

Homework 4 Build Topics Using BERTopic.

BERTopic is a topic modeling technique that leverages 🧠 transformers and c-TF-IDF

(<https://maartengr.github.io/BERTopic/api/ctfidf.html>) to create dense clusters allowing for easily interpretable topics whilst keeping important words in the topic descriptions.

Load the wine reievw data and train a BerTopic model. Answer all questions within this notebook. Upload notebook to Moodle Homework 4. Sample code is provided along with references to complet the assignment. You can work in groups but each studen should make a their own submission.

Student Name: Frances LeMond-Glasser April, 2025

✓ Load in Data

```
## Install BerTopic https://maartengr.github.io/BERTopic/index.html
import sys
!{sys.executable} -m pip install BERTopic
```

```
Requirement already satisfied: BERTopic in /usr/local/lib/python3.11/dist-packages (0.17.0)
Requirement already satisfied: hdbscan>=0.8.29 in /usr/local/lib/python3.11/dist-packages (from BERTopic) (0.8.40)
Requirement already satisfied: numpy>=1.20.0 in /usr/local/lib/python3.11/dist-packages (from BERTopic) (2.0.2)
Requirement already satisfied: pandas>=1.1.5 in /usr/local/lib/python3.11/dist-packages (from BERTopic) (2.2.2)
Requirement already satisfied: plotly>=4.7.0 in /usr/local/lib/python3.11/dist-packages (from BERTopic) (5.24.1)
Requirement already satisfied: scikit-learn>=1.0 in /usr/local/lib/python3.11/dist-packages (from BERTopic) (1.6.1)
Requirement already satisfied: sentence-transformers>=0.4.1 in /usr/local/lib/python3.11/dist-packages (from BERTopic) (3.4.1)
Requirement already satisfied: tqdm>=4.41.1 in /usr/local/lib/python3.11/dist-packages (from BERTopic) (4.67.1)
Requirement already satisfied: umap-learn>=0.5.0 in /usr/local/lib/python3.11/dist-packages (from BERTopic) (0.5.7)
Requirement already satisfied: scipy>=1.0 in /usr/local/lib/python3.11/dist-packages (from hdbscan>=0.8.29->BERTopic) (1.14.1)
Requirement already satisfied: joblib>=1.0 in /usr/local/lib/python3.11/dist-packages (from hdbscan>=0.8.29->BERTopic) (1.4.2)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.11/dist-packages (from pandas>=1.1.5->BERTopic) (2.8)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-packages (from pandas>=1.1.5->BERTopic) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-packages (from pandas>=1.1.5->BERTopic) (2025.2)
Requirement already satisfied: tenacity>=6.2.0 in /usr/local/lib/python3.11/dist-packages (from plotly>=4.7.0->BERTopic) (9.1.2)
Requirement already satisfied: packaging in /usr/local/lib/python3.11/dist-packages (from plotly>=4.7.0->BERTopic) (24.2)
Requirement already satisfied: threadpoolctl>=3.1.0 in /usr/local/lib/python3.11/dist-packages (from scikit-learn>=1.0->BERTopic) (3)
Requirement already satisfied: transformers<5.0.0,>=4.41.0 in /usr/local/lib/python3.11/dist-packages (from sentence-transformers>=0.4.1->BERTopic) (4.41.0)
Requirement already satisfied: torch>=1.11.0 in /usr/local/lib/python3.11/dist-packages (from sentence-transformers>=0.4.1->BERTopic) (2.5.1)
Requirement already satisfied: huggingface-hub>=0.20.0 in /usr/local/lib/python3.11/dist-packages (from sentence-transformers>=0.4.1->BERTopic) (0.20.0)
Requirement already satisfied: Pillow in /usr/local/lib/python3.11/dist-packages (from sentence-transformers>=0.4.1->BERTopic) (11.1)
Requirement already satisfied: numba>=0.51.2 in /usr/local/lib/python3.11/dist-packages (from umap-learn>=0.5.0->BERTopic) (0.60.0)
Requirement already satisfied: pynndescent>=0.5 in /usr/local/lib/python3.11/dist-packages (from umap-learn>=0.5.0->BERTopic) (0.5.1)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.20.0->sentence-transformers) (3.16.0)
Requirement already satisfied: fsspec>=2023.5.0 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.20.0->sentence-transformers) (2025.2)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.20.0->sentence-transformers) (6.0.2)
Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.20.0->sentence-transformers) (2.32.3)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.20.0->sentence-transformers) (4.12.0)
Requirement already satisfied: llvmlite<0.44,>=0.43.0dev0 in /usr/local/lib/python3.11/dist-packages (from numba>=0.51.2->umap-learn) (0.44.0)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.8.2->pandas) (1.16.0)
Requirement already satisfied: networkx in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (3.4.2)
Requirement already satisfied: Jinja2 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (3.1.4)
Requirement already satisfied: nvidia-cuda-nvrtc-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (12.4.127)
Requirement already satisfied: nvidia-cuda-runtime-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (12.4.127)
Requirement already satisfied: nvidia-cuda-cupti-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (12.4.127)
Requirement already satisfied: nvidia-cudnn-cu12==9.1.0.70 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (9.1.0.70)
Requirement already satisfied: nvidia-cublas-cu12==12.4.5.8 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (12.4.5.8)
Requirement already satisfied: nvidia-cufft-cu12==11.2.1.3 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (11.2.1.3)
Requirement already satisfied: nvidia-curand-cu12==10.3.5.147 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (10.3.5.147)
Requirement already satisfied: nvidia-cusolver-cu12==11.6.1.9 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (11.6.1.9)
Requirement already satisfied: nvidia-cusparselt-cu12==0.6.2 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (0.6.2)
Requirement already satisfied: nvidia-cusparse-cu12==12.3.1.170 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (12.3.1.170)
Requirement already satisfied: nvidia-cusparselt-cu12==0.6.2 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (0.6.2)
Requirement already satisfied: nvidia-nccl-cu12==2.21.5 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (2.21.5)
Requirement already satisfied: nvidia-nvtx-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (12.4.127)
Requirement already satisfied: nvidia-nvjitlink-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (12.4.127)
Requirement already satisfied: triton==3.2.0 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (3.2.0)
Requirement already satisfied: sympy==1.13.1 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transformers) (1.13.1)
Requirement already satisfied: mpmath<1.4,>=1.1.0 in /usr/local/lib/python3.11/dist-packages (from sympy==1.13.1->torch) (1.13.1)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.11/dist-packages (from transformers<5.0.0,>=4.41.0->BERTopic) (2024.11.6)
Requirement already satisfied: tokenizers<0.22,>=0.21 in /usr/local/lib/python3.11/dist-packages (from transformers<5.0.0,>=4.41.0->BERTopic) (0.21.0)
Requirement already satisfied: safetensors>=0.4.3 in /usr/local/lib/python3.11/dist-packages (from transformers<5.0.0,>=4.41.0->BERTopic) (0.5.0)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.11/dist-packages (from Jinja2->torch) (3.0.2)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests->huggingface-hub) (3.4.0)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-packages (from requests->huggingface-hub) (3.10.1)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests->huggingface-hub) (2.3.0)
```

```
## Other packages
import pandas as pd
from bertopic import BERTopic
from sentence_transformers import SentenceTransformer, util
from umap import UMAP
from hdbscan import HDBSCAN
from sklearn.feature_extraction.text import CountVectorizer
from bertopic.vectorizers import ClassTfidfTransformer
import numpy as np
```

```
#create a pandas dataframe
df = pd.read_csv('winemag-data-130k-v2.csv')

#check the first 5 dataframe rows
df.head(5)
```

