Extração de tópicos em Notas Fiscais Eletrônicas (NF-e) não rotuladas: uma análise utilizando BERTopic

Apresentação de Trabalho de Conclusão de Curso

Antônio Oss Boll

Universidade Federal do Rio Grande do Sul

20 de fevereiro de 2024

Sumário

Introdução

BERTopic

Resultados

Conclusão

Introdução

- Há uma maior dificuldade em trabalhar com dados não rotulados, uma vez que não há um resultado comparável ao predito.
- O objetivo desse trabalho é fazer o agrupamento de tópicos para diferentes produtos utilizando as descrições contidas nas NF-e sem rotulação. Os dados foram disponibilizados pela SEFAZ-RS.

Modelo BERTopic

- Trabalhando sem dados rotulados, o BERTopic se torna uma alternativa aos modelos tradicionais de *topic modeling* como o LDA (Blei et al. [2003]) (Matriz de frequência) e o CTM (Blei and Lafferty [2007]).
- Diferentemente dos tradicionais, o BERTopic obtém a semântica da frase dos modelos BERT (Devlin et al. [2018]), fazendo assim, uma modelagem considerando o contexto da frase. Ele é dividido em três partes que serão explicadas a seguir.

Word Embeddings

Word Embeddings são a representação de palavras como vetores (Mikolov et al. [2013]).
Dessa maneira, palavras similares são posicionadas próximas umas às outras.

- Rei - Homem + Mulher = Rainha

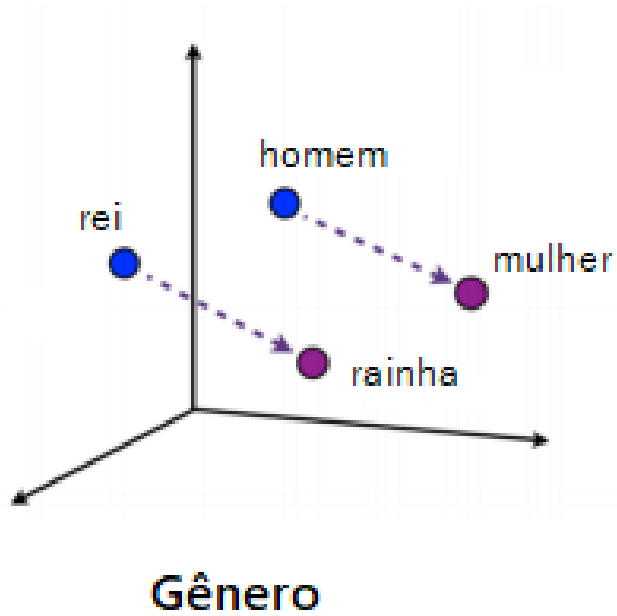


Figure: Exemplo vetorial de um Word Embedding.[1]

[1] Fonte: Artigo da Medium

Document Embedding

Inicialmente, o BERTopic trabalha com o S-BERT (Reimers and Gurevych [2019]) para gerar *sentence embeddings*. Ele é uma variação do BERT.

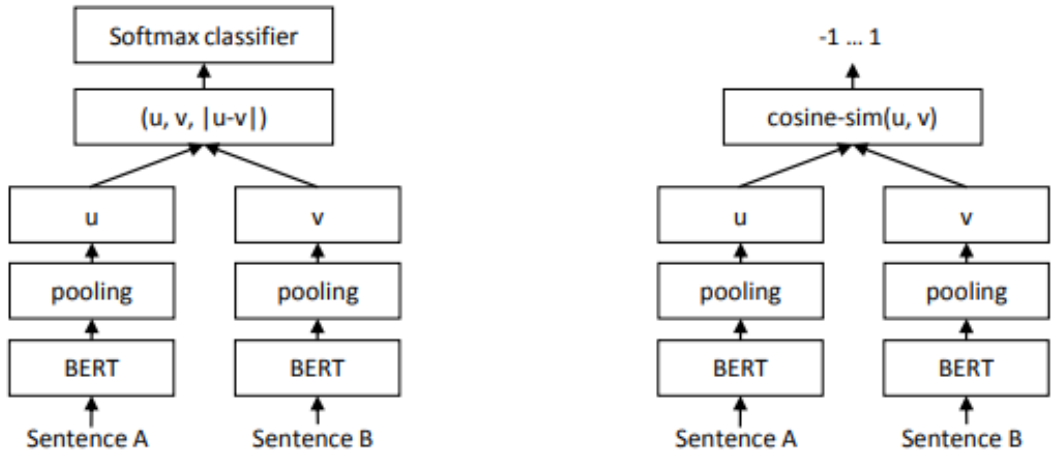


Figure: Representação do S-BERT.[2]

Os resultados do S-BERT são utilizados para criar *clusters*, o próximo passo do BERTopic.

[2]Fonte: Reimers and Gurevych [2019]

Document Clustering UMAP

Os vetores gerados dos *Document Embedding* possuem muitas dimensões. Dessa forma, o UMAP (McInnes and Healy [2018]) é utilizado para uma redução de dimensionalidade, que seria a projeção de dados de uma dimensão maior para uma menor.

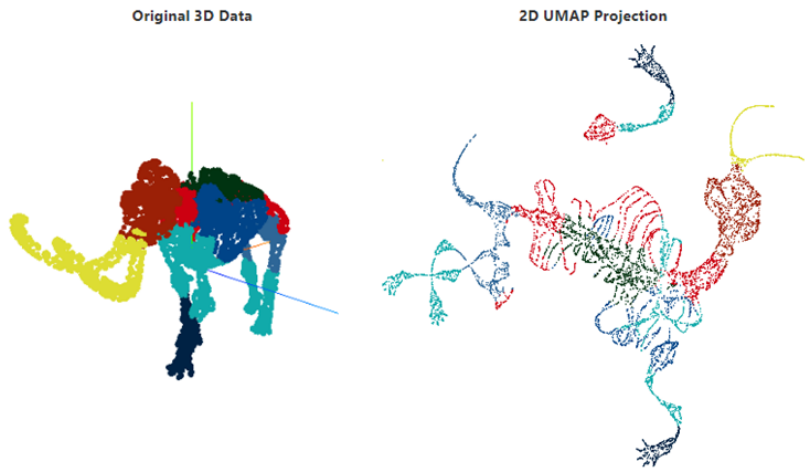


Figure: Representação do UMAP.[3]

Ele possui como objetivo replicar os *clusters* das maiores dimensões às menores sem perder os padrões.

[3] Fonte: Khadivi et al. [2023]

Document Clustering HDBSCAN

A partir do UMAP, é utilizado o HDBSCAN (Campello et al. [2013]) para *clusterizar* os *embeddings*. Ele realiza a clusterização hierárquica seguindo o padrão do DBSCAN (Ester et al. [1996])

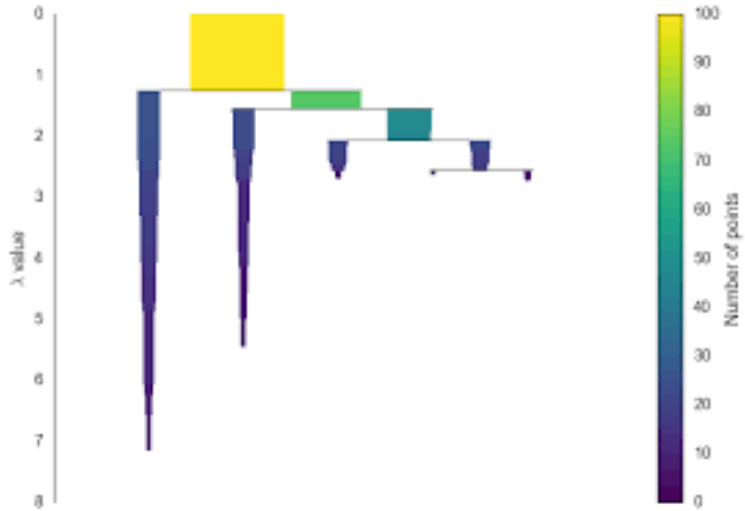


Figure: Representação do HBDSCAN.[4]

[4] Fonte: https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html

Topic Representation

Após o HDBSCAN, o algoritmo passa pelo cTF-IDF, um método análogo ao TF-IDF que obtém as palavras chave de cada *cluster*. Ele realiza o cálculo do TF-IDF para cada agrupamento, obtendo as palavras mais importantes por tópico.

$$\textbf{cTF(term)} = \left(\frac{n_{\text{aparição do termo no } \textit{cluster}}}{n_{\text{total de termos no } \textit{cluster}}} \right). \tag{1}$$

$$\textbf{cIDF(term)} = \ln \left(1 + \frac{n_{\text{médio de palavras por } \textit{cluster}}}{n_{\textit{clusters} \text{ contendo o termo}}} \right). \tag{2}$$

$$\textbf{cTF-IDF(term)} = \text{cTF(term)} \times \text{cIDF(term)}. \tag{3}$$

Modelo BERTopic

Uma ideia desenvolvida pelo autor do artigo foi a abertura para a aplicação de diferentes métodos e modelos de representação, podendo utilizar métodos como o **K-Means, PCA, LLMs**, entre outros. Alguns modelos de representação usados serão explicados.

