Libraries

```
import requests
from bs4 import BeautifulSoup


import spacy
```

Data Collection

Webscraping/raspagem

Utilizamos um texto coletado do website wikipedia. O texto se trata das informações sobre Ciência de Dados na língua portuguesa.

```
url = "https://pt.wikipedia.org/wiki/Ci%C3%AAncia_de_dados"

# Fazendo a requisição para a página
response = requests.get(url)

# Verificando se a requisição foi bem-sucedida (código 200)
if response.status_code == 200:
    # Parseando o conteúdo da página com BeautifulSoup
    soup = BeautifulSoup(response.text, 'html.parser')

    # Agora você pode procurar por elementos HTML e extrair informações
    # Vamos extrair o título da página como exemplo
    titulo_pagina = soup.title.text
    print(f"Título da página: {titulo_pagina}")

    # Também podemos extrair o texto do conteúdo principal da página
    conteudo_principal = soup.find('div', {'id': 'mw-content-text'})
    texto = conteudo_principal.get_text()
    print(texto)
else:
    print(f"A requisição falhou com o código de status {response.status_code}")
```

↑ Zaidi, Deena (7 de outubro de 2017). «Data Analytics in Banking». Consultado em 29 de novembro de 2017

↑ Pereira, Tiago (24 de junho de 2017). «Cientista de Dados – por onde começar em 8 passos». Consultado em 25 de novembr

↑ «50 Best Jobs in America». 2017. Consultado em 29 de novembro de 2017

↑ Zhang, Vivian (14 de abril de 2017). «3 razões pelas quais o cientista de dados continua sendo o principal emprego na

↑ Columbus, Louis (13 de maio de 2017). «IBM prevê demanda por dados Os cientistas aumentarão em 28% até 2020». Consulta

↑ Efraim Turban, Dursun Delen, Ramesh Sharda (2017). Business Intelligence, Analytics, and Data Science: A Managerial Pe

Portal da ciência Portal da matemática Portal de probabilidade e estatística Portal das tecnologias de informação

