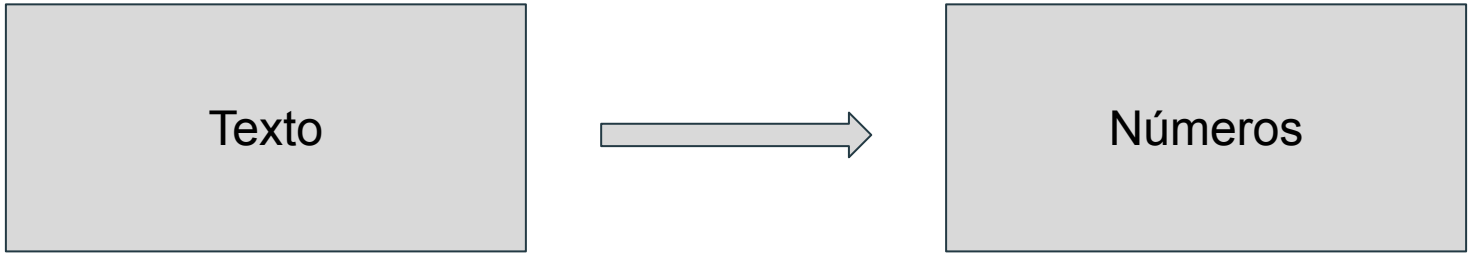
Como visualizar Word Embeddings

Feature extraction

Problema



Soluções

- Bag of Words;
- Bag of n-gram;
- Word2vec; e
- Paragraph2vec

Bag of Words

Frase	Brasil	acima	de	todos	Deus	tudo
Brasil acima de todos	1	1	1	1	0	0
Deus acima de tudo	0	1	1	0	1	1

As palavras não são ordenadas, logo se perde a noção do contexto.

Bag of n-grams

Frase	Brasil acima	acima de	de todos	Deus acima	de tudo
Brasil acima de todos	1	1	1	0	0
Deus acima de tudo	0	1	0	1	1

Quando expandimos os n-grams obtemos matrizes esparsas de alta dimensionalidade. Ainda assim, é capaz apenas de observar pequenos contextos.

Word2Vec

Se tomar vacina [e] virar jacaré não tenho nada a ver com isso...

Maximizar a seguinte função de verossimilhança:

$$\frac{1}{T}\sum_{t=k}^{T-k}\log p(w_t|w_{t-k},...,w_{t+k})$$

Ou seja, encontrar o melhor contexto a partir da palavra central.

Paragraph2vec

Memória distribuída (PV-DM):

Se tomar vacina [e] virar jacaré não tenho nada a ver com isso...

ID Parágrafo (Contém informações faltantes no contexto) {Em verde}

A partir do contexto predizemos o centro. O modelo matemático é semelhante ao Word2Vec, porém há a adição de um ente que se refere a memória do parágrafo.

Paragraph2vec

Bag of word distribuídas
(PV-DBOW):

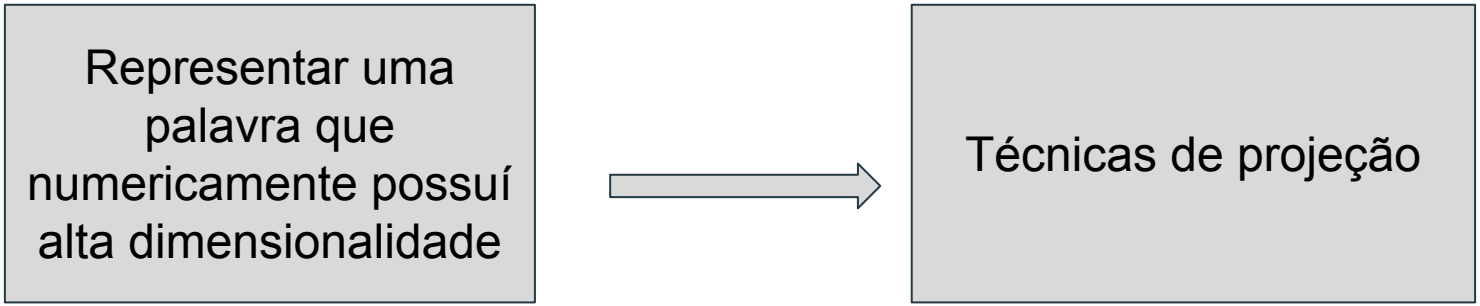
Se tomar vacina [e] virar jacaré não tenho nada a ver com isso...

