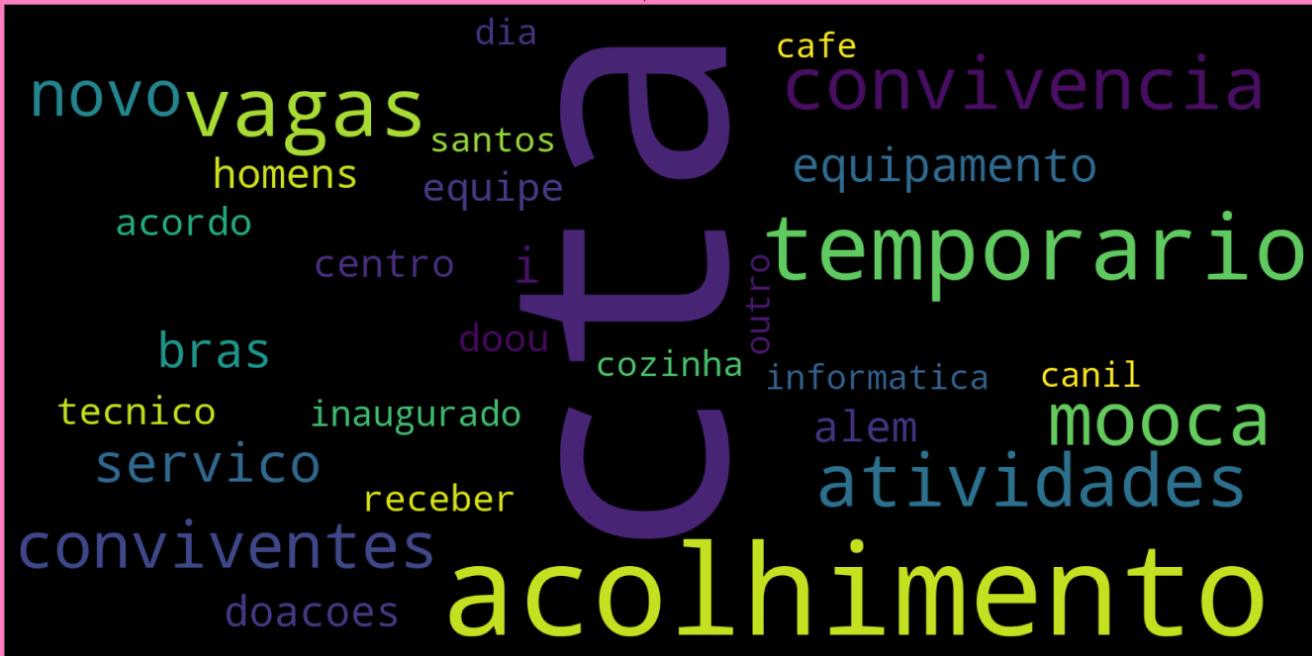
Dória-Covas  
(2017-2020)