Modelo de Representação: BART

- O modelo BART (Lewis et al. [2019]) foi o mais utilizado, mais especificamente realizando um *Zero-Shot Learning* com alguns tópicos possíveis. Esse método é uma forma de realizar um ajuste fino à rede neural buscando a otimização de nenhum exemplo.

Zero-Shot Learning

- São entregues palavras chave juntamente àquelas geradas pelo c-TF-IDF. Se algum tópico for semelhante àquelas palavras chave entregues, ele recebe o nome sugerido. Se não, o nome do c-TF-IDF é mantido.
- Exemplificando, o modelo BART recebe um input, como [“leite”, “carne”] e nomeia tópicos relacionados a esses temas com base em seu “conhecimento prévio” e as palavras destacadas.

Modelos de Representação

Existem vários tipos de modelos, pré treinados ou não, que ajudam no ajuste fino do BERTopic. São eles:

- KeyBERTInspired
- Zephyr Mistral 7B

Entre outros métodos e modelos.

Corpus

O corpora utilizado é composto por descrições contidas de NF-e disponibilizados pela SEFAZ-RS. Ele é dividido em:

- Corpus refinado por uma cadeia de markov (Markov [1906]) para compreender observações de leite e carne. É composto por aproximadamente 384 mil observações e passou, além da cadeia de Markov, pela remoção de acentuação e conversão de letras maiúsculas para minúsculas;
- Corpus sem refinamento composto por aproximadamente 203 mil mercadorias de NF-e da SEFAZ-RS. Ele recebeu o pré-processamento semelhante ao outro banco.

Métrica

A métrica utilizada foi a de silhueta (Rousseeuw [1987]). Ela consiste em um valor que identifica o quão cada *embedding* é similar ao seu *cluster* comparado à outros *clusters*. Quanto maior o valor da silhueta, melhor o algoritmo realizou as clusterizações. Sua fórmula é descrita a partir da Equação (4):

$$\frac{r - s}{\max(s, r)}, \tag{4}$$

No qual **s** é a média da distância intra-*cluster* e **r** é a média do *cluster* mais próximo. O resultado final é a média de todos os pontos calculados.

Outra métrica

Além da silhueta, foi considerado o número de tópicos gerados.

Resultados Corpus refinado "carne e leite"

Modelo	Modelo de Embedding	Palavras Chave	Tamanho mínimo	Número de tópicos	Modelo de Representação	Tempo	Silhueta	Banco utilizado (%)
3	"all"	5	500	X (77)	X	1d 4h 36 Min	0,70	100
12	"p-m"	5	500	15 (13)	KeyBERT Inspired	8 Min	0,67	20
13	"p-m"	5	500	10 (10)	Mistral Zephyr 7B	8h 37 Min	0,01	100
15	"p-m"	10	X	X (5470)	BART	1h 17 Min	0,93	50
16	**"p-m-MiniLM-L12-v2"**	**10**	**1000**	**X (7)**	**BART**	**9 Min**	**0,67**	**20**

Configuração do corpus refinado: Melhor modelo

A melhor configuração do primeiro corpus foi a seguinte:

Modelo	Modelo de Embedding	Palavras Chave	Tamanho mínimo	Número de tópicos	Modelo de Representação	Tempo	Silhueta	Banco utilizado (%)
17	"paraphrase-multilingual-MiniLM-L12-v2"	10	1000	X (21)	BART	8h 1 Min	0,62	100

Zero-Shot do Corpus com refinamento “carne e leite”

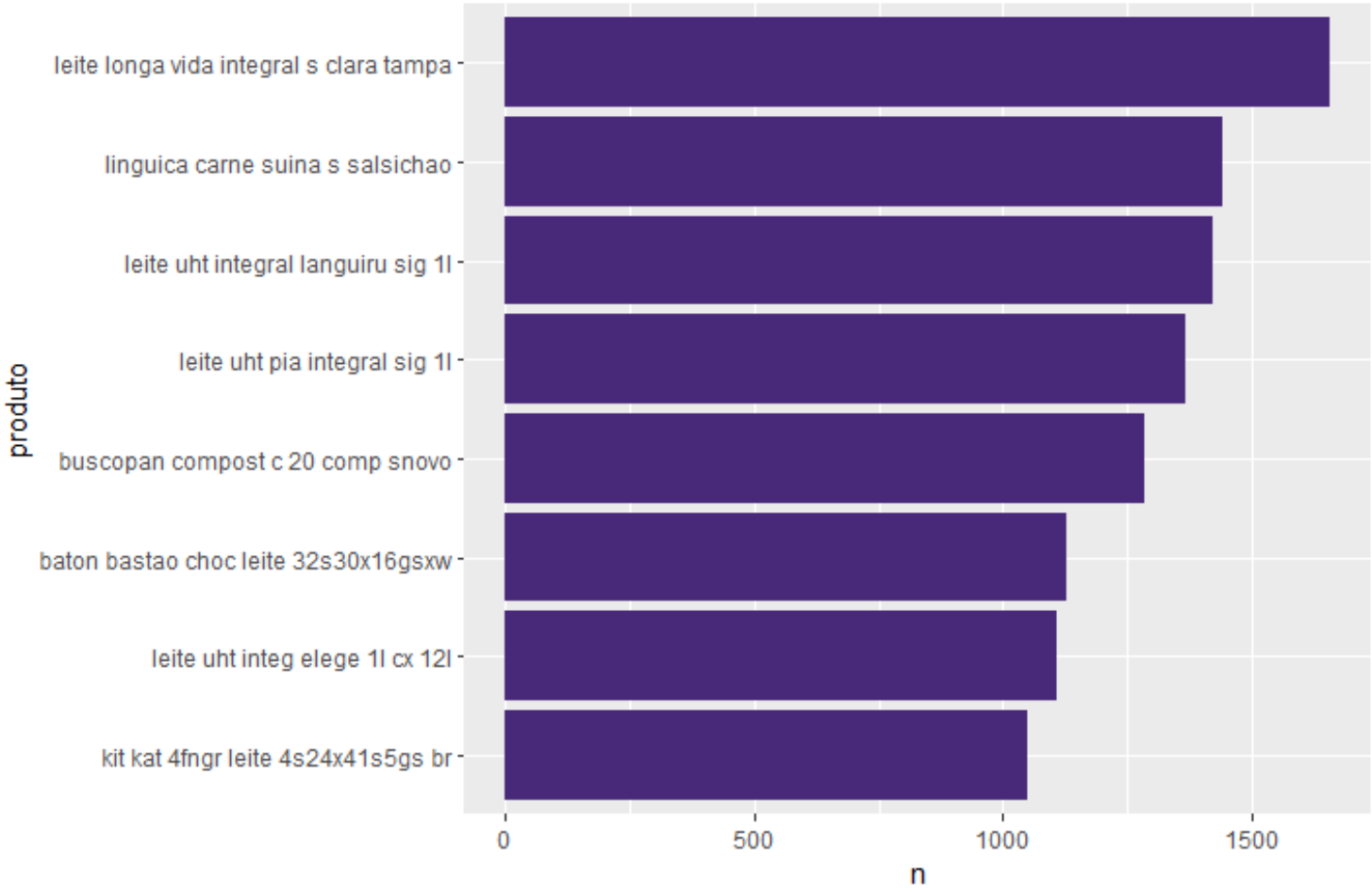
Zero Shot:
[“leite”, “carne”, “frango”]