ID Parágrafo (Contém informações faltantes no contexto) {Em verde}

Treinamos um vetor que representa o parágrafo para predizer todas as palavras no contexto em questão.

Representação Gráfica

Problema - Solução



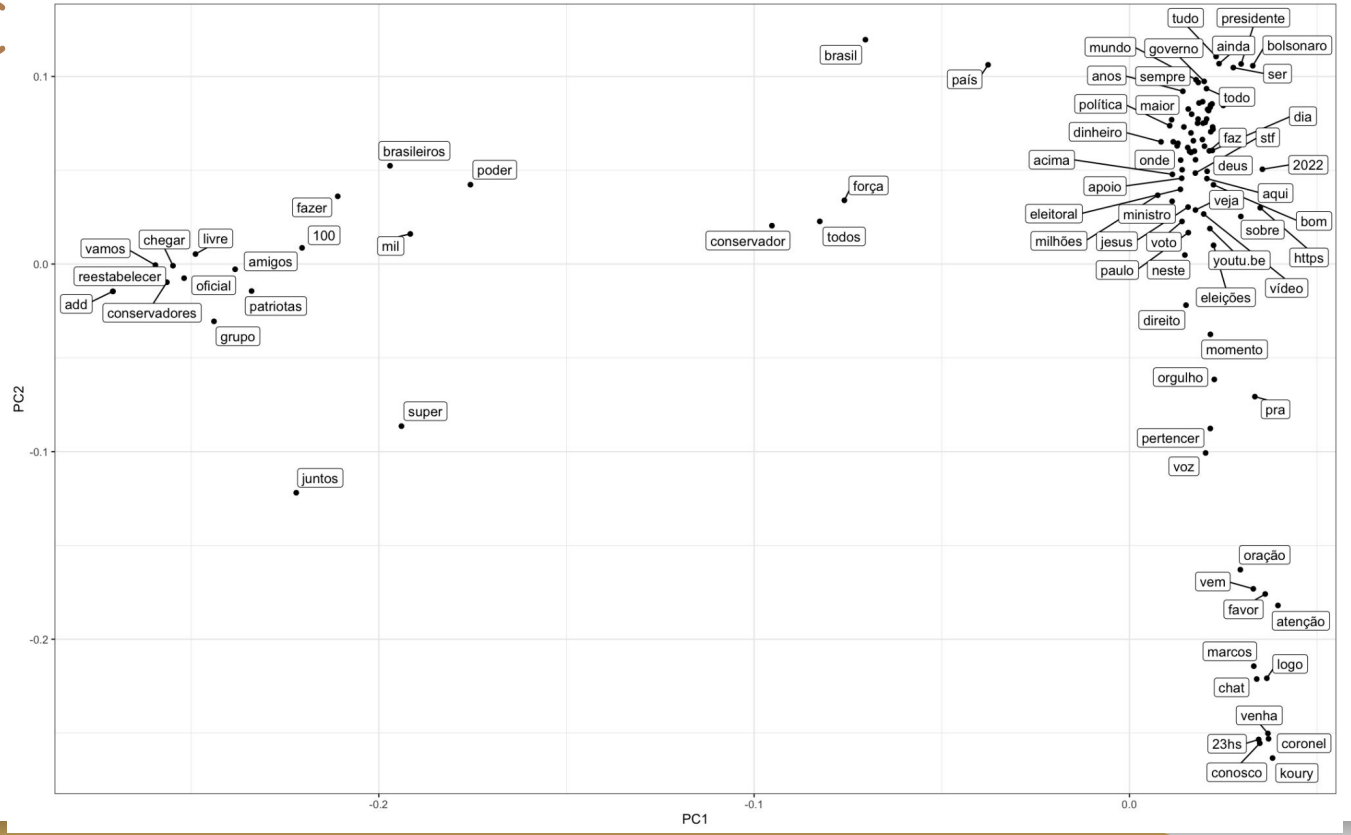
Soluções

- PCA;
- Umap;
- T-sne;

PCA

Objetivo: Encontrar um novo conjunto de variáveis menor que o conjunto original que preserve a maior parte da variabilidade presente nos dados (pautado na covariância).

