Extração de informação sobre notícias falsas e verdadeiras

Thiago Schinda Bubniak
Waine Oliveira Junior

Tema

Compreender quais os temas de fake news e notícias verdadeiras mais relevantes em determinado período e comparar com eventos relacionados no período

Perguntas e Hipóteses

Perguntas de pesquisa:

- Quais foram os tópicos mais comentados nas notícias?
- Como esses tópicos evoluíram com o passar do tempo?
- Qual a relação desses tópicos com eventos políticos?

Hipóteses

- Os tópicos das notícias evoluíram conforme eventos políticos (e.g. perto do julgamento do ex-presidente, as notícias falsas sobre ele aumentaram)
- Os tópicos das notícias falsas e verdadeiras, num mesmo intervalo de tempo, tendem a ser os mesmos

Dados utilizados

- corpus Fake.br
- 3600 notícias falsas + 3600 notícias verdadeiras

Processamentos aplicados

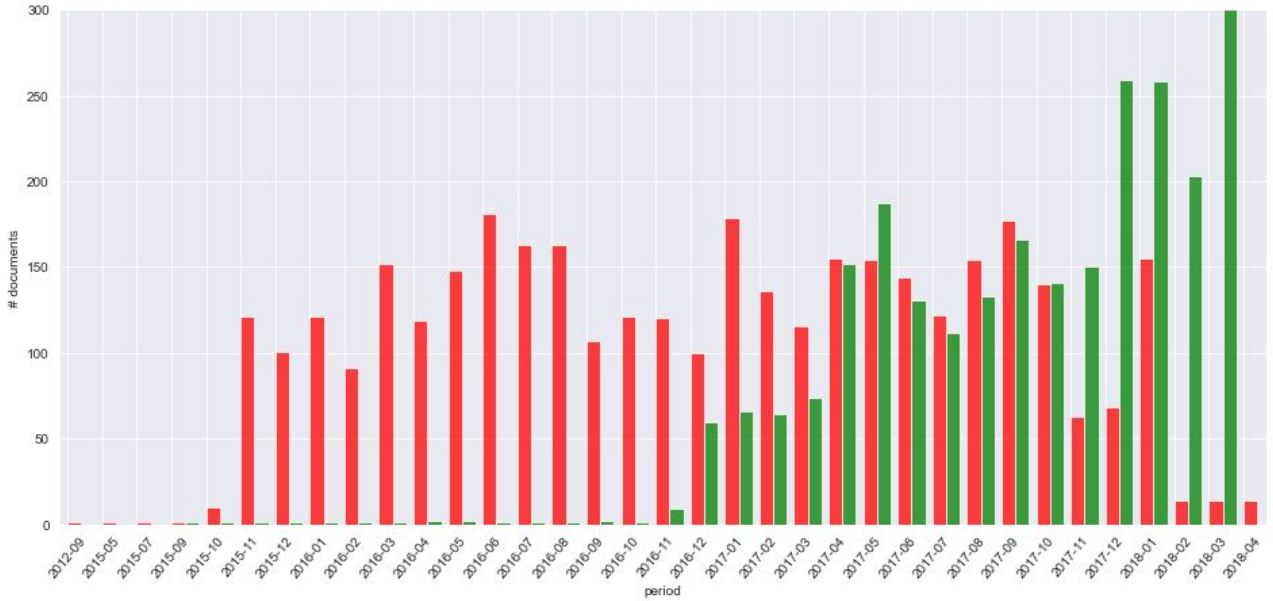
- Remoção de acentos e pontuação
- Remoção de stopwords
- Normalização de texto usando NFKD
- Tokenization
- Lemmatization
- Criação de matriz termo-documento

Análise exploratória

politica	4180
tv_celebridades	1544
sociedade_cotidiano	1276
ciencia_tecnologia	112
economia	44
religiao	44

link	label	
diariodobrasil.org	fake	3337
g1.globo.com	true	2300
politica.estadao.com.br	true	753
afolhabrasil.com.br	fake	174
internacional.estadao.com.br	true	114
cultura.estadao.com.br	true	95
www1.folha.uol.com.br	true	91
thejornalbrasil.com.br	fake	66
economia.estadao.com.br	true	51
brasil.estadao.com.br	true	46
esportes.estadao.com.br	true	30
alias.estadao.com.br	true	29
ceticismopolitico.com	fake	16
sao-paulo.estadao.com.br	true	16
saude.estadao.com.br	true	15
sustentabilidade.estadao.com.br	true	13
opiniao.estadao.com.br	true	11
topfivetv.com	fake	7
estadao.com.br	true	7
educacao.estadao.com.br	true	7
ciencia.estadao.com.br	true	5
viagem.estadao.com.br	true	5
emais.estadao.com.br	true	3
link.estadao.com.br	true	2
f5.folha.uol.com.br	true	2
territorioeldorado.limao.com.br	true	2
datafolha.folha.uol.com.br	true	1
blogdofred.blogfolha.uol.com.br	true	1
acervo.estadao.com.br	true	1