Melhor configuração: Características básicas

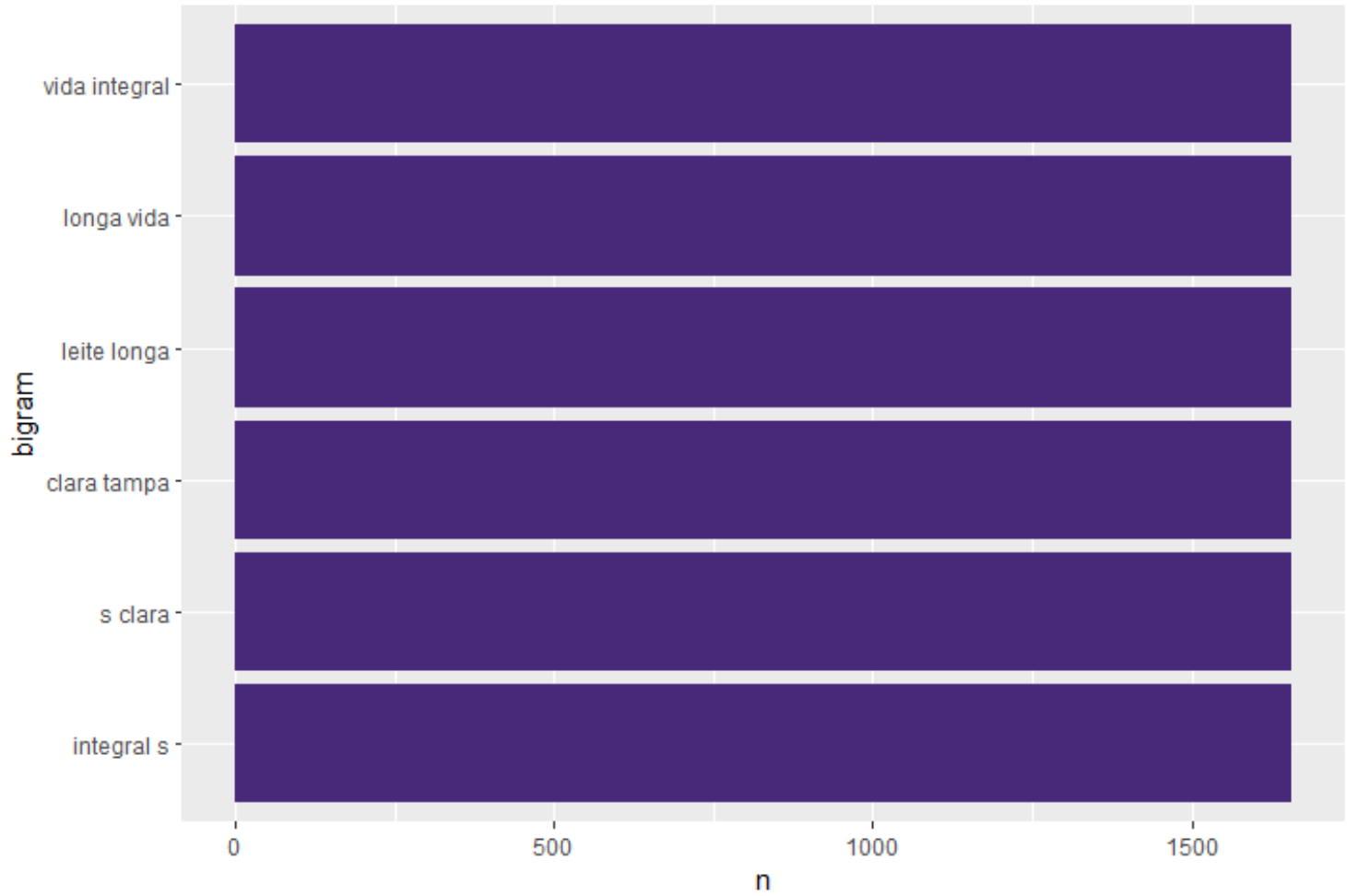
```
-1              -1_kg_leite_frango_carne
0                           0_leite___
1         1_compost_buscopan_snovo_comp
2                           2_leite___
3                           3_leite___
4                          4_frango___
5                           5_leite___
6                           6_leite___
7                           7_carne___
8                           8_leite___
9                           9_leite___
10                         10_leite___
11          11_pao_pullman_integral_500g
12    12_salsichao_linguica_suina_carne
13                         13_carne___
14                         14_leite___
15                         15_leite___
16                         16_carne___
17        17_osso_3303speito_suino_bovi
18                  18_kg_po_sabao_tixan
19                      19_tipo_po_sx_ss
Name: Name, dtype: object
```

```
-1     317843
0        1706
1        1284
2        1127
3        1492
4        3999
5        1247
6        1680
7        1933
8        1072
9        2248
10       1169
11       2104
12       1443
13       1007
14       1197
15       3909
16       2578
17       1151
18       2333
19      31732
Name: Count, dtype: int64
```

Melhor configuração: Palavras mais frequentes



Melhor configuração: Bigrams mais frequentes



Melhor configuração - “Tópico Lixo”

```
Topic                                                                -1
Count                                                            317843
Name                                          -1_kg_leite_frango_carne
Representation          [kg, leite, frango, carne, po, integral, peito...
Representative_Docs     [cortes frango coxa sobrecoxa osso cong pct da...
Name: 0, dtype: object
Topico -1 - -1_kg_leite_frango_carne - Top 5 de cada topico:
['kg', 'leite', 'frango', 'carne', 'po', 'integral', 'peito', 'cong', 'cx', 'pao']
```

Melhor configuração - “Leite”

```
Topic                                                                0
Count                                                             1706
Name                                                       0_leite___
Representation                                  [leite, , , , , , , , , , ]
Representative_Docs    [leite uht pia integral sig 1l, leite uht pia ...
Name: 1, dtype: object
Topico 0 - 0_leite___ - Top 5 de cada topico:
['leite', '', '', '', '', '', '', '', '', '']
Topic                                                                3
Count                                                             1492
Name                                                       3_leite___
Representation                                  [leite, , , , , , , , , , ]
Representative_Docs    [leite uht integral languiru sig 1l, leite uht...
Name: 4, dtype: object
Topico 3 - 3_leite___ - Top 5 de cada topico:
['leite', '', '', '', '', '', '', '', '', '']
 Topic                                                               6
 Count                                                            1680
 Name                                                       6_leite___
 Representation                                  [leite, , , , , , , , , , ]
 Representative_Docs    [leite longa vida integral s clara tampa, leit...
 Name: 7, dtype: object
 Topico 6 - 6_leite___ - Top 5 de cada topico:
 ['leite', '', '', '', '', '', '', '', '', '']
 Topic                                                              10
 Count                                                            1169
 Name                                                      10_leite___
 Representation                                  [leite, , , , , , , , , , ]
 Representative_Docs    [leite uht integ elege 1l cx 12l, leite uht in...
 Name: 11, dtype: object
 Topico 10 - 10_leite___ - Top 5 de cada topico:
 ['leite', '', '', '', '', '', '', '', '', '']
```

Melhor configuração - “Leite”

```
Topic                                                                    2
Count                                                                 1127
Name                                                           2_leite___
Representation                                  [leite, , , , , , , , , , ]
Representative_Docs    [baton bastao choc leite 32s30x16gsxw, baton b...
Name: 3, dtype: object
Topico 2 - 2_leite___ - Top 5 de cada topico:
['leite', '', '', '', '', '', '', '', '', '']
```

```
Topic                                                                    5
Count                                                                 1247
Name                                                           5_leite___
Representation                                  [leite, , , , , , , , , , ]
Representative_Docs    [leite magnesia phil 350ml, leite magnesia phi...
Name: 6, dtype: object
Topico 5 - 5_leite___ - Top 5 de cada topico:
['leite', '', '', '', '', '', '', '', '', '']
```

```
Topic                                                                    8
Count                                                                 1072
Name                                                           8_leite___
Representation                                  [leite, , , , , , , , , , ]
Representative_Docs    [kit kat 4fngr leite 4s24x41s5gs br, kit kat 4...
Name: 9, dtype: object
Topico 8 - 8_leite___ - Top 5 de cada topico:
['leite', '', '', '', '', '', '', '', '', '']
```

```
Topic                                                                    9
Count                                                                 2248
Name                                                           9_leite___
Representation                                  [leite, , , , , , , , , , ]
Representative_Docs    [danoninho leite fermentado 450g 1x450gr, dano...
Name: 10, dtype: object
Topico 9 - 9_leite___ - Top 5 de cada topico:
['leite', '', '', '', '', '', '', '', '', '']
```

Melhor configuração - “Leite”

```
Topic                                                              14
Count                                                            1197
Name                                                      14_leite___
Representation                                  [leite, , , , , , , , , , ]
Representative_Docs    [leite po 1 kg, leite po 1 kg, leite po 1 kg]
Name: 15, dtype: object
Topico 14 - 14_leite___ - Top 5 de cada topico:
['leite', '', '', '', '', '', '', '', '', '']


Topic                                                              15
Count                                                            3909
Name                                                      15_leite___
Representation                                  [leite, , , , , , , , , , ]
Representative_Docs    [cuca doce leite, cuca doce leite, cuca doce l...
Name: 16, dtype: object
Topico 15 - 15_leite___ - Top 5 de cada topico:
['leite', '', '', '', '', '', '', '', '', '']
```