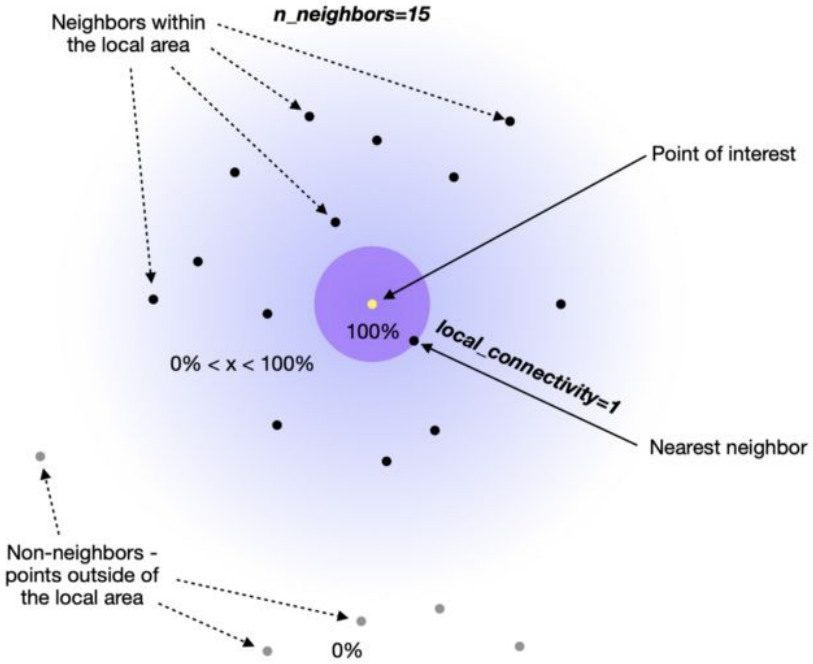
PC


PC2
PC1
0.1
0.0
-0.1
-0.2
-0.2
-0.1
0.0
brasileiros
poder
fazer
100
mil
vamos
chegar
livre
amigos
reestabelecer
oficial
add
conservadores
patriotas
grupo
super
juntos
brasil
país
força
conservador
todos
tudo
presidente
mundo
governo
ainda
bolsonaro
anos
sempre
ser
todo
política
maior
dia
dinheiro
faz
stf
acima
onde
deus
2022
apoio
aqui
veja
eleitoral
ministro
bom
sobre
milhões
jesus
voto
https
paulo
neste
youtu.be
vídeo
eleições
direito
momento
orgulho
pra
pertencer
voz
oração
vem
favor
atenção
marcos
logo
chat
venha
23hs
coronel
conosco
koury