Tokenização

Aqui iniciamos o processo de limpeza do texto com os devidos tratamentos a fim de transformá-lo em tokens.

```
!python -m spacy download pt_core_news_sm
```

```
Collecting pt-core-news-sm==3.7.0
  Downloading https://github.com/explosion/spacy-models/releases/download/pt_core_news_sm-3.7.0/pt_core_news_sm-3.7.0-py
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 13.0/13.0 MB 27.0 MB/s eta 0:00:00
Requirement already satisfied: spacy<3.8.0,>=3.7.0 in /usr/local/lib/python3.10/dist-packages (from pt-core-news-sm==3.7
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in /usr/local/lib/python3.10/dist-packages (from spacy<3.8.0,
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.8.0,
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.8.0,>=
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in /usr/local/lib/python3.10/dist-packages (from spacy<3.8.0,>=3.7.0-
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in /usr/local/lib/python3.10/dist-packages (from spacy<3.8.0,>=3.7.
Requirement already satisfied: thinc<8.3.0,>=8.1.8 in /usr/local/lib/python3.10/dist-packages (from spacy<3.8.0,>=3.7.0-
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in /usr/local/lib/python3.10/dist-packages (from spacy<3.8.0,>=3.7.0
Requirement already satisfied: srsly<3.0.0,>=2.4.3 in /usr/local/lib/python3.10/dist-packages (from spacy<3.8.0,>=3.7.0-
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in /usr/local/lib/python3.10/dist-packages (from spacy<3.8.0,>=3.
Requirement already satisfied: weasel<0.4.0,>=0.1.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.8.0,>=3.7.0
Requirement already satisfied: typer<0.10.0,>=0.3.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.8.0,>=3.7.0
Requirement already satisfied: smart-open<7.0.0,>=5.2.1 in /usr/local/lib/python3.10/dist-packages (from spacy<3.8.0,>=3
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.8.0,>=3.7.0-
Requirement already satisfied: requests<3.0.0,>=2.13.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.8.0,>=3.
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4 in /usr/local/lib/python3.10/dist-packages (from spa
Requirement already satisfied: jinja2 in /usr/local/lib/python3.10/dist-packages (from spacy<3.8.0,>=3.7.0->pt-core-news
Requirement already satisfied: setuptools in /usr/local/lib/python3.10/dist-packages (from spacy<3.8.0,>=3.7.0->pt-core-
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.8.0,>=3.7.0->pt-
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.8.0,>=3.
Requirement already satisfied: numpy>=1.19.0 in /usr/local/lib/python3.10/dist-packages (from spacy<3.8.0,>=3.7.0->pt-co
Requirement already satisfied: annotated-types>=0.4.0 in /usr/local/lib/python3.10/dist-packages (from pydantic!=1.8,!=1
Requirement already satisfied: pydantic-core==2.16.1 in /usr/local/lib/python3.10/dist-packages (from pydantic!=1.8,!=1.
Requirement already satisfied: typing-extensions>=4.6.1 in /usr/local/lib/python3.10/dist-packages (from pydantic!=1.8,
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages (from requests<3.0.0,
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests<3.0.0,>=2.13.0->sp
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.10/dist-packages (from requests<3.0.0,>=2.13
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-packages (from requests<3.0.0,>=2.13
Requirement already satisfied: blis<0.8.0,>=0.7.8 in /usr/local/lib/python3.10/dist-packages (from thinc<8.3.0,>=8.1.8->
Requirement already satisfied: confection<1.0.0,>=0.0.1 in /usr/local/lib/python3.10/dist-packages (from thinc<8.3.0,>=8
Requirement already satisfied: click<9.0.0,>=7.1.1 in /usr/local/lib/python3.10/dist-packages (from typer<0.10.0,>=0.3.0
Requirement already satisfied: cloudpathlib<0.17.0,>=0.7.0 in /usr/local/lib/python3.10/dist-packages (from weasel<0.4.0
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.10/dist-packages (from jinja2->spacy<3.8.0,>=3.
Installing collected packages: pt-core-news-sm
Successfully installed pt-core-news-sm-3.7.0
✔ Download and installation successful
You can now load the package via spacy.load('pt_core_news_sm')
```

```
# definindo modelo de linguagem em português nível médio
nlp = spacy.load("pt_core_news_sm")

# Carregar o modelo de linguagem em português do spaCy
nlp = spacy.load('pt_core_news_sm')

# Aplicar o modelo ao texto
doc = nlp(texto)

# Inicializar a lista de tokens
tokens = []

# Iterar sobre os tokens no documento
for token in doc:
    # Verificar se o token é um verbo, é alfabético e não é uma stop word
    if token.pos_ == "VERB" and token.is_alpha and not token.is_stop:
        tokens.append(str.lower(token.lemma_))
    elif not token.pos_ == "VERB" and token.is_alpha and not token.is_stop:
        tokens.append(str.lower(token.text))

# Exibir os tokens resultantes
print(tokens)
```

```
['ciência', 'dados', 'inglês', 'data', 'science', 'interdisciplinar', 'interface', 'estatística', 'ciência', 'computação
```

```
len(tokens)
```

```
1119
```