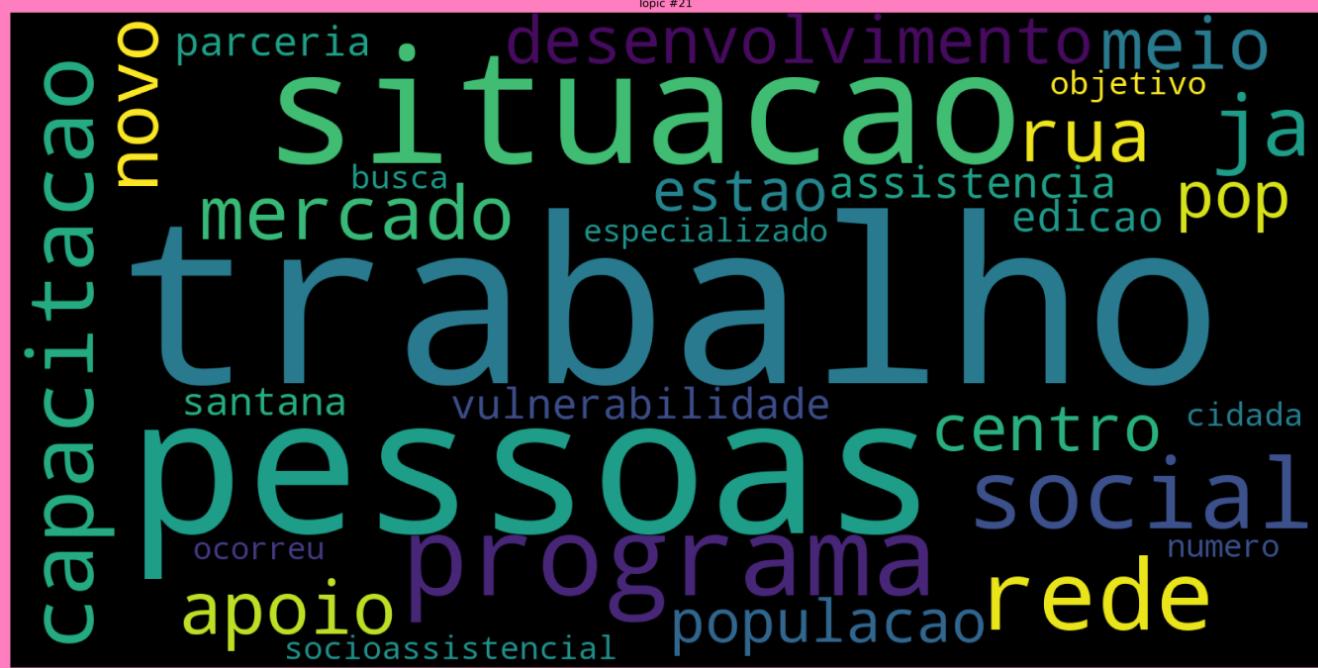


Topic #28

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

Topic #13





Topic #5

projeto profissionais gestao  
acoes politicas sobre alimentos secretario protecao  
**assistencia**  
desenvolvimento publico  
servicos encontro infancia **social** unico espaco publica  
chefe smads municipa  
primeira sistema sociais capacitação

ficando  
gostei  
bela  
beleza  
talentos  
paulistano  
externas  
obstaculo  
simpatia  
instalado  
raquel  
todos"  
pele  
domingues  
ganham  
valente  
cuidados  
vaidade  
"uma"  
parabens  
compreenderam  
margarida  
plantas  
mensagem  
fantasia  
cobertura  
idosas  
mulheres"  
beneficiarias  
mostraram

# // Visualizações //

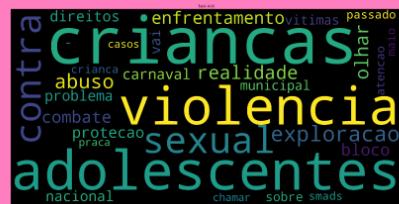
comparação gestão Haddad (2013-2016) e Dória-Covas (2017-2020)