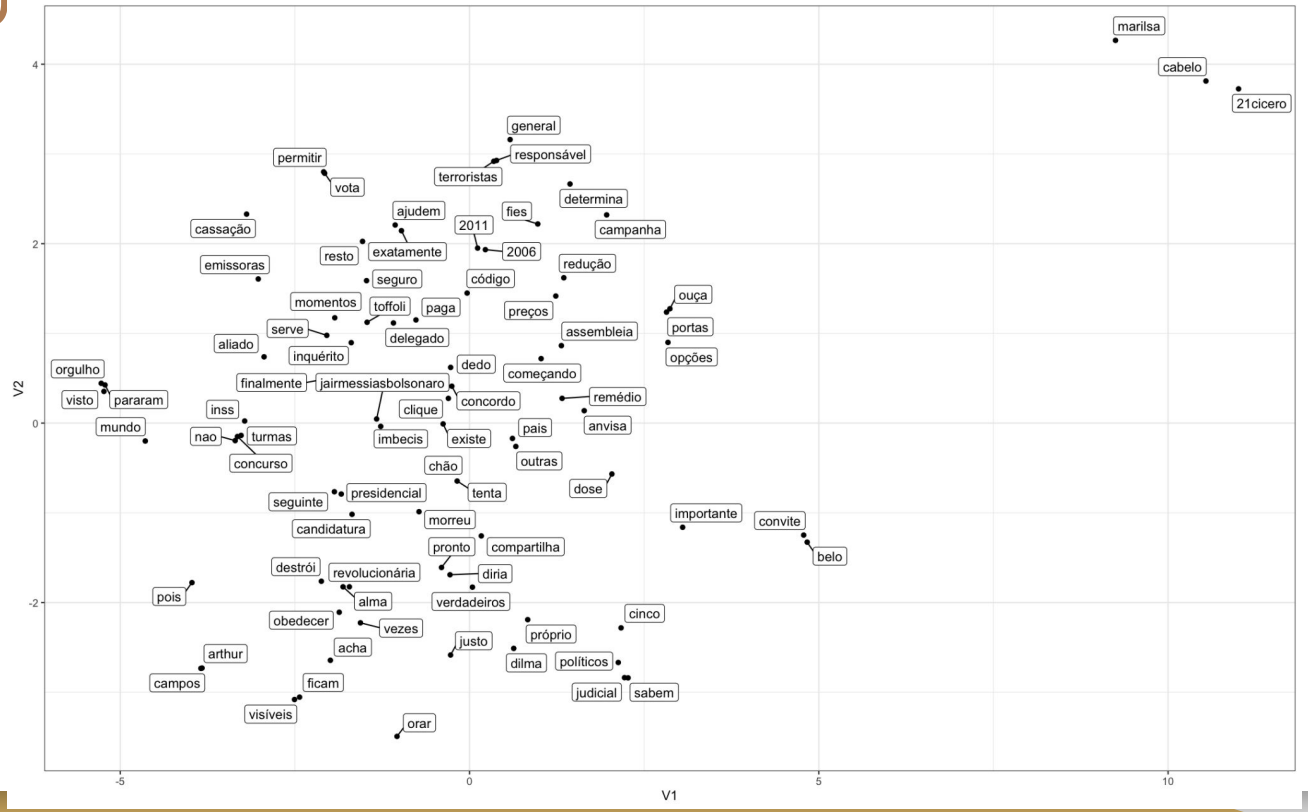
UMAP

UMAP

UMAP


Neighbors within
the local area
n_neighbors=15
Point of interest
100%
local_connectivity=1
0% < x < 100%
Nearest neighbor
Non-neighbors -
points outside of
the local area
0%

UMAP


marilsa
cabelo
21cicero
general
responsável
permitir
vota
terroristas
determina
ajudem
fies
campanha
cassação
2011
2006
exatamente
resto
emissoras
redução
seguro
código
ouça
momentos
toffoli
paga
preços
portas
serve
assembleia
delegado
aliado
inquérito
opções
orgulho
dedo
começando
finalmente
jairmessiasbolsonaro
remédio
visto
pararam
concordo
inss
clique
anvisa
mundo
pais
nao
turmas
imbecis
existe
concurso
outras
chão
dose
seguinte
presidencial
tenta
importante
convite
candidatura
morreu
pronto
compartilha
belo
destrói
revolucionária
diria
pois
alma
verdadeiros
cinco
obedecer
vezes
próprio
justo
arthur
acha
dilma
políticos
campos
ficam
judicial
sabem
visíveis
orar
V1
V2
4
2
0
-2
-5
0
5
10