Modelagem

Instalação e importação da biblioteca BERTopic

```
!pip install bertopic
```

```
Collecting bertopic
  Downloading bertopic-0.16.0-py2.py3-none-any.whl (154 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 154.1/154.1 kB 4.9 MB/s eta 0:0
Requirement already satisfied: numpy>=1.20.0 in /usr/local/lib/python3.10/dis
Collecting hdbscan>=0.8.29 (from bertopic)
  Downloading hdbscan-0.8.33.tar.gz (5.2 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 5.2/5.2 MB 23.2 MB/s eta 0:00:0
  Installing build dependencies ... done
  Getting requirements to build wheel ... done
  Preparing metadata (pyproject.toml) ... done
Collecting umap-learn>=0.5.0 (from bertopic)
  Downloading umap-learn-0.5.5.tar.gz (90 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 90.9/90.9 kB 14.5 MB/s eta 0:00
  Preparing metadata (setup.py) ... done
Requirement already satisfied: pandas>=1.1.5 in /usr/local/lib/python3.10/dis
Requirement already satisfied: scikit-learn>=0.22.2.post1 in /usr/local/lib/p
Requirement already satisfied: tqdm>=4.41.1 in /usr/local/lib/python3.10/dist
Collecting sentence-transformers>=0.4.1 (from bertopic)
  Downloading sentence_transformers-2.3.1-py3-none-any.whl (132 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 132.8/132.8 kB 21.0 MB/s eta 0:
Requirement already satisfied: plotly>=4.7.0 in /usr/local/lib/python3.10/dis
Collecting cython<3,>=0.27 (from hdbscan>=0.8.29->bertopic)
  Using cached Cython-0.29.37-cp310-cp310-manylinux_2_17_x86_64.manylinux2014
Requirement already satisfied: scipy>=1.0 in /usr/local/lib/python3.10/dist-p
Requirement already satisfied: joblib>=1.0 in /usr/local/lib/python3.10/dist-
Requirement already satisfied: python-dateutil>=2.8.1 in /usr/local/lib/pytho
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist
Requirement already satisfied: tenacity>=6.2.0 in /usr/local/lib/python3.10/d
Requirement already satisfied: packaging in /usr/local/lib/python3.10/dist-pa
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3
Requirement already satisfied: transformers<5.0.0,>=4.32.0 in /usr/local/lib/
Requirement already satisfied: torch>=1.11.0 in /usr/local/lib/python3.10/dis
Requirement already satisfied: nltk in /usr/local/lib/python3.10/dist-package
Requirement already satisfied: sentencepiece in /usr/local/lib/python3.10/dis
Requirement already satisfied: huggingface-hub>=0.15.1 in /usr/local/lib/pyth
Requirement already satisfied: Pillow in /usr/local/lib/python3.10/dist-packa
Requirement already satisfied: numba>=0.51.2 in /usr/local/lib/python3.10/dis
Collecting pynndescent>=0.5 (from umap-learn>=0.5.0->bertopic)
  Downloading pynndescent-0.5.11-py3-none-any.whl (55 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 55.8/55.8 kB 9.9 MB/s eta 0:00:
Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-pac
Requirement already satisfied: fsspec>=2023.5.0 in /usr/local/lib/python3.10/
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-pac
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.10/dist-
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/p
Requirement already satisfied: llvmlite<0.42,>=0.41.0dev0 in /usr/local/lib/p
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-pac
Requirement already satisfied: sympy in /usr/local/lib/python3.10/dist-packag
Requirement already satisfied: networkx in /usr/local/lib/python3.10/dist-pac
Requirement already satisfied: jinja2 in /usr/local/lib/python3.10/dist-packa
Requirement already satisfied: triton==2.1.0 in /usr/local/lib/python3.10/dis
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.10
Requirement already satisfied: tokenizers<0.19,>=0.14 in /usr/local/lib/pytho
Requirement already satisfied: safetensors>=0.3.1 in /usr/local/lib/python3.1
Requirement already satisfied: click in /usr/local/lib/python3.10/dist-packag
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.10/d
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/pyt
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.1
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.1
Requirement already satisfied: mpmath>=0.19 in /usr/local/lib/python3.10/dist
Building wheels for collected packages: hdbscan, umap-learn
  Building wheel for hdbscan (pyproject.toml) ... done
  Created wheel for hdbscan: filename=hdbscan-0.8.33-cp310-cp310-linux_x86_64
  Stored in directory: /root/.cache/pip/wheels/75/0b/3b/dc4f60b7cc455efaefb62
  Building wheel for umap-learn (setup.py) ... done
  Created wheel for umap-learn: filename=umap_learn-0.5.5-py3-none-any.whl si
  Stored in directory: /root/.cache/pip/wheels/3a/70/07/428d2b58660a1a3b431db
Successfully built hdbscan umap-learn
Installing collected packages: cython, pynndescent, hdbscan, umap-learn, sent
  Attempting uninstall: cython
    Found existing installation: Cython 3.0.8
    Uninstalling Cython-3.0.8:
      Successfully uninstalled Cython-3.0.8
Successfully installed bertopic-0.16.0 cython-0.29.37 hdbscan-0.8.33 pynndesc
WARNING: The following packages were previously imported in this runtime:
  [Cython,cython]
You must restart the runtime in order to use newly installed versions.
```