Melhor configuração - “Frango”

```
Topic                                                                  4
Count                                                               3999
Name                                                          4_frango___
Representation                                  [frango, , , , , , , , , , ]
Representative_Docs    [frango resfriado s miudos spacotes, frango re...
Name: 5, dtype: object
Topico 4 - 4_frango___ - Top 5 de cada topico:
['frango', '', '', '', '', '', '', '', '', '']
```

Melhor configuração - “Carne”

```
Topic                                                                  7
Count                                                               1933
Name                                                          7_carne___
Representation                                      [carne, , , , , , , , , , ]
Representative_Docs    [file peito resfriado s pct, file peito resfri...
Name: 8, dtype: object
Topico 7 - 7_carne___ - Top 5 de cada topico:
['carne', '', '', '', '', '', '', '', '', '']
Topic                                                                 13
Count                                                               1007
Name                                                         13_carne___
Representation                                      [carne, , , , , , , , , , ]
Representative_Docs    [carne bovina moida, carne moida bovina i, car...
Name: 14, dtype: object
Topico 13 - 13_carne___ - Top 5 de cada topico:
['carne', '', '', '', '', '', '', '', '', '']
```

Melhor configuração - "Outros"

```
Topic                                                              1
Count                                                           1284
Name                                        1_compost_buscopan_snovo_comp
Representation          [compost, buscopan, snovo, comp, 20, 7p, polim...
Representative_Docs     [buscopan compost c 20 comp snovo, buscopan co...
Name: 2, dtype: object
Topico 1 - 1_compost_buscopan_snovo_comp - Top 5 de cada topico:
['compost', 'buscopan', 'snovo', 'comp', '20', '7p', 'polim', 'basic', 'disco', 'ceramico']

Topic                                                             11
Count                                                           2104
Name                                       11_pao_pullman_integral_500g
Representation          [pao, pullman, integral, 500g, 1x400gr, rap10,...
Representative_Docs     [pao integral 500g pullman, pao integral 500g ...
Name: 12, dtype: object
Topico 11 - 11_pao_pullman_integral_500g - Top 5 de cada topico:
['pao', 'pullman', 'integral', '500g', '1x400gr', 'rap10', '450g', 'visconti', 'casca', '330g']


Topic                                                             12
Count                                                           1443
Name                                    12_salsichao_linguica_suina_carne
Representation          [salsichao, linguica, suina, carne, salsichaos...
Representative_Docs     [linguica carne suina s salsichao, linguica ca...
Name: 13, dtype: object
Topico 12 - 12_salsichao_linguica_suina_carne - Top 5 de cada topico:
['salsichao', 'linguica', 'suina', 'carne', 'salsichaos', 'dd169', 'borrussia', 'divikrek', 'flocos', 'display'
```

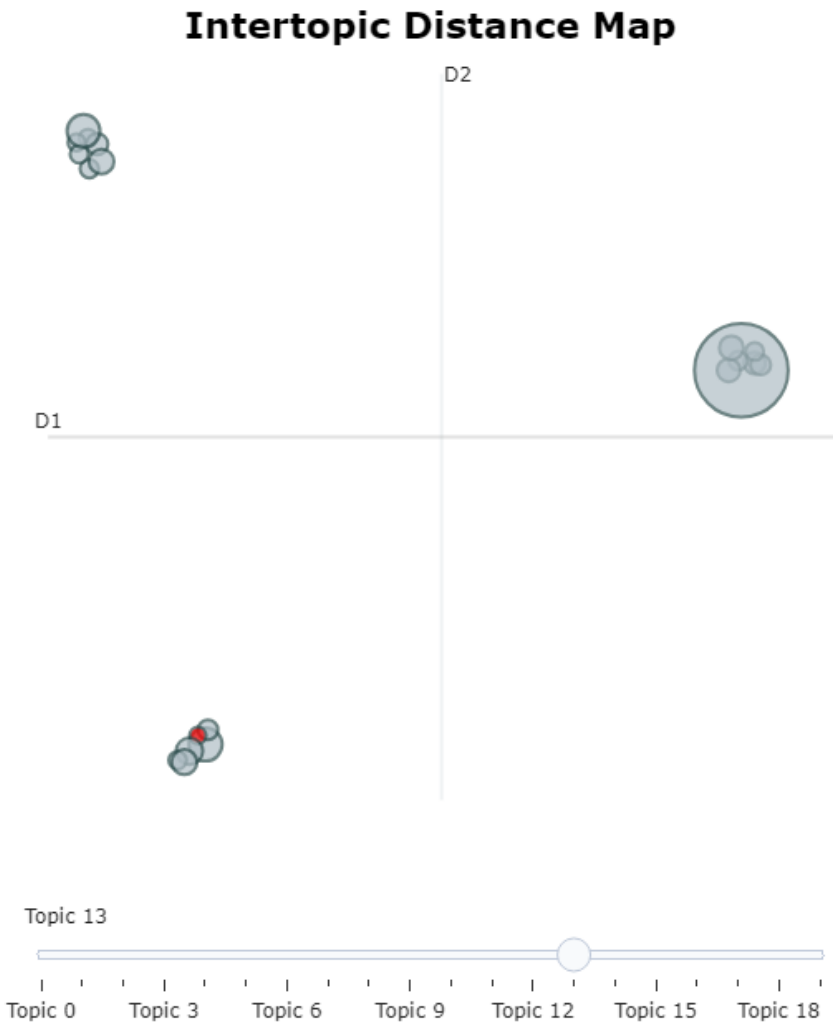
Melhor configuração - "Outros"

```
Topic                                                              17
Count                                                            1151
Name                                          17_osso_3303speito_suino_bovi
Representation          [osso, 3303speito, suino, bovi, palito, bob, c...
Representative_Docs                              [osso kg, osso kg, osso kg]
Name: 18, dtype: object
Topico 17 - 17_osso_3303speito_suino_bovi - Top 5 de cada topico:
['osso', '3303speito', 'suino', 'bovi', 'palito', 'bob', 'cong', 'pernil', 'flex', 'dianteiro']

Topic                                                              18
Count                                                            2333
Name                                                18_kg_po_sabao_tixan
Representation          [kg, po, sabao, tixan, det, ype, maciez, peso,...
Representative_Docs     [sabao po 5 kg, sabao po s kg s, sabao po 1 kg...
Name: 19, dtype: object
Topico 18 - 18_kg_po_sabao_tixan - Top 5 de cada topico:
['kg', 'po', 'sabao', 'tixan', 'det', 'ype', 'maciez', 'peso', 'larg', 'pes']

Topic                                                              19
Count                                                           31732
Name                                                19_tipo_po_sx_ss
Representation          [tipo, po, sx, ss, luva, ma, abracadeira, zb, ...
Representative_Docs     [po sx unc 7s16 chv 5s8 zb, po sx 5s8 unc 1s4 ...
Name: 20, dtype: object
Topico 19 - 19_tipo_po_sx_ss - Top 5 de cada topico:
['tipo', 'po', 'sx', 'ss', 'luva', 'ma', 'abracadeira', 'zb', 'cs', 'polo']
```