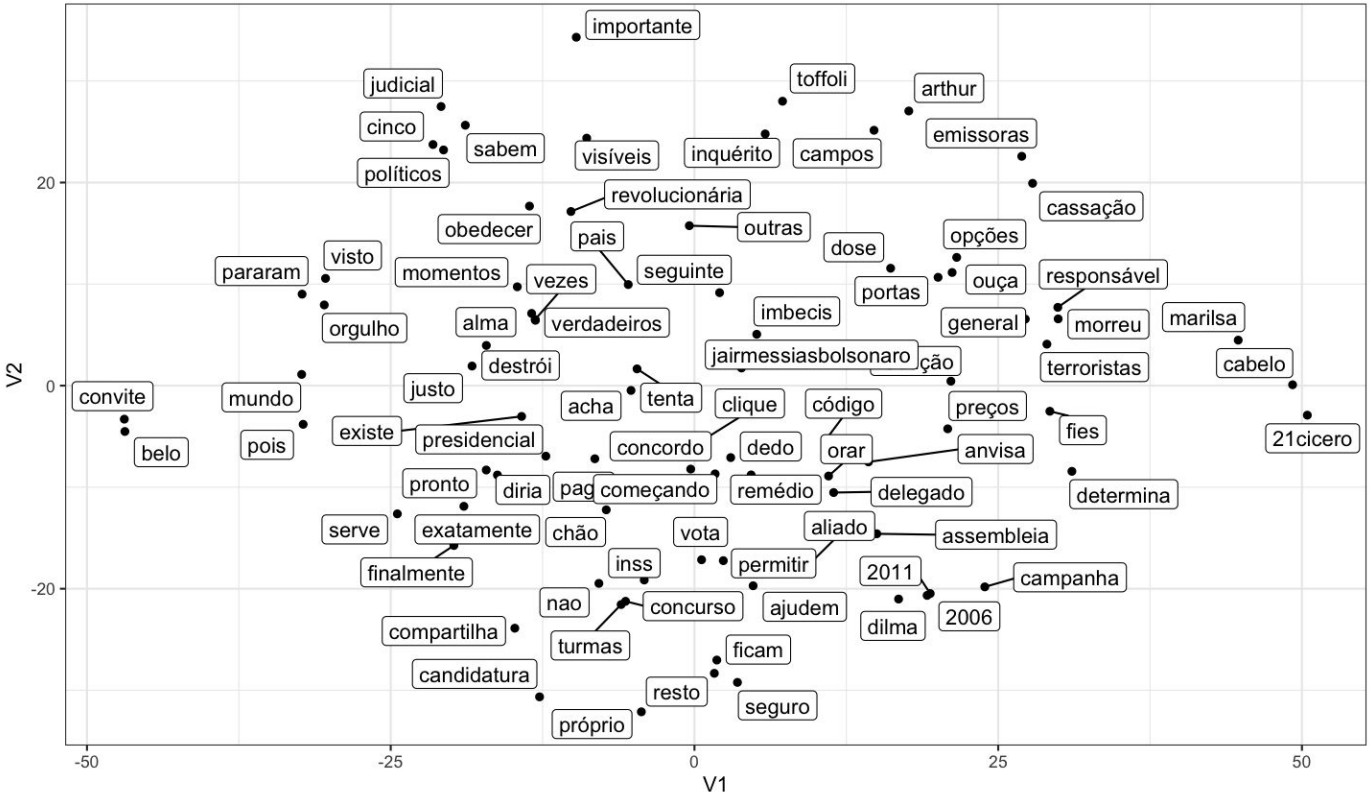
TSNE

T-SNE

Para transformarmos distâncias em probabilidades, imaginamos cada ponto como o centro de uma distribuição e usamos a distância entre este ponto e os outros como a probabilidade de ocorrência daquela distância dentro da distribuição escolhida.

b)

TSNE


importante
judicial
toffoli
arthur
cinco
sabem
visíveis
inquérito
campos
emissoras
políticos
cassação
revolucionária
obedecer
pais
outras
dose
opções
pararam
visto
momentos
vezes
seguinte
portas
ouça
responsável
orgulho
alma
verdadeiros
imbecis
general
morreu
marilsa
destrói
jairmessiasbolsonaro
terroristas
cabelo
convite
mundo
justo
tenta
acha
clique
código
preços
fies
belo
pois
existe
presidencial
concordo
dedo
orar
anvisa
21cicero
pronto
diria
começando
remédio
delegado
determina
serve
exatamente
chão
vota
aliado
assembleia
finalmente
inss
permitir
2011
campanha
nao
concurso
ajudem
2006
compartilha
turmas
dilma
ficam
candidatura
resto
seguro
próprio
V2
V1
20
0
-20
-50
-25
0
25
50