RESTART SESSION

```
from bertopic import BERTopic
```

Definição do modelo

Como utilizamos a página do wikipedia em português, foi necessário definir a língua portuguesa para instanciar o modelo. Também utilizamos um embedding_model pronto, em vez de nenhum, para reduzir o tempo de ajuste do modelo.

```
# instantiate model
topic_model = BERTopic(language="portuguese", embedding_model="all-MiniLM-L6-v2")
```

```
# fit and transform
topics, probs = topic_model.fit_transform(tokens)
```

```
modules.json: 100%                                349/349 [00:00<00:00, 14.5kB/s]
config_sentence_transformers.json: 100%           116/116 [00:00<00:00, 7.34kB/s]
README.md: 100%                                   10.6k/10.6k [00:00<00:00, 458kB/s]
sentence_bert_config.json: 100%                   53.0/53.0 [00:00<00:00, 2.86kB/s]
config.json: 100%                                 612/612 [00:00<00:00, 26.9kB/s]
pytorch_model.bin: 100%                           90.9M/90.9M [00:00<00:00, 256MB/s]
tokenizer_config.json: 100%                       350/350 [00:00<00:00, 22.0kB/s]
vocab.txt: 100%                                   232k/232k [00:00<00:00, 1.93MB/s]
```

Model Visualization

Informações sobre os tópicos

```
#topics
```

```
probs
```

```
array([1.        , 0.86631835, 0.        , ..., 1.        , 0.        ,
       0.        ])
```

Com o comando abaixo, nós entendemos a essência de saída do modelo. Vemos cada tópico/cluster, a quantidade de tokens de cada tópico e sua representação indicando os tokens que estão contidos em cada tópico/cluster.

- O valor -1 indica outliers. Logo, temos 289 outliers.

```
topic_model.get_topic_info()
```

	Topic	Count	Name	Representation
0	-1	289	-1_empresas_outubro_negócios_programação	[empresas, outubro, negócios, programação, pro...
1	0	135	0_association_business_sites_principal	[association, business, sites, principal, jour...
2	1	60	1_usuário_nate_ramo_william	[usuário, nate, ramo, william, sendo, maio, fa...
3	2	50	2_the_of_in_big	[the, of, in, big, and, to, what, gap, good, o...
4	3	43	3_editar_transformar_tomar_mudar	[editar, transformar, tomar, mudar, aumentar, ...
5	4	39	4_data_statistical_volume_analytics	[data, statistical, volume, analytics, statist...
6	5	34	5_dados_contextos__	[dados, contextos, , , , , , ,]
7	6	29	6_ciência_ciências_referências_	[ciência, ciências, referências, , , , , ,]
8	7	23	7_utilizar_melhores_melhor_otimizar	[utilizar, melhores, melhor, otimizar, úteis, ...
9	8	23	8_abril_aplicações_acordo_ações	[abril, aplicações, acordo, ações, aplicar, ab...
10	9	23	9_gerar_principais_produtos_procurar	[gerar, principais, produtos, procurar, probab...
11	10	22	10_dados_genéticos_famílias_dias	[dados, genéticos, famílias, dias, , , , , ,]
12	11	20	11_aprendizado_auxiliar_aprendizagem_suficientes	[aprendizado, auxiliar, aprendizagem, suficien...
13	12	20	12_consultar_viagens__	[consultar, viagens, , , , , , ,]
14	13	20	13_science_buzzword_statement_artificial	[science, buzzword, statement, artificial, , ,...
15	14	20	14_cientistas_cientista_artigo_mecanismos	[cientistas, cientista, artigo, mecanismos, in...

```
topic_model.get_topic(0)
```

```
[('association', 0.06510295043959169),
 ('business', 0.06510295043959169),
 ('sites', 0.054594128284917866),
 ('principal', 0.054594128284917866),
 ('journal', 0.054594128284917866),
 ('intelligence', 0.054594128284917866),
 ('insights', 0.054594128284917866),
 ('academia', 0.04197353102305506),
 ('forbes', 0.04197353102305506),
 ('buscas', 0.04197353102305506)]
```

```
topic_model.get_representative_docs(0)
```

```
['business', 'association', 'business']
```

```
top = topic_model.get_topic_info()
```

```
for item in top.values:
    idTopico = item[0]
    contagem = item[1]
    conteudo = item[2]

    print("#########################")
    print("ID do tópico:", idTopico)
    print("Contagem do tópico:", contagem)
    print("Conteúdo do tópico:", conteudo)
```

```
Contagem do tópico: 17
Conteúdo do tópico: 18_serviços_sociais_hoje_virtuais
#########################
ID do tópico: 19
Contagem do tópico: 17
Conteúdo do tópico: 19_algoritmos_lançar_algoritmo_últimos
#########################
ID do tópico: 20
Contagem do tópico: 16
Conteúdo do tópico: 20_anos_tipos_voos_passos
#########################
ID do tópico: 21
Contagem do tópico: 16
Conteúdo do tópico: 21_sugestões_suporte_sistemas_sugerir
#########################
ID do tópico: 22
Contagem do tópico: 16
Conteúdo do tópico: 22_monitora_matemática_problemas_lista
#########################
ID do tópico: 23
Contagem do tópico: 15
Conteúdo do tópico: 23_extrair_existir_substituir_subir
#########################
ID do tópico: 24
Contagem do tópico: 14
Conteúdo do tópico: 24_capacidade_passado_necessidade_habilidades
#########################
ID do tópico: 25
Contagem do tópico: 14
Conteúdo do tópico: 25_estatística_estatístico_estatísticas_
#########################
ID do tópico: 26
Contagem do tópico: 14
Conteúdo do tópico: 26_novembro_setembro__
#########################
ID do tópico: 27
Contagem do tópico: 13
Conteúdo do tópico: 27_segundos_semelhanças_seguro_segurança
#########################
ID do tópico: 28
Contagem do tópico: 13
Conteúdo do tópico: 28_atualmente_similares_anteriormente_similarmente
#########################
ID do tópico: 29
Contagem do tópico: 13
Conteúdo do tópico: 29_conteúdo_complementares_conversar_continuar
#########################
ID do tópico: 30
Contagem do tópico: 13
Conteúdo do tópico: 30_termo_técnicas_tecnologias_teoria
#########################
ID do tópico: 31
Contagem do tópico: 12
Conteúdo do tópico: 31_reconhecimento_conhecimento_aumento_surgimento
#########################
ID do tópico: 32
Contagem do tópico: 11
Conteúdo do tópico: 32_análise_analistas_systematic_approaches
```