Análise exploratória



Análise exploratória

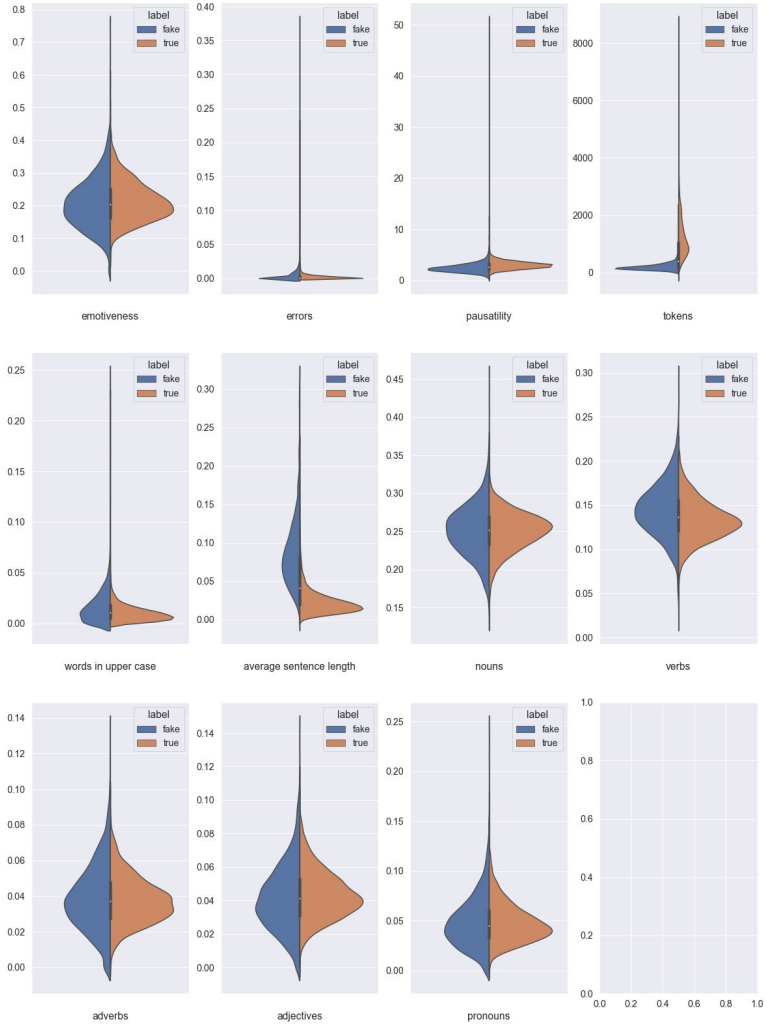
Top 100 Words in Fake News

Top 100 Words in True News

Análise exploratória

Maiores diferenças:

- Tamanho médio de sentenças
- Número de palavras (tokens)



Análise exploratória

```
Árvore de regressão (max_depth = 2)
F1-Score Train: 0.9455687369155616
F1-Score Test: 0.9437939110070257
|--- nouns <= 111.50
|   |--- nouns <= 86.50
|   |   |--- class: 0.0
|   |--- nouns >  86.50
|   |   |--- class: 0.0
|--- nouns >  111.50
|   |--- nouns <= 147.50
|   |   |--- class: 1.0
|   |--- nouns >  147.50
|   |   |--- class: 1.0
```

```
Árvore de regressão (max_depth = 2)
F1-Score Train: 0.8978178039487357
F1-Score Test: 0.8907913003239242
|--- average sentence length <= 0.04
|   |--- average sentence length <= 0.03
|   |   |--- class: 1.0
|   |--- average sentence length >  0.03
|   |   |--- class: 1.0
|--- average sentence length >  0.04
|   |--- average sentence length <= 0.06
|   |   |--- class: 0.0
|   |--- average sentence length >  0.06
|   |   |--- class: 0.0
```

```
Árvore de regressão (max_depth = 2)
F1-Score Train: 0.7437185929648241
F1-Score Test: 0.7492076906824424
|--- pausatility <= 2.20
|   |--- words in upper case <= 0.02
|   |   |--- class: 0.0
|   |--- words in upper case >  0.02
|   |   |--- class: 0.0
|--- pausatility >  2.20
|   |--- words in upper case <= 0.02
|   |   |--- class: 1.0
|   |--- words in upper case >  0.02
|   |   |--- class: 0.0
```