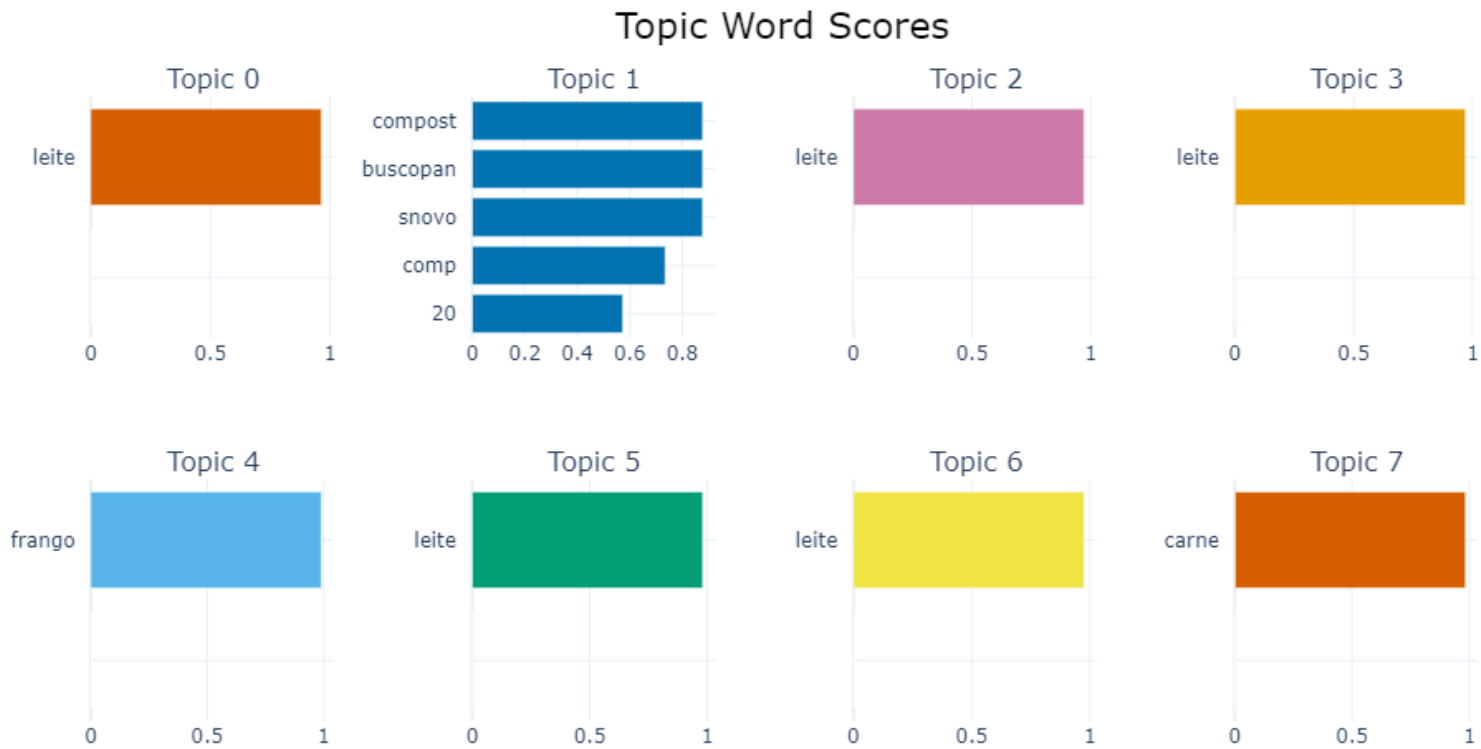
Melhor configuração - Espaço vetorial



Melhor configuração - Espaço vetorial

- Tópicos 13 e 16 (Carne), 12 (Linguiças), 4 (Frango), 17 (Ossos) e 18 (Geral) no canto inferior esquerdo;
- Tópicos 2 (Leite - Chocolate), 5 (Leite - Magnésia), 6 (Leite - Santa Clara), 8 (Leite - Chocolate), 9 (Leite - Iogurtes), 14 (Leite - Pó e Doces) e 15 (Leite - Doces) no canto superior esquerdo;
- Tópicos 0 (Leite - Pia), 1 (Remédios), 3 (Leite - Languiru), 7 (Carne - Peito), 10 (Leite - Elege), 11 (Pães) e 19 (Geral).

Melhor configuração - Escores dos tópicos

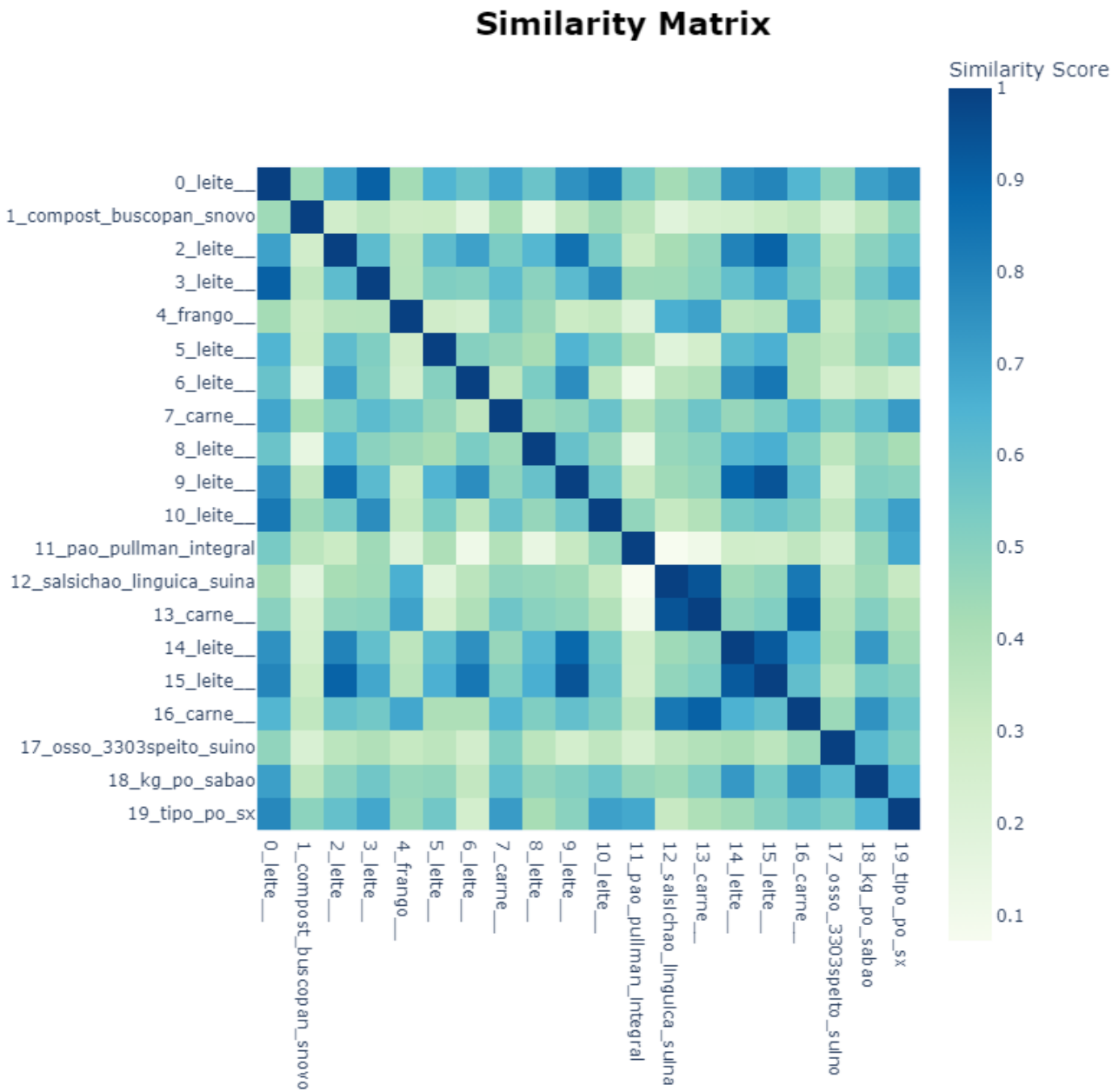


Melhor configuração - Escores dos tópicos

Melhor configuração - Escores dos tópicos

- Tópicos 11 (Pães), 18 (Geral) e 19 (Geral) são os menos específicos;
- Tópicos *"zero-shot learning"* obtiveram valores maiores que 0,94;
- Tópicos 1 (Remédios), 12 (Linguiças) e 17 (Ossos) são os mais específicos sem *"zero-shot learning"*.

Melhor configuração - Matriz de similaridade



Melhor configuração - Matriz de similaridade

- Carnes são mais similares entre si. O grupo 4 (Frango) é mais similar aos tópicos 7 (Carne - Peito), 12 (Linguiças), 13 e 16 (Carnes);
- Leites mais similares entre si;
- Grupos 1 (Remédios), 11 (Pães) e 17 (Ossos) não possuem alta similaridade;
- Grupos 18 e 19 são mais gerais.

Resultados Corpus sem refinamento

Modelo	Modelo de Embedding	Palavras Chave	Tamanho mínimo	Número de tópicos	Modelo de Representação	Tempo	Silhueta	Banco utilizado (%)
3	"p-m"	10	500	X (13)	X	8 Min	0,47	50
5	"p-m"	10	X	X (2686)	BART (1)	41 Min	0,82	100
9	"p-m"	10	500	X (10)	Key BERT I	3h 4 Min	0,57	50
10	"p-m"	10	500	X (36)	BART (1)	26 Min	0,64	100
12	**"p-m"**	**10**	**500**	**X (34)**	**BART (2)**	**27 Min**	**0,65**	**100**