#1

WordClouds

SMADS

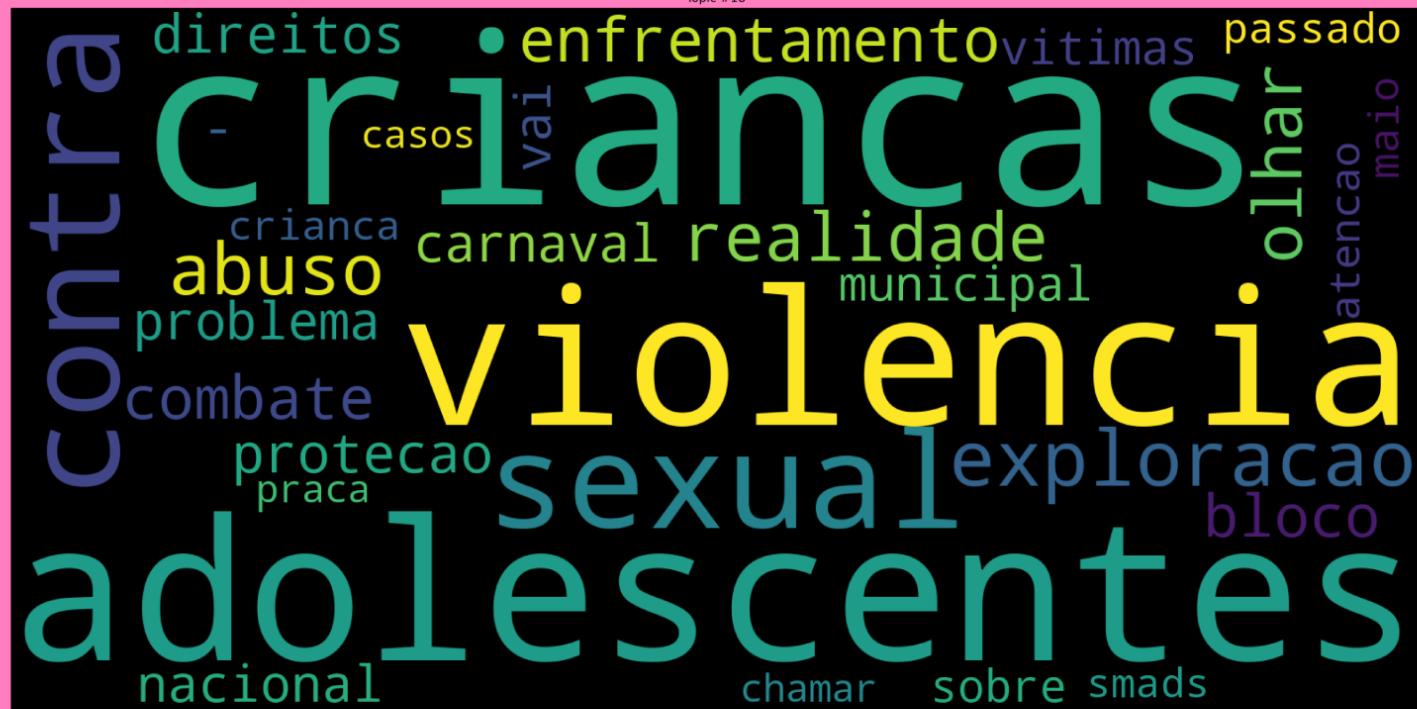
Haddad  
(2013-2016)



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  
Social  
assistencia





Topic #20

especializados  
comprometido  
carater  
maio  
atendimento  
william  
federativo  
rua  
vai  
jantar  
abordagem  
prestados  
intensifica  
civil  
desprotecoes  
vagas  
abrigos  
banho  
discutidos  
central  
situacao  
encaminhamento  
centros  
acolhida  
oc  
reciclageis

**baixas**

**operacao**

**temperaturas**



# // Visualizações //

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

#1

WordClouds

SMADS

Dória-Covas  
(2017-2020)

# // Visualizações //

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

#1

WordClouds

SMADS

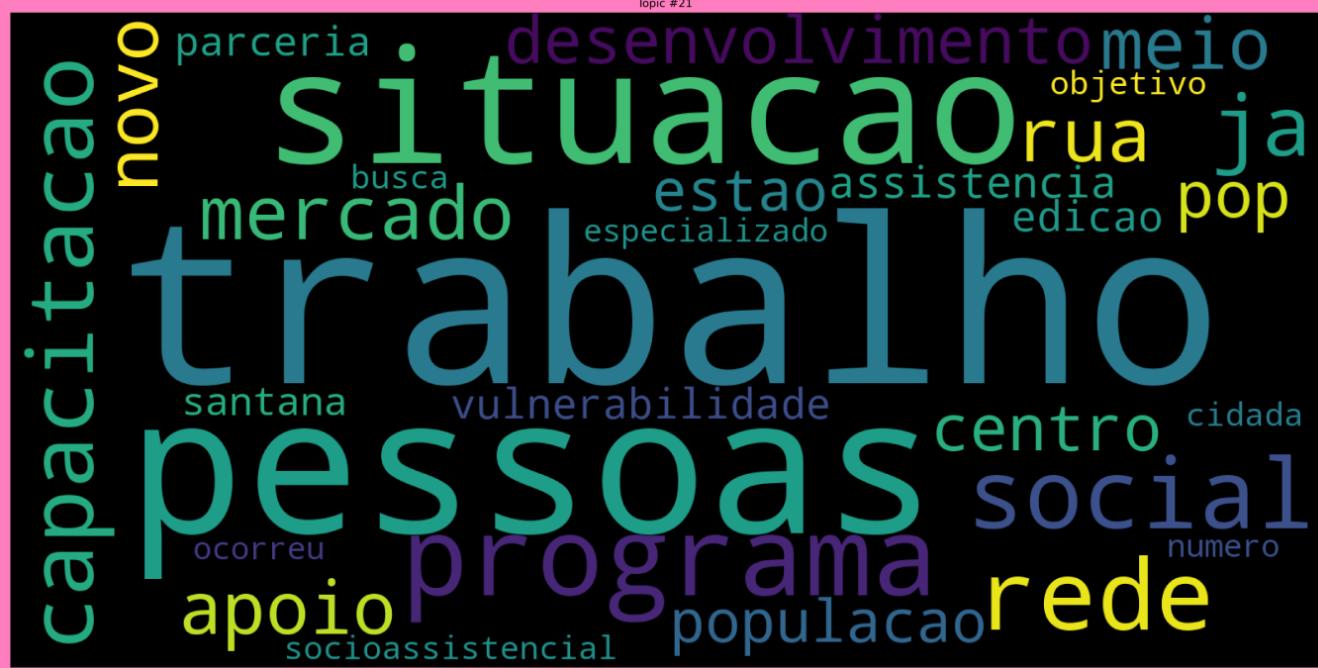
Dória-Covas  
(2017-2020)