Unnamed: 0	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle	title	
0	0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	NaN	Kerin O'Keefe	@kerinokeefe	Nico 2013 Vu Bian (Etr
1	1	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87	15.0	Douro	NaN	NaN	Roger Voss	@vossroger	Quinta c Avidag 2C Avidag R (Dou
2	2	US	Tart and snappy, the flavors of lime flesh and...	NaN	87	14.0	Oregon	Willamette Valley	Willamette Valley	Paul Gregutt	@paulgwine	Rainsto 2013 Pii C (Willame Vall
3	3	US	Pineapple rind, lemon pith and orange blossom ...	Reserve Late Harvest	87	13.0	Michigan	Lake Michigan Shore	NaN	Alexander Peartree	NaN	St. Jul 2C Reser L: Harv Riesling
Much like the ...												


Next steps:

Generate code with df

View recommended plots

New interactive sheet

```
# Modify this code to read only wines from Italy
df = df.loc[df['country'] == 'Italy']
df.head(5)
```



	Unnamed: 0	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle	title
0	0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	NaN	Kerin O'Keefe	@kerinokeefe	Nicos 20' Vull Bian (Etn
6	6	Italy	Here's a bright, informal red that opens with ...	Belsito	87	16.0	Sicily & Sardinia	Vittoria	NaN	Kerin O'Keefe	@kerinokeefe	Terre Giur 20' Belsi Frappa (Vittori
13	13	Italy	This is dominated by oak and oak-driven aromas...	Rosso	87	NaN	Sicily & Sardinia	Etna	NaN	Kerin O'Keefe	@kerinokeefe	Masser Settepor 20' Ros: (Etn
22	22	Italy	Delicate aromas recall white flower and citrus...	Ficiligno	87	19.0	Sicily & Sardinia	Sicilia	NaN	Kerin O'Keefe	@kerinokeefe	Baglio Pianet 20(Ficiligr Whi (Sicili

Aromas of

```
#create an array of descriptions
docs = df.description.values
```

Begin with BERTopic:

<https://python.plainenglish.io/topic-modeling-for-beginners-using-bertopic-and-python-aaf1b421afeb> The algorithm has five primary steps:

- 1. Extract embeddings - Uses the sentence transformer library
- 2. Reduce dimensionality - Uses the UMAP aglorithm
- 3. Cluster reduced embeddings - Uses HBSCAN to form clustes of different shapes
- 4. Tokenize topics - Uses bag of words
- 5. Extract topic words - class bsaed TF-IDF

For this homework you will train an automated BerTopic model and later perform each of these steps individually. Use the last model for the visulizations.

Research Questions -- Enter you answers within this markdown cell. 10 points out the 60.

1. Do language models like BERTopic require applying stop words. Yes or No? Why?

Yes. because if we dont they end up in the outliers, 'topic -1', or cloud the meaningful words we want the model to analyse.

2. Roughly how many parameters does GPT4 have?

1.76 Trillion. Source: <https://en.wikipedia.org/wiki/GPT-4#:~:text=Rumors%20claim%20that%20GPT%2D4,running%20and%20by%20George%20Hotz.>

3. The python package for imputaton in Sklearn is SimpleImputer. True or False?

True!

4. What is the difference between stemming and lematization?

Stemming is reducing words ending in -ing, -ed, -s which is less accurate but faster than using a lematization technique which reduces words to their real root word in the dictionary using words near the word being analyzed to account for the part of speech and context of the word.

5. Is topic modeling an unsupervised machine learning technique? Yes or No?

Yes! This is because we don't have topic labels we want to train the model to predict/classify. Instead the model is generating it's own classficiation of the topics.

```
#instantiate BERTopic
topic_model = BERTopic(language="english", calculate_probabilities=True,
                        verbose=True)

# Generate the topics --
topics, probs = topic_model.fit_transform(docs)

2025-04-10 01:34:25,190 - BERTopic - Embedding - Transforming documents to embeddings.
Batches: 100% 313/313 [05:32<00:00, 2.22it/s]
2025-04-10 01:39:59,264 - BERTopic - Embedding - Completed ✓
2025-04-10 01:39:59,266 - BERTopic - Dimensionality - Fitting the dimensionality reduction algorithm
2025-04-10 01:40:09,297 - BERTopic - Dimensionality - Completed ✓
2025-04-10 01:40:09,299 - BERTopic - Cluster - Start clustering the reduced embeddings
2025-04-10 01:40:15,334 - BERTopic - Cluster - Completed ✓
2025-04-10 01:40:15,342 - BERTopic - Representation - Fine-tuning topics using representation models.
```

Topic Output

```
topic_model.get_topic_info()

# How many topics are there?
# There are 0,1,2,...96 topics plus the outliers and stopwords in -1, in
# total there are 98. There are also 98 rows so 98 individual topics identified.
```