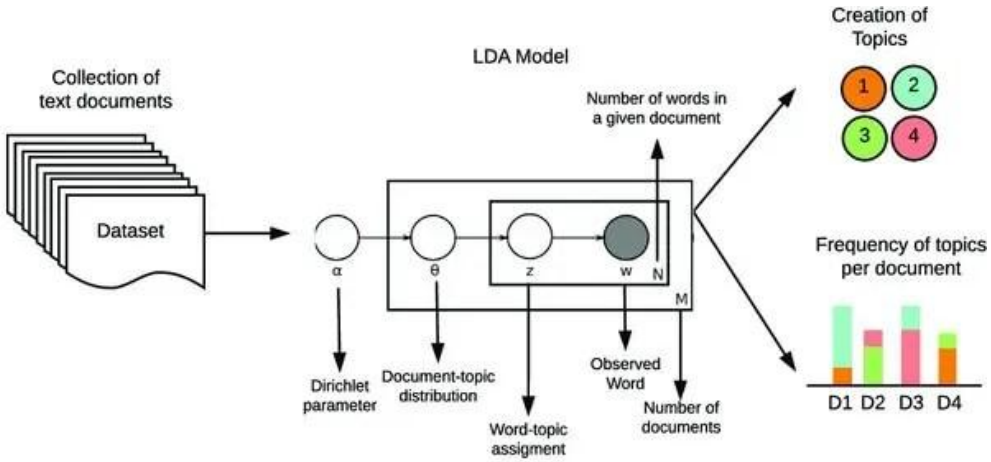
Topic Discovery - LDA

Entrada:

- Conjunto de documentos
- Número de tópicos

Saída:

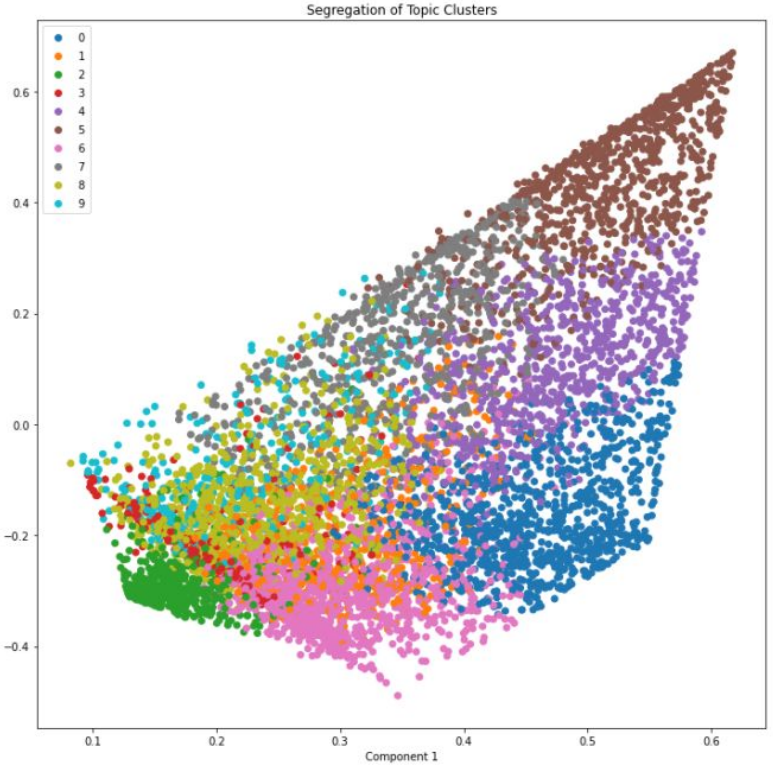
- Frequência de tópicos por documento
- Conjunto de palavras por tópico



LDA - Todos Documentos

- LDA topic discovery
- [pyLDAvis]