Topic #28

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





Topic #5

projeto profissionais gestao  
acoes politicas sobre alimentos secretario protecao  
**assistencia**  
desenvolvimento publico  
servicos encontro infancia **social** unico espaço publica  
chefe **smads** educacao programas crianca  
primeira sistema sociais capacitacao  
municipal

Topic #8

# // Visualizações //

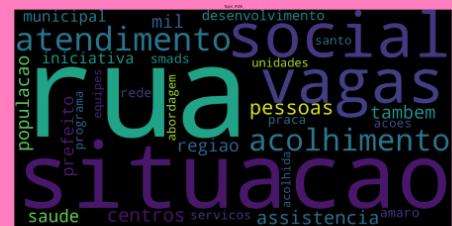
comparação gestão Haddad (2013-2016) e Dória-Covas (2017-2020)

#1

## WordClouds

SMADS

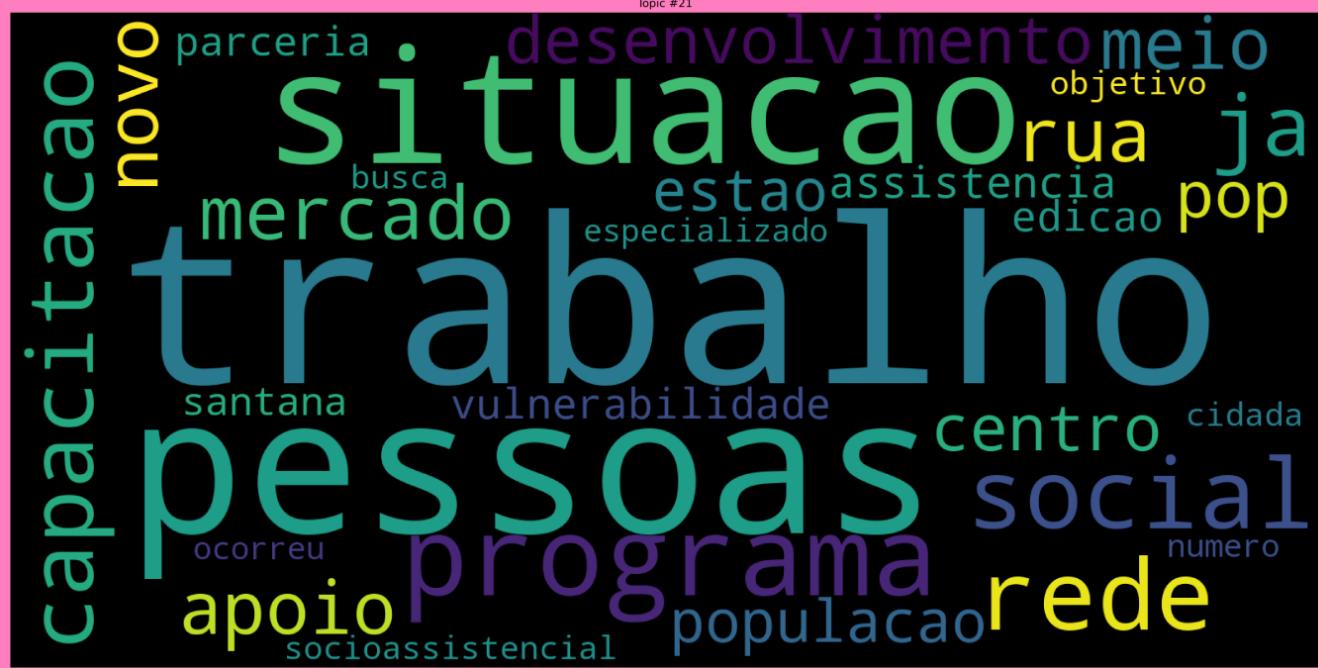
Dória-Covas  
(2017-2020)