	Topic	Count	Name	Representation	Representative_Docs
0	-1	1457	-1_wine_is_with_and	[wine, is, with, and, of, the, this, in, that,...	[Here's a blend of Chardonnay, Sauvignon Blanc...
1	0	1898	0_tannins_cherry_red_palate	[tannins, cherry, red, palate, drink, berry, b...	[Enticing aromas of blue flower, ripe berry, l...
2	1	653	1_apple_yellow_white_pear	[apple, yellow, white, pear, citrus, peach, ac...	[This opens with delicate orchard fruit, citru...
3	2	537	2_nose_lead_tannins_palate	[nose, lead, tannins, palate, the, alongside, ...	[Toasted oak, espresso, clove and dark berry a...
4	3	505	3_cabernet_sauvignon_merlot_blend	[cabernet, sauvignon, merlot, blend, franc, sa...	[A blend of Sangiovese, with 10% Cabernet Sauv...
...
93	92	12	92_zibibbo_dessert_apricot_grapes	[zibibbo, dessert, apricot, grapes, fig, 100, ...	[Made with 100% dried Zibibbo grapes, this ele...
94	93	11	93_roero_key_rendered_winning	[roero, key, rendered, winning, arneis, ensemb...	[Streamlined and bright, this offers the aroma...
95	94	11	94_fried_trebbiano_seafood_vongole	[fried, trebbiano, seafood, vongole, con, verm...	[Straightforward and bright, this is the perfe...

What does the -1 topic refer to?

The -1 Topic is meant to recognize the outliers, or in this case because BERTopic doesn't get rid of stop words/most common words with less meaning ('with','is', 'of',...) are in the -1 Topic.