	Word 0	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Word 10	Word 11	Word 12	Word 13
Topic 0	pessoa	pai	politicar	vidar	querer	passar	formar	brasileiro	mundo	ficar	direito	social	achar	casar
Topic 1	militar	casar	pessoa	publicar	ficar	pai	policiar	gente	acontecer	video	passar	querer	contar	deixar
Topic 2	cidade	policiar	rir	janeiro	pessoa	saude	morto	governar	escola	morte	prefeitura	feirar	tres	publicar
Topic 3	presidente	temer	deputar	governar	ministrar	camara	senador	pmdb	senado	politicar	afirmar	votar	parlamentar	psdb
Topic 4	aumentar	mercar	empresar	produto	resultar	aguar	saude	brasileiro	medir	programar	analisar	lei	paises	reduzir
Topic 5	falar	globo	jornalista	musicar	artista	filmar	cantor	show	amigo	tv	cinema	domingo	ficar	contar
Topic 6	coreia	norte	trump	americano	nuclear	pai	armar	guerra	presidente	coreano	sul	militar	forcar	atacar
Topic 7	empresar	milhoes	dinheiro	odebrecht	contar	receber	pagamento	propinar	campanha	pagar	comprar	afirmar	obrar	acordar
Topic 8	policiar	apo	preso	hospital	familia	medicar	informar	filho	mulher	feirar	vitimar	ficar	presar	afirmar
Topic 9	presidente	federal	lavar	jato	pedir	defeso	juiz	publicar	advogar	processar	prisao	operacao	justica	denunciar

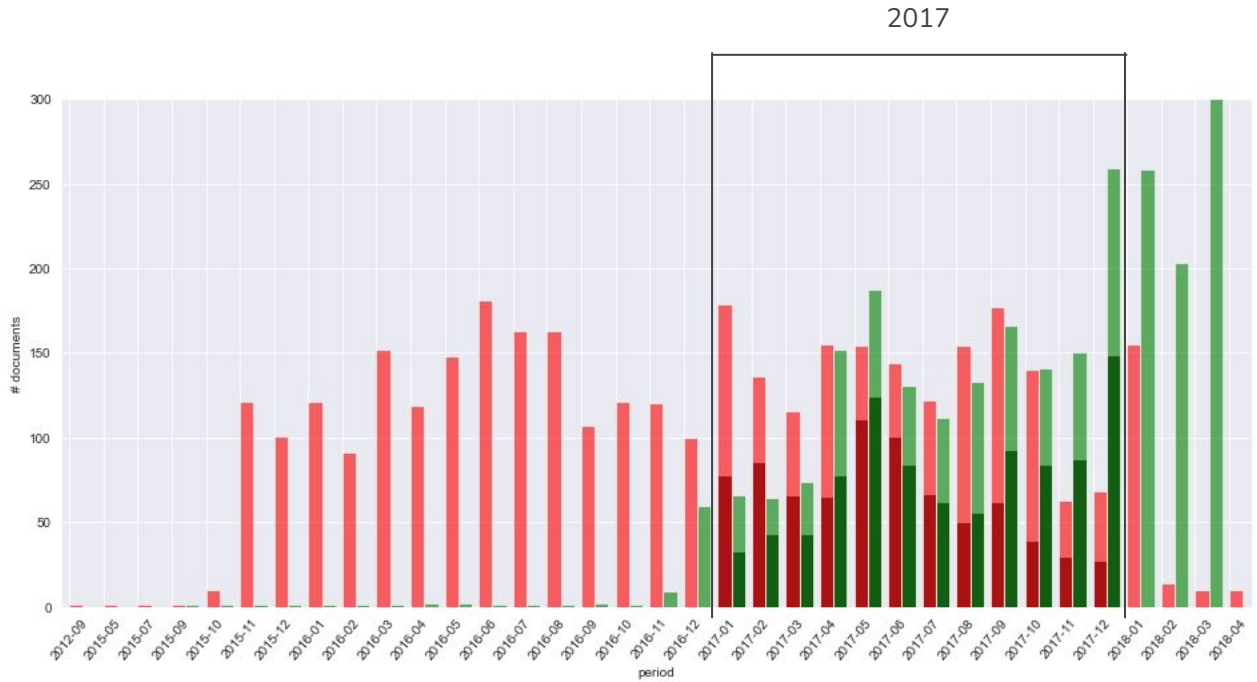


Estreitando a análise

- Focando as análises em:
  - Notícias de 2017
  - Apenas notícias de cunho político

Próximos passos:

- Aplicar LDA novamente
- GridSearch para achar os melhores parâmetros e número de tópicos