Topic #28

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





Topic #5

projeto acoes politicas sobre alimentos profissionais gestao secretario protecao  
assistencia desenvolvimento publico  
servicos encontro infancia social unico espaco publica  
chefe smads municipal  
primeira sistema sociais capacitação

ficando  
gostei  
pele  
domingues  
simpatia  
instalado  
raquel  
todos"  
deveria  
parabens  
compreenderam  
margarida  
mostraram  
cuidados  
vaidade  
"uma"  
valente  
ganham  
obstaculo  
externas  
cobertura  
mensagem  
fantasia  
idosas  
mulheres"  
beneficiarias  
plantas  
**beleza**  
**talentos**

# // Visualizações //

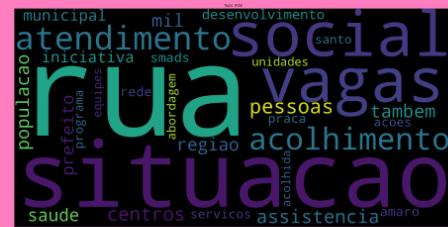
comparação gestão Haddad (2013-2016) e Dória-Covas (2017-2020)

#1

WordClouds

SMADS

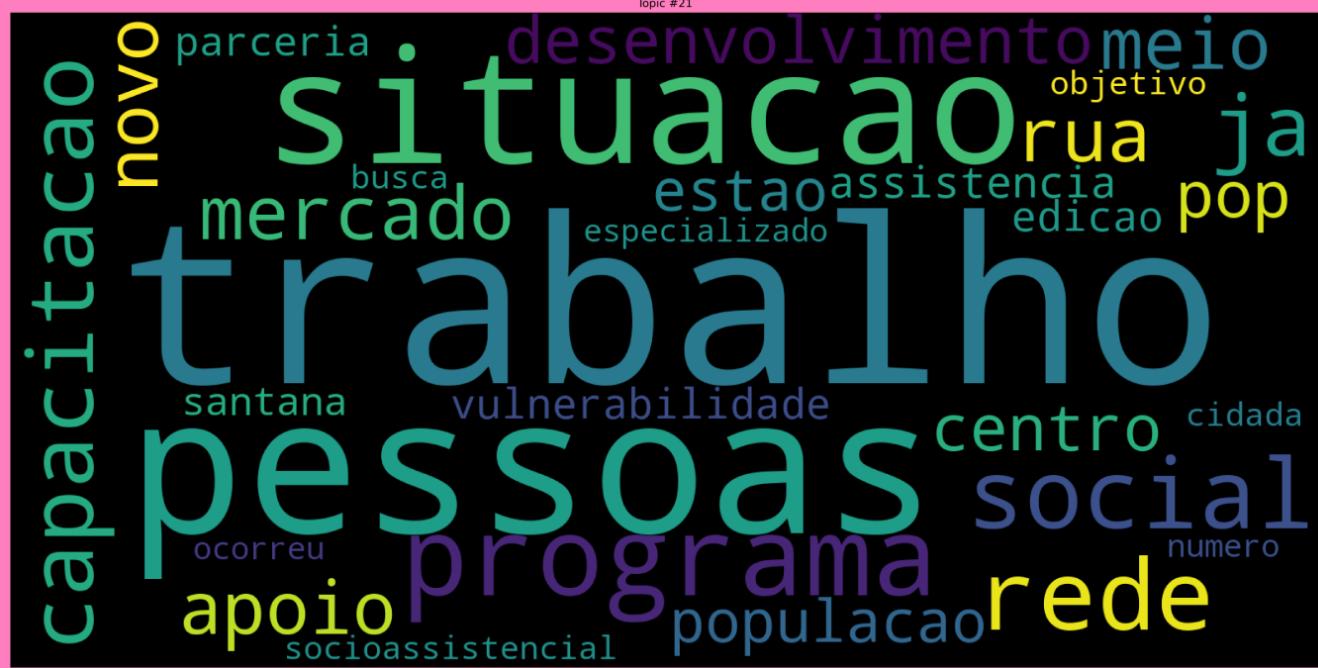
Dória-Covas  
(2017-2020)