Visualizações

Distância 2D entre tópicos/clusters

Abaixo, a distância que os tópicos estão entre eles em plano 2D.

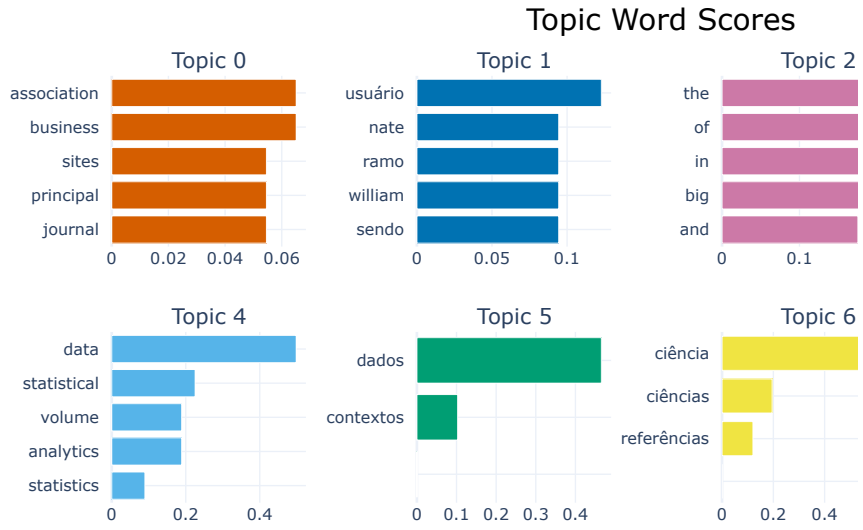
```
Quanto maior o círculo, maior o cluster em número de tokens.
Quanto mais próximos os círculos estão entre si, mais semanticamente similar são as palavras que compõem os círculos.
```

```
topic_model.visualize_topics()
```

Gráfico de barras dos tópicos/clusters

Abaixo, os tópicos agrupados por barra.