LDA para notícias políticas de 2017

3 TÓPICOS

	Word 0	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Word 10
Topic 0	temer	senador	presidente	ministrar	aecio	federal	policiar	deputar	pmdb	denunciar	pedir
Topic 1	presidente	federal	pedir	afirmar	odebrecht	lavar	dinheiro	jato	defeso	empresar	propinar
Topic 2	presidente	governar	pai	temer	politicar	deputar	pessoa	afirmar	publicar	brasileiro	passar

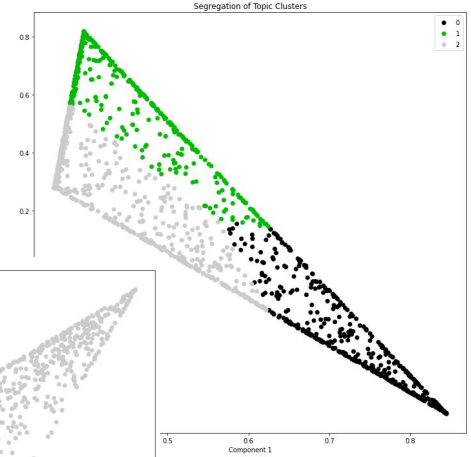
4 TÓPICOS

	Word 0	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Word 10
Topic 0	temer	deputar	presidente	governar	camara	policiar	federal	senador	pmdb	casar	senado
Topic 1	presidente	ministrar	decisao	temer	juiz	processar	jato	lavar	crime	condenar	morar
Topic 2	pai	pessoa	politicar	brasileiro	ficar	publicar	passar	querer	governar	contar	direito
Topic 3	presidente	pedir	afirmar	federal	odebrecht	dinheiro	empresar	lavar	jato	milhoes	propinar

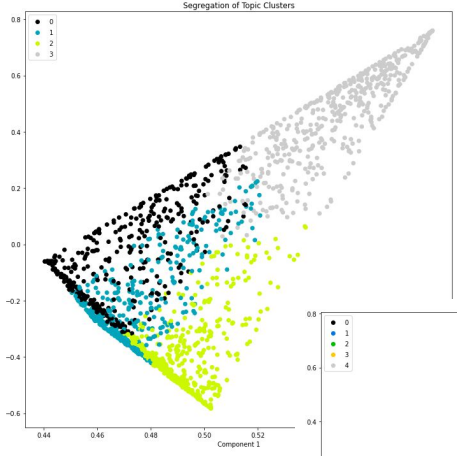
5 TÓPICOS

	Word 0	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Word 10
Topic 0	empresar	governar	publicar	cidade	pai	brasileiro	contar	obrar	alar	milhoes	rir
Topic 1	pai	pessoa	ficar	querer	policiar	brasileiro	casar	militar	apo	publicar	mulher
Topic 2	presidente	juiz	morar	federal	processar	defeso	lavar	jato	palocci	lula	prisao
Topic 3	presidente	temer	deputar	camara	governar	votar	politicar	parlamentar	senado	senador	casar
Topic 4	presidente	ministrar	pedir	temer	federal	afirmar	jato	lavar	delacao	odebrecht	campanha

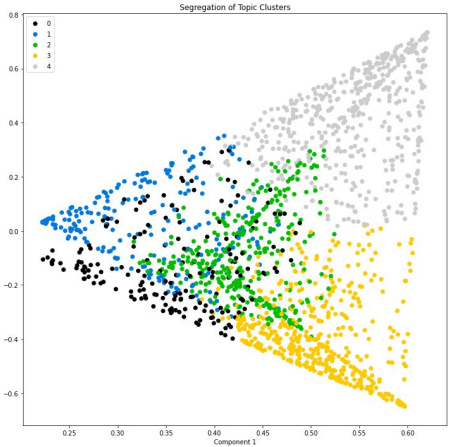
3 TÓPICOS (Log Likelihood -621096)



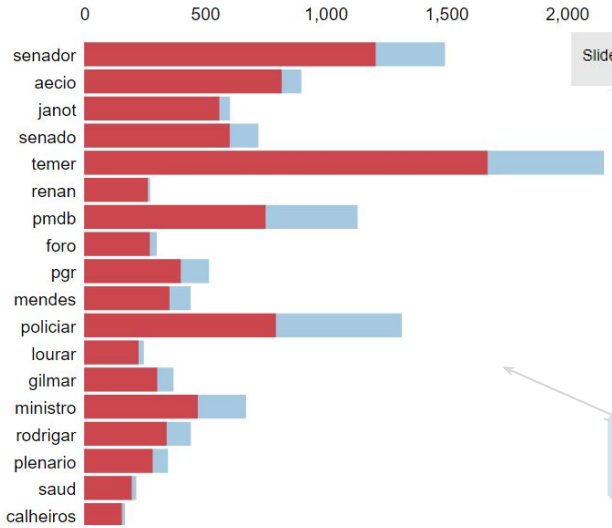
4 TÓPICOS (-624739)