Topic #28

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





Topic #5

projeto profissionais gestao  
acoes politicas sobre alimentos secretario protecao  
**assistencia**  
desenvolvimento publico  
servicos encontro infancia **social** unico espaco publica  
chefe **smads** educacao programas crianca  
primeira sistema sociais capacitacao  
**municipal**

ficando  
paulistano  
**beleza**  
talentos  
gostei  
obstaculo  
simpatia  
instalado  
raquel  
todos"  
deveria  
parabens  
compreenderam  
margarida  
mostraram  
cuidados  
vaidade  
"uma"  
valente  
ganham  
planta's  
mensagem  
fantasia  
cobertura  
ídoras  
mulheres"  
beneficiarias  
externas

# // Visualizações //

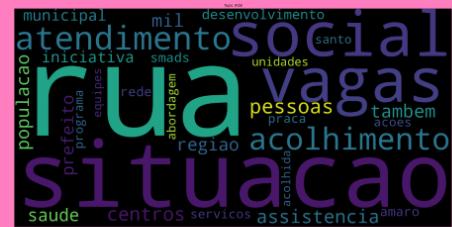
comparação gestão Haddad (2013-2016) e Dória-Covas (2017-2020)

#1

WordClouds

SMADS

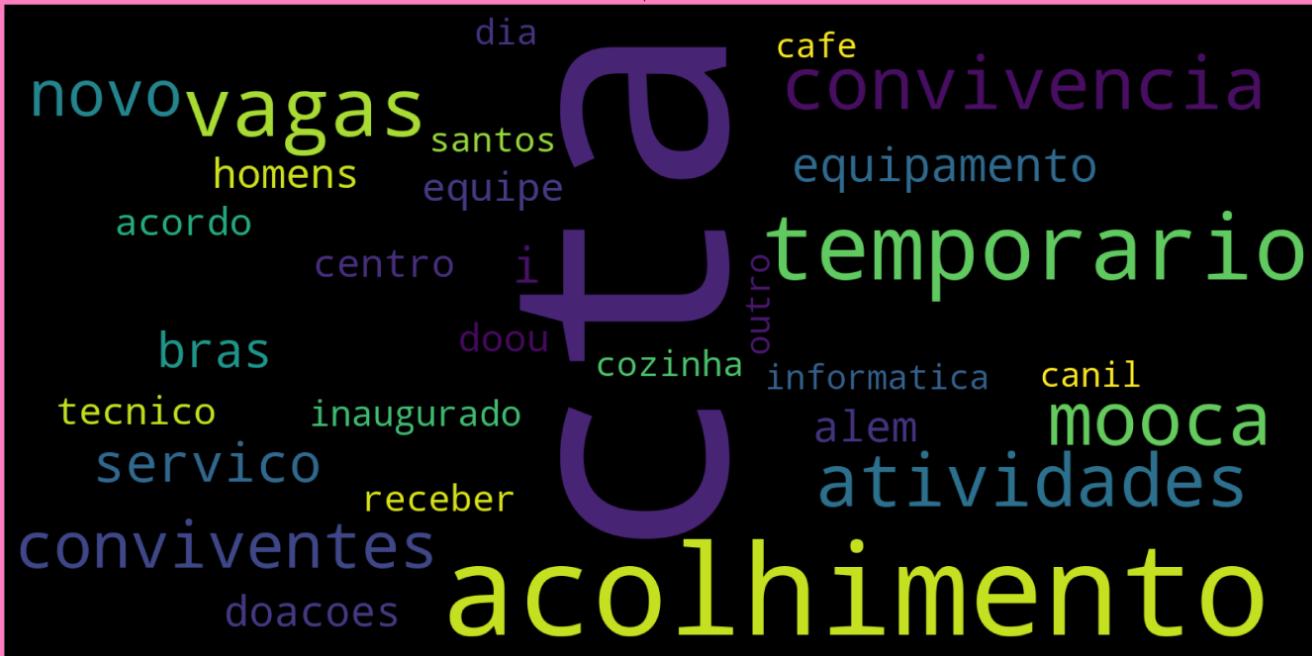
Dória-Covas  
(2017-2020)