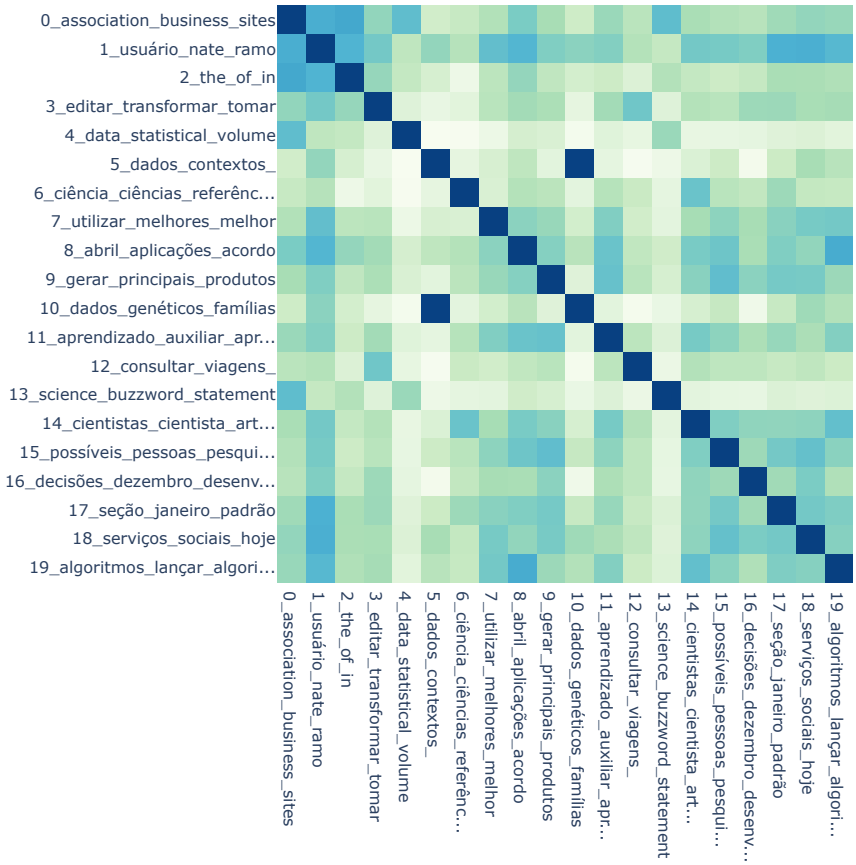
```
topic_model.visualize_barchart()
```



Matriz de correlações dos tópicos/clusters

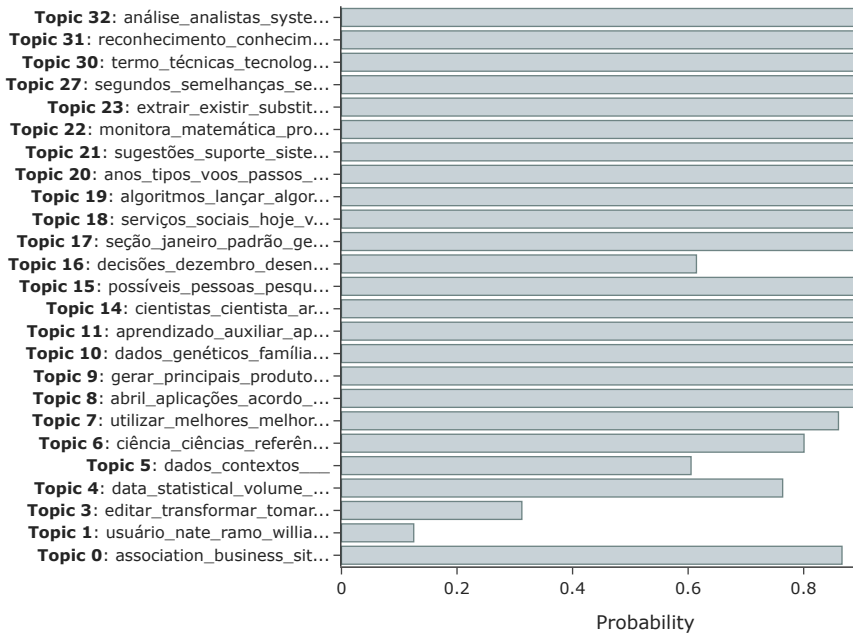
Eu quero ver os top 20 tópicos em um heatmap. Assim, conseguimos ver a correlação entre os tópicos.

```
topic_model.visualize_heatmap(top_n_topics=20)
```



Distribuição de probabilidade

```
topic_model.visualize_distribution(probs)
```



```
topic_model.visualize_hierarchy()
```