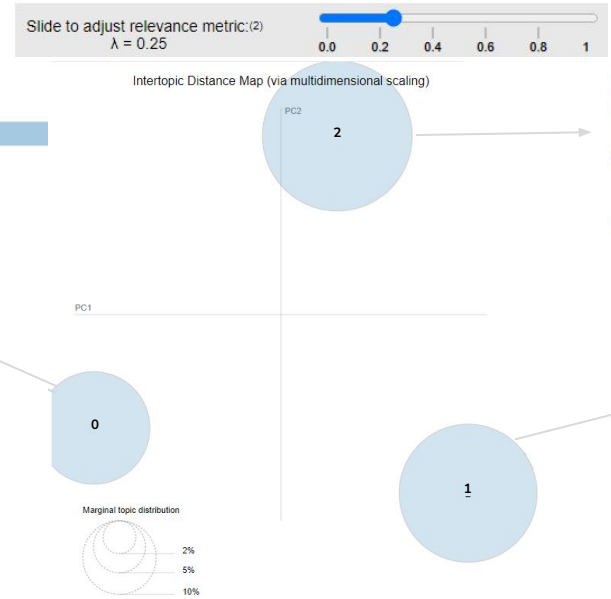


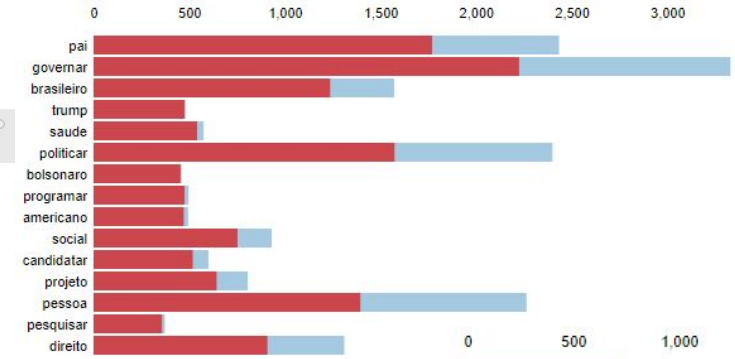
5 TÓPICOS (-628583)

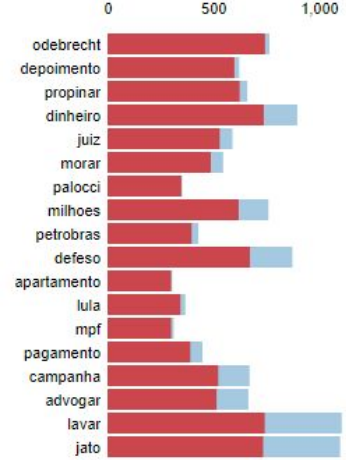


LDA para notícias políticas de 2017
(3 Tópicos)









- [pyLDAvis notebook]

Overall term frequency
Estimated term frequency within the selected topic

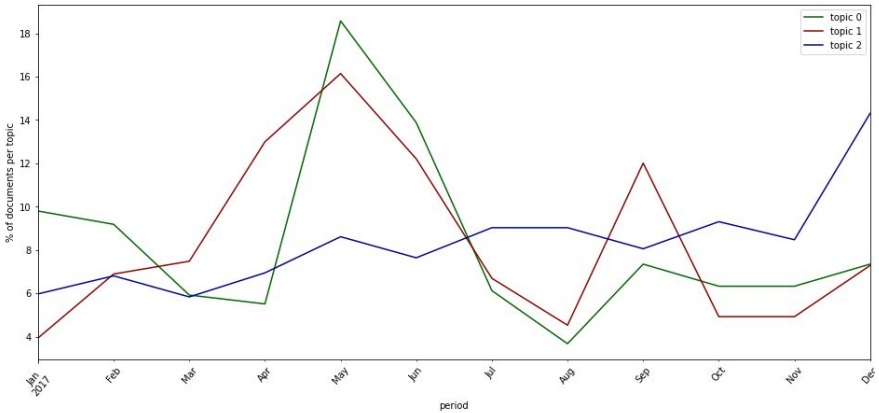
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Fatos ocorridos em 2017