Zero-Shot do Corpus sem refinamento

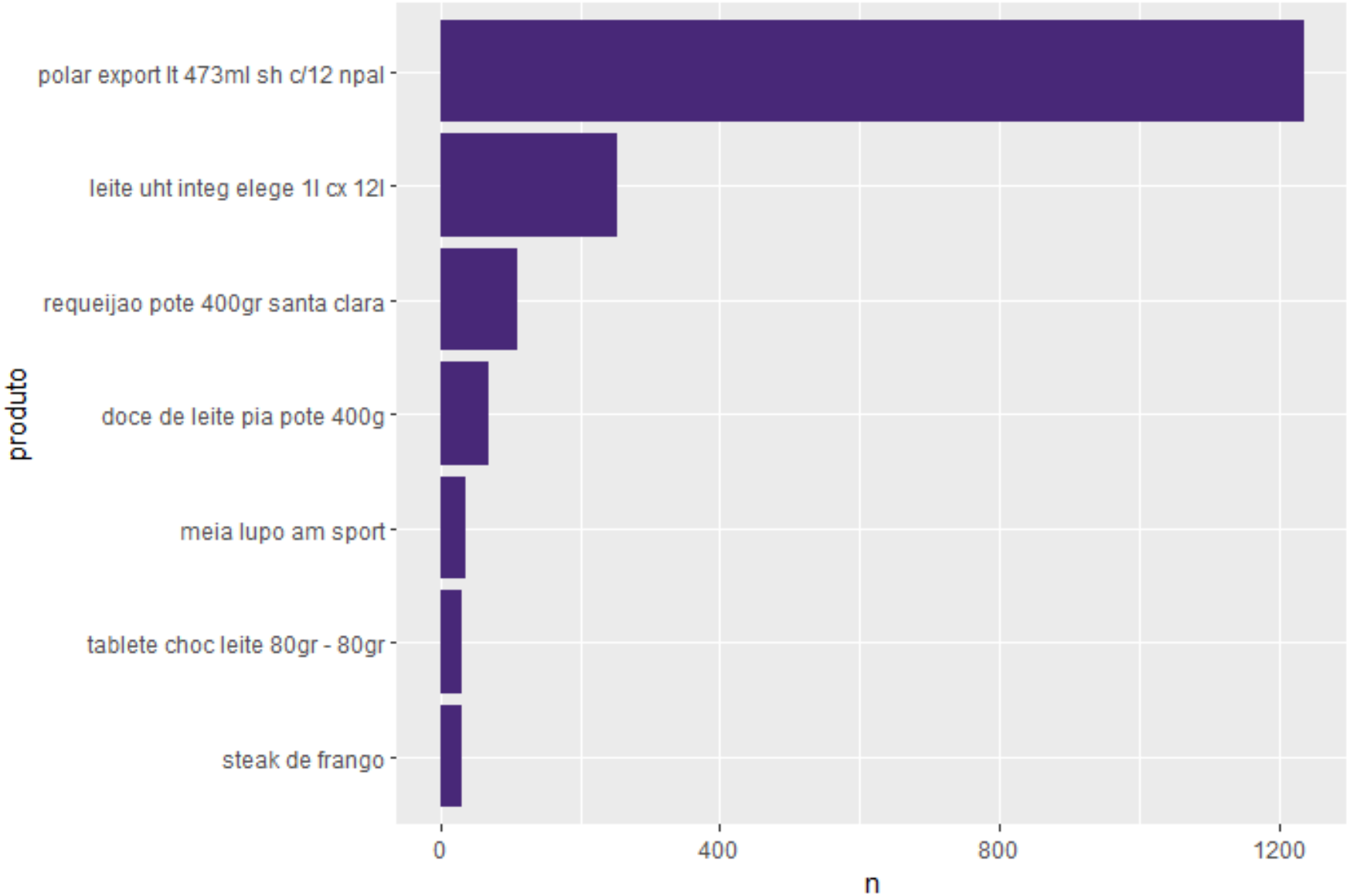
Zero Shot (2):
[“leite”, “carne”, “frango”, “shampoo”, “porta”, “linguica”, “pao”]

Melhor configuração: Características básicas

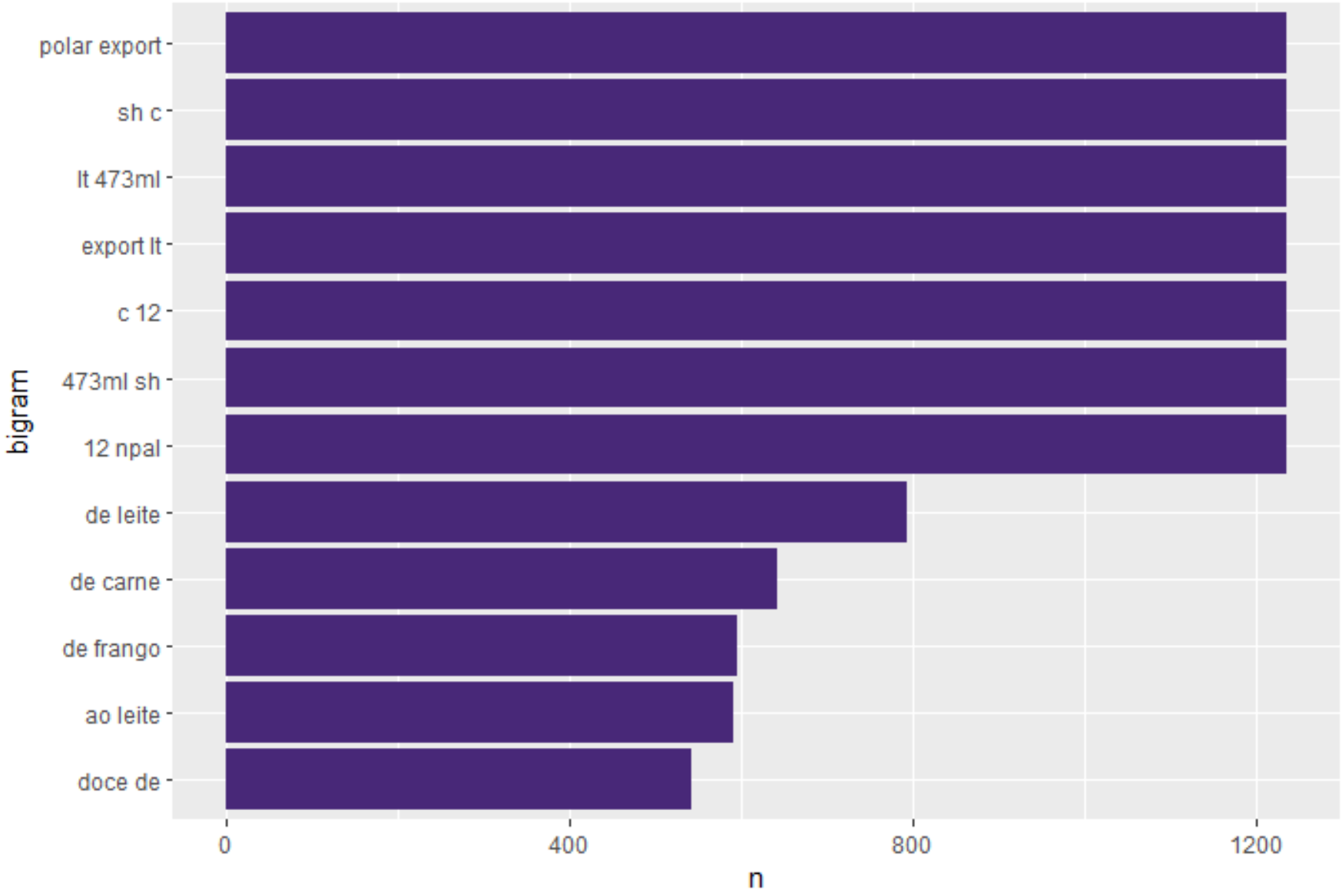
```
-1                      -1_de_leite_po_integral
0                                  0_porta___
1                    1_carne_moida_bife_cx
2                                  2_porta___
3                           3_kg_po_1kg_cx
4                                  4_porta___
5          5_poliester_100_tecido_poliuretano
6                                  6_leite___
7                        7_porca_sext_zb_ma
8                                8_shampoo___
9                      9_rosa_color_cor_pink
10          10_50mg_losartana_gen_hipoclorito
11                 11_congelado_pao_gel_gelo
12                       12_npal_473ml_sh_lt
13                  13_lampada_led_12v_polo
14             14_sport_meia_sports_esportiva
15                               15_frango___
16       16_esponja_esfrebom_multiuso_brilhus
17                               17_frango___
18                                18_porta___
19                                19_leite___
20    20_porcelana_porcelanato_caneca_grafite
21                    21_polia_poli_ceditop_ar
22                  22_grampo_26_5000_grampos
23                                23_leite___
24              24_camisa_camiseta_polo_manga
25                 25_cafe_melitta_moido_500g
26                      26_arroz_t1_lf_polido
27                27_pastel_forno_mini_torta
28    28_aerossol_spray_aerossois_desodorante
29                                29_porta___
30                30_flor_floral_flores_beija
31                         31_inox_304_ri_aco
32                      32_integ_12l_elege_uht
Name: Name, dtype: object
```

```
-1    147752
0      10316
1       4548
2       4017
3       3398
4       3204
5       2683
6       2492
7       2004
8       1801
9       1374
10      1354
11      1284
12      1237
13      1214
14      1192
15      1114
16      1057
17      1048
18       793
19       788
20       767
21       742
22       722
23       665
24       660
25       644
26       618
27       610
28       579
29       575
30       559
31       540
32       509
Name: Count, dtype: int64
```