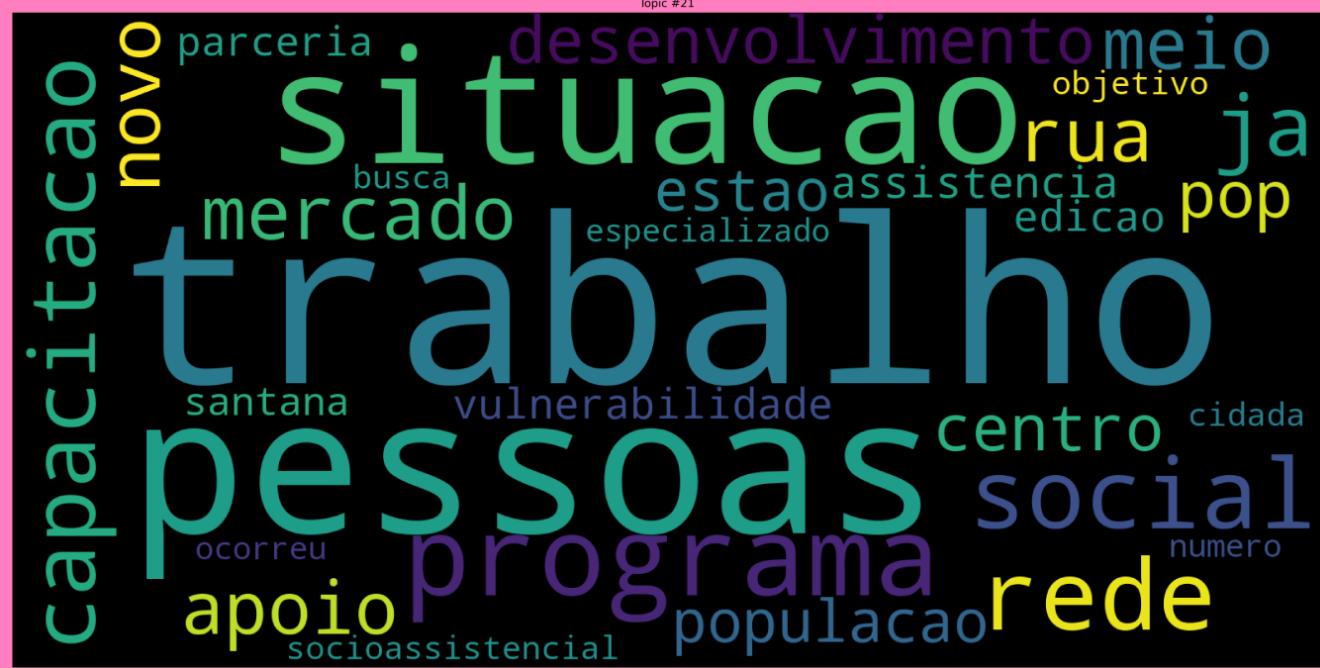


Topic #28

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

Topic #13





Topic #5

projeto profissionais gestao  
acoes politicas sobre alimentos secretario protecao  
**assistencia**  
desenvolvimento publico  
servicos encontro infancia **social** unico espaço publica  
chefe smads municipa  
primeira sistema sociais capacitação

A word cloud visualization titled "Topic #8" featuring large, bold words in green, teal, and yellow, set against a black background. The words are arranged in a roughly circular pattern. The central words are "beleza" and "talentos". Other prominent words include "gostei", "fiquei", "margarida", "obstaculo", "valente", "cuidados", "vaidade", "uma", "ganham", "raquel", "toda", "instalado", "simpatia", "pele", "domingues", "instalado", "raquel", "toda", "deveria", "parabens", "compreenderam", "margarida", "mostraram", "beneficiarias", "mulheres", "idosas", "fantasia", "mensagem", "cobertura", "externas", "paulistano", and "plantas". Smaller words are scattered around the perimeter of the main cluster.

# Modelo de Tópicos e a Prefeitura de SP

// Topic  
Model //

//  
Resultados //

// Workflow //

// Futuro //

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