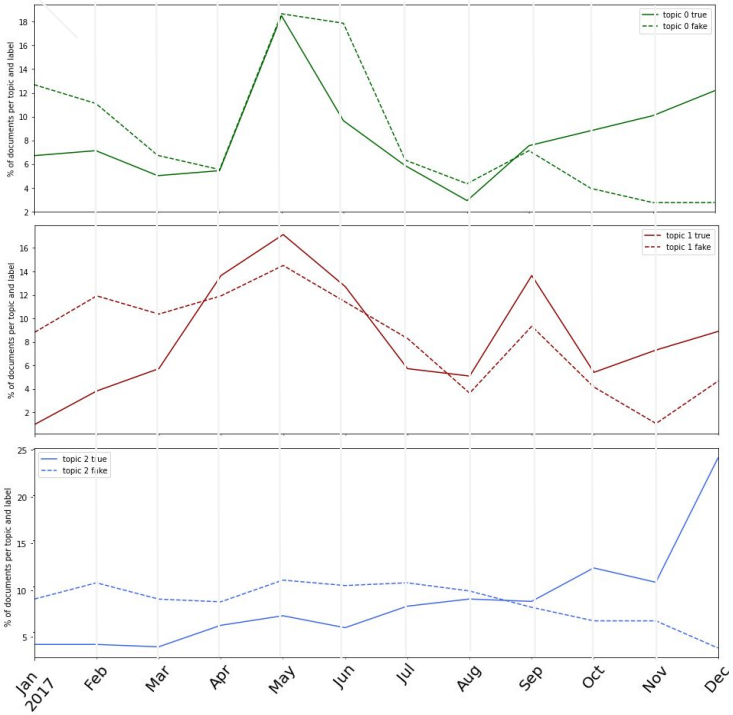
Tópico 0: Notícias relacionadas ao presidente Temer e a deputados e senadores
Tópico 1: Notícias sobre a investigação Lava-Jato
Tópico 2: Notícias sobre os EUA e pré-candidatura de Bolsonaro (discurso parecido com o de Trump)

dominant_topic	label	
0	fake	252
	true	238
1	fake	193
	true	315
2	fake	337
	true	383

	Word 0	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Word 10
Topic 0	temer	senador	presidente	ministrar	aecio	federal	policiar	deputar	pmdb	denunciar	pedir
Topic 1	presidente	federal	pedir	afirmar	odebrecht	lavar	dinheiro	jato	defeso	empresar	propinar
Topic 2	presidente	governar	pai	temer	politicar	deputar	pessoa	afirmar	publicar	brasileiro	passar



Acontecimentos políticos 2017



Tópico 0 (Presidência/Câmara):

- Janeiro: Morte de Teori Zavaski
- Maio: divulgação do áudio de Joesley Batista com Michel Temer
- Setembro: Afastamento de Aecio Neves

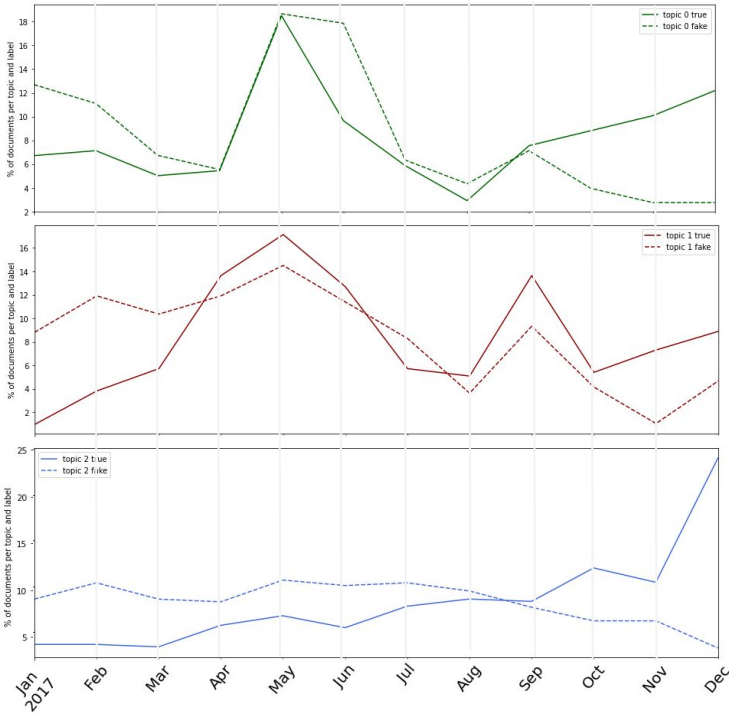
Tópico 1 (Operação Lava-Jato):

- Abril: Projeto de abuso de autoridade é aprovado
- Julho: Lula condenado a 9 anos por Moro
- Setembro: 51 milhões de Geddel

Tópico 2 (Trump e Bolsonaro):

- Janeiro: Trump assume a presidência dos EUA
- Dezembro: Trump reconhece Jerusalém como capital de Israel
- Ao longo do ano: Trump e Bolsonaro quebrando protocolos, comparações entre os dois.