Melhor configuração: Palavras mais frequentes



Melhor configuração: Bigrams mais frequentes



Melhor configuração - “Leite”

```
Topic                                                                  6
Count                                                               2492
Name                                                       6_leite___
Representation                                  [leite, , , , , , , , , ]
Representative_Docs    [doce de leite pia 400g, doce de leite 4.5kg, ...
Name: 7, dtype: object
Topico 6 - 6_leite___ - Top 5 de cada topico:
['leite', '', '', '', '', '', '', '', '', '']

Topic                                                                 19
Count                                                                788
Name                                                      19_leite___
Representation                                  [leite, , , , , , , , , ]
Representative_Docs    [leite uht integral, leite uht integral       ...
Name: 20, dtype: object
Topico 19 - 19_leite___ - Top 5 de cada topico:
['leite', '', '', '', '', '', '', '', '', '']

Topic                                                                 23
Count                                                                665
Name                                                      23_leite___
Representation                                  [leite, , , , , , , , , ]
Representative_Docs    [chocolate po 200g, chocolate em po 50% 5kg, c...
Name: 24, dtype: object
Topico 23 - 23_leite___ - Top 5 de cada topico:
['leite', '', '', '', '', '', '', '', '', '']


Topic                                                                 32
Count                                                                509
Name                                             32_integ_12l_elege_uht
Representation         [integ, 12l, elege, uht, 1l, 400gr, cx, requei...
Representative_Docs    [leite uht integ elege 1l cx 12l, leite uht in...
Name: 33, dtype: object
Topico 32 - 32_integ_12l_elege_uht - Top 5 de cada topico:
['integ', '12l', 'elege', 'uht', '1l', '400gr', 'cx', 'requeijao', 'santa', '80gr']
```

Melhor configuração - "Roupas"

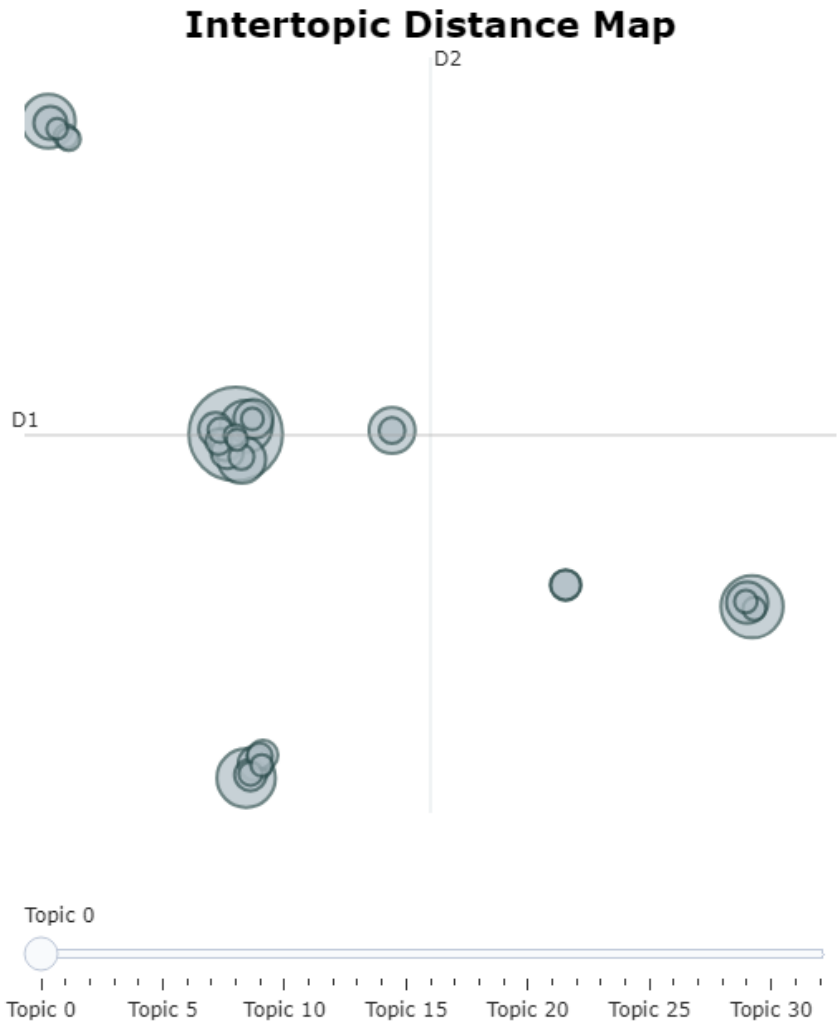
```
Topic                                                                  5
Count                                                               2683
Name                                      5_poliester_100_tecido_poliuretano
Representation          [poliester, 100, tecido, poliuretano, m2, elas...
Representative_Docs     [tecido poliester, tecido poliester beta polie...
Name: 6, dtype: object
Topico 5 - 5_poliester_100_tecido_poliuretano - Top 5 de cada topico:
['poliester', '100', 'tecido', 'poliuretano', 'm2', 'elastano', 'larg', 'algodao', 'polietileno', 'polipropilen

Topic                                                                  9
Count                                                               1374
Name                                                9_rosa_color_cor_pink
Representation          [rosa, color, cor, pink, cores, pimpolho, cora...
Representative_Docs     [hav top cor rosa porcelana 33/34, meia colori...
Name: 10, dtype: object
Topico 9 - 9_rosa_color_cor_pink - Top 5 de cada topico:
['rosa', 'color', 'cor', 'pink', 'cores', 'pimpolho', 'cora', 'glitter', 'esmalte', 'incolor']

Topic                                                                 14
Count                                                               1192
Name                                     14_sport_meia_sports_esportiva
Representation          [sport, meia, sports, esportiva, sportage, tam...
Representative_Docs     [meia lupo am sport, meia lupo am sport, meia ...
Name: 15, dtype: object
Topico 14 - 14_sport_meia_sports_esportiva - Top 5 de cada topico:
['sport', 'meia', 'sports', 'esportiva', 'sportage', 'tam', 'wg', 'lupo', 'pto', 'fila']

Topic                                                                 24
Count                                                                660
Name                                      24_camisa_camiseta_polo_manga
Representation          [camisa, camiseta, polo, manga, malha, tam, cu...
Representative_Docs                [camisa polo, camisa polo, camisa polo]
Name: 25, dtype: object
Topico 24 - 24_camisa_camiseta_polo_manga - Top 5 de cada topico:
['camisa', 'camiseta', 'polo', 'manga', 'malha', 'tam', 'curta', 'gola', 'gg', 'mc']
```