Distribuição pelo tempo dos 3 tópicos

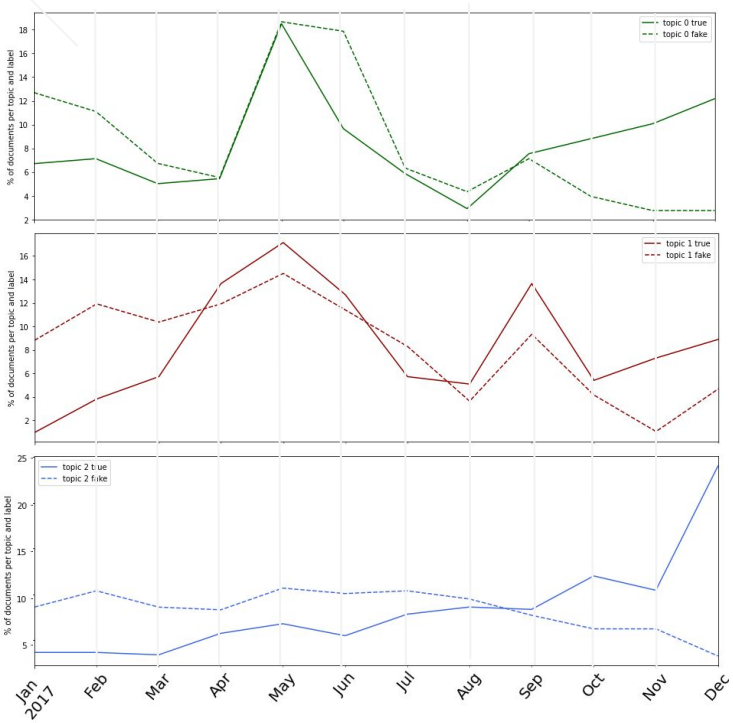


Notícias verdadeiras e falsas apresentam uma tendência semelhante em alguns casos, outros não.

Não é possível fazer uma conclusão forte sobre a relação entre notícias verdadeiras/falsas e os tópicos, pois isso pode ser uma coincidência das notícias obtidas pelo corpus.

Para uma análise mais detalhada, seria necessário a definição de uma metodologia com relação à obtenção de notícias e seus tópicos.

Acontecimentos políticos 2017



Apesar das relações apresentadas entre tópicos e fatos, há vários fatos que ocorreram no período e não causaram mudanças significativas nos tópicos.
As várias reformas feitas em 2017 (trabalhista, previdência, política) são exemplos.

Na análise foi feito o caminho reverso de, a partir da mudança da frequência de tópicos, encontrar uma justificativa para tal.
Acreditamos que encontrar a tendência de tópicos a partir de eventos políticos seja tão ou mais frutífero que o processo que foi feito.

Dificuldades

- Filtro na lematização
  - Algumas palavras perdem sentido (“Moro” se torna “morar”)
- Encontrar parâmetros ideais para o LDA
  - Número de tópicos e parâmetros de aprendizagem tomam muito tempo de teste
- Fazer sentido dos dados obtidos das descobertas de tópicos
  - Após a descoberta dos tópicos, é necessário investigar “a mão” as palavras e documentos
  - A partir disso, é possível definir uma descrição para o tópico
- Documentos muito variados
  - Para resolver isso, foi restringido apenas documentos do ano de 2017, com tema de cunho político.
- Relacionar aumento de documentos sobre tópicos com eventos políticos

Limitações

- O modelo utilizado foi o suficiente para um espaço de tempo curto, porém para intervalo maiores seria necessário utilizar uma modelagem de tópicos que levasse em conta a dimensão temporal dos documentos, como Topics Over Time.
- Como fonte de análise foram utilizadas notícias, porém tweets ou outras fontes mais “instantâneas” devem refletir melhor a volatilidade de tópicos com o passar do tempo
- Não é apresentada uma metodologia para obtenção dos dados pelo corpus, o que pode fazer com que os resultados obtidos sejam incidentais.
- Um modelo para relacionar os tópicos e os eventos políticos seria uma grande contribuição para análises nesse sentido