Melhor configuração - Espaço vetorial



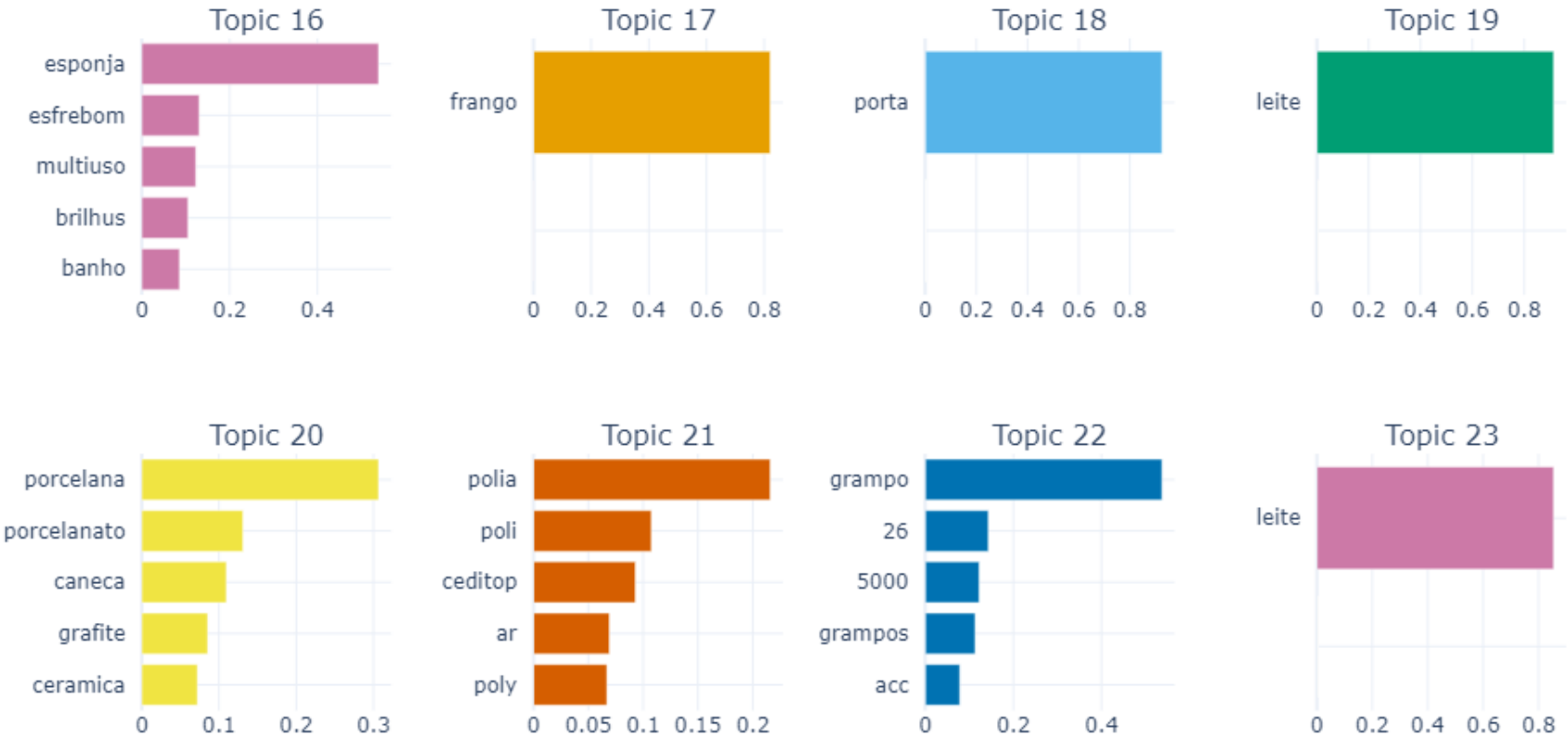
Melhor configuração - Espaço vetorial

- Grupos 3 (pós e carnes), 11 (produtos congelados), 23 (chocolates de leite), 25 (cafés) e 32 (produtos gerais de leite) no canto superior esquerdo;
- Tópicos 6 (doces de leite) e 19 (doces feitos de leite) na região central;
- Grupos 1 (carnes gerais), 7 (porcas), 26 (arroz) e 27 (pastéis) no canto inferior esquerdo.

Melhor configuração - Escores dos tópicos

Melhor configuração - Escores dos tópicos

Melhor configuração - Escores dos tópicos

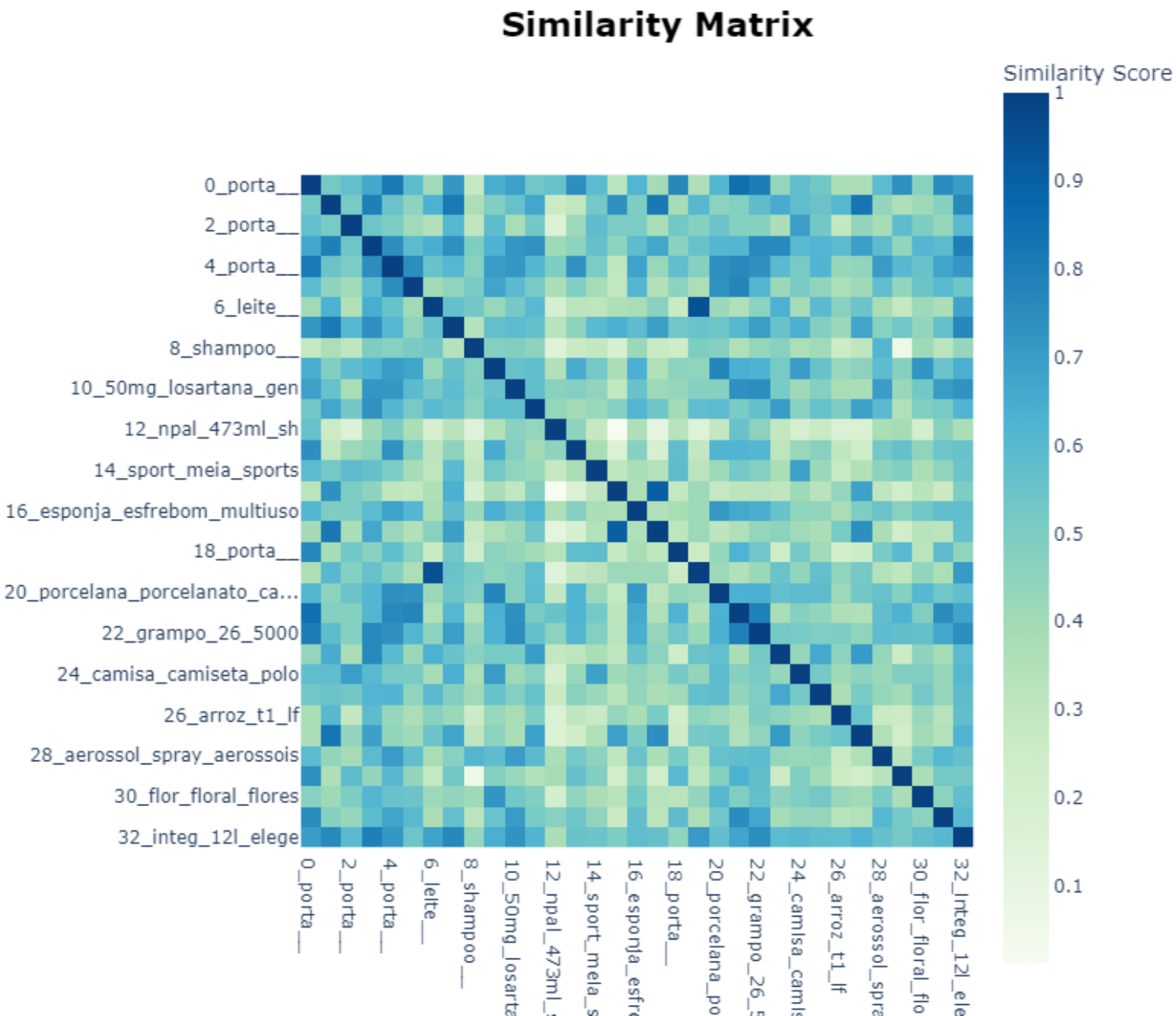


Melhor configuração - Escores dos tópicos

Melhor configuração - Escores dos tópicos

- Tópicos obtidos a partir do *"zero-shot learning"* obtiveram escores maiores que 0,80;
- Agrupamentos 16 (esponjas), 22 (grampos) e 26 (arroz) obtiveram escores maiores que 0,50.

Melhor configuração - Matriz de similaridade



Melhor configuração - Matriz de similaridade

- Grupos 6 (doces de leite) e 19 (doces feitos de leite) possuem uma similaridade muito alta (0,94);
- Grupos 1 (carne) e 27 (pastel) possuem uma similaridade de 0,83.

Conclusão

- Menor certeza para os agrupamentos do segundo corpus, além da presença de mais *outliers*.
- Agrupamentos em geral foram bem divididos, com produtos variados em ambos os corpus.
- Tópicos similares em ambos corpus, indicando robustez do modelo.

Conclusão

- Obter agora características de tópico como o preço para observar, por exemplo, fraudes.
- Comparar preços praticados no RS com àqueles do Sistema de Registro de Preços (SRP).

Trabalhos futuros

- Estudar o que poderia ser retirado nos corpus;
- Modelos de representação atualizados;
- Variar e melhorar *prompts* para os LLMs;
- Avaliar os modelos que estratificaram muito o banco de dados;
- Criar uma métrica com penalidade de número de tópicos.

Referências I

D. M. Blei and J. D. Lafferty. Correlated topic models. *Advances in Neural Information Processing Systems*, 20:147–154, 2007.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

R. J. Campello, D. Moulavi, and J. Sander. Density-based clustering by means of core samples. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 160–168. SIAM, 2013.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 226–231. AAAI Press, 1996.

M. Khadivi, S. Akbarpour, M. R. Feizi Derakhshi, and B. Anari. Persian topic detection based on human word association and graph embedding, 02 2023.

M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019.

Referências II

A. Markov. Rasprostranenie zakona bol'shih chisel na velichiny, zavisyaschie drug ot druga. *Izvestiya Fiziko-matematicheskogo obschestva pri Kazanskom universitete*, 15:135–156, 1906.

L. McInnes and J. Healy. Umap: Uniform manifold approximation and projection for dimension reduction. 02 2018.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. ISSN 0377-0427. doi: https://doi.org/10.1016/0377-0427(87)90125-7. URL `https://www.sciencedirect.com/science/article/pii/0377042787901257`.

Obrigado!