Interpret 'n' Topics

```
topic_model.get_topic(14) #TOPIC 14 looks like it's about a sparkling white wine

# This topic is about a sparkling white wine with a palette of apple, pear, and
# peach flavor. From the adjectives provided, it seems to have some acidity with
# crisp citrusy smell or flavor profile.

[('sparkler', np.float64(0.08363027524799128)),
 ('bubbles', np.float64(0.028501739071226433)),
 ('apple', np.float64(0.028441509632721077)),
 ('pear', np.float64(0.019430628763348383)),
 ('peach', np.float64(0.01832446975348749)),
```

```

('citrus', np.float64(0.01743262330024686)),
('perlage', np.float64(0.015562322029397974)),
('crisp', np.float64(0.014718575002264441)),
('green', np.float64(0.014153575461125861)),
('acidity', np.float64(0.01396885371853972))]]

topic_model.get_topic(40) #TOPIC 40

# This wine seems to be a pale red wine made from mascalese grade variety.
# Rosato means rose, so the wine is a berry blend of pink color with
# raspberry flavor.

🔄 [(['rosato', np.float64(0.12072411420880161)),
 ('made', np.float64(0.019525884521877177)),
 ('raspberry', np.float64(0.018813132313443552)),
 ('pink', np.float64(0.018761342225685294)),
 ('berry', np.float64(0.016735636621047968)),
 ('colored', np.float64(0.016192552385040596)),
 ('red', np.float64(0.015741363190516028)),
 ('mascalese', np.float64(0.014475671575706405)),
 ('wild', np.float64(0.014042672506815288)),
 ('pale', np.float64(0.01399518869873092))]]

## Run this example of training BERTopic sequentially.

# Step 1 - Extract embeddings
embedding_model = SentenceTransformer("all-MiniLM-L6-v2")

# Step 2 - Reduce dimensionality
umap_model = UMAP(n_neighbors=15, n_components=5,
                  min_dist=0.0, metric='cosine', random_state=42)

# Step 3 - Cluster reduced embeddings
hdbscan_model = HDSCAN(min_cluster_size=15, metric='euclidean',
                       cluster_selection_method='eom', prediction_data=True)

# Step 4 - Tokenize topics
vectorizer_model = CountVectorizer(stop_words="english")

# Step 5 - Create topic representation
ctfidf_model = ClassTfidfTransformer()

# All steps together
topic_model = BERTopic(
    embedding_model=embedding_model,      # Step 1 - Extract embeddings
    umap_model=umap_model,               # Step 2 - Reduce dimensionality
    hdbscan_model=hdbscan_model,         # Step 3 - Cluster reduced embeddings
    vectorizer_model=vectorizer_model,   # Step 4 - Tokenize topics
    ctfidf_model=ctfidf_model,           # Step 5 - Extract topic words

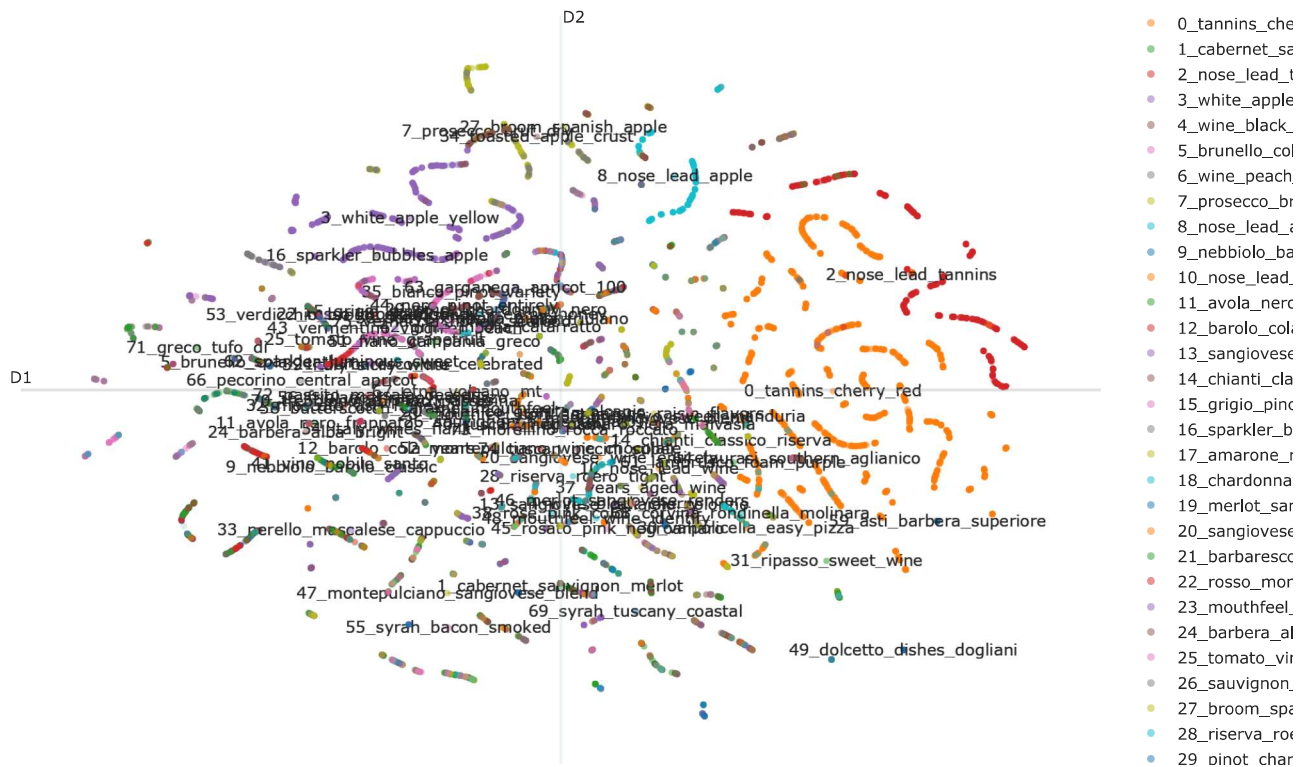
    calculate_probabilities=True,
    verbose=True
)

# Render the following visualizations and answer the questions in a markdown
# cell you insert just above the viz.
# Run the visualization with the original embeddings
# Fit the model
#topics, probs = topic_model.fit_transform(docs)
embeddings = embedding_model.encode(docs, show_progress_bar=False)
reduced_embeddings = UMAP(n_components=2, random_state=42).fit_transform(embeddings)
topic_model.visualize_documents(docs, embeddings=reduced_embeddings)

```



Documents and Topics



1. 2D representation of the topics - What are some parent and subtopics?

I see that there is a large group of wine reviews with white, apple, and yellow -group 3- and tannins, cherry, and red -group 0-.

The group 3 is a parent with many observations of reviews that talk about white fruity wines that are light. The reviews in group 3 make up several subgroups that look like strings, one being focuses on reviews that call the wine pear like and another for reviews that taked about apples. I think it's really cool that the HDB Scan can pick out the overall theme of a wine and then make subgroups for the flavor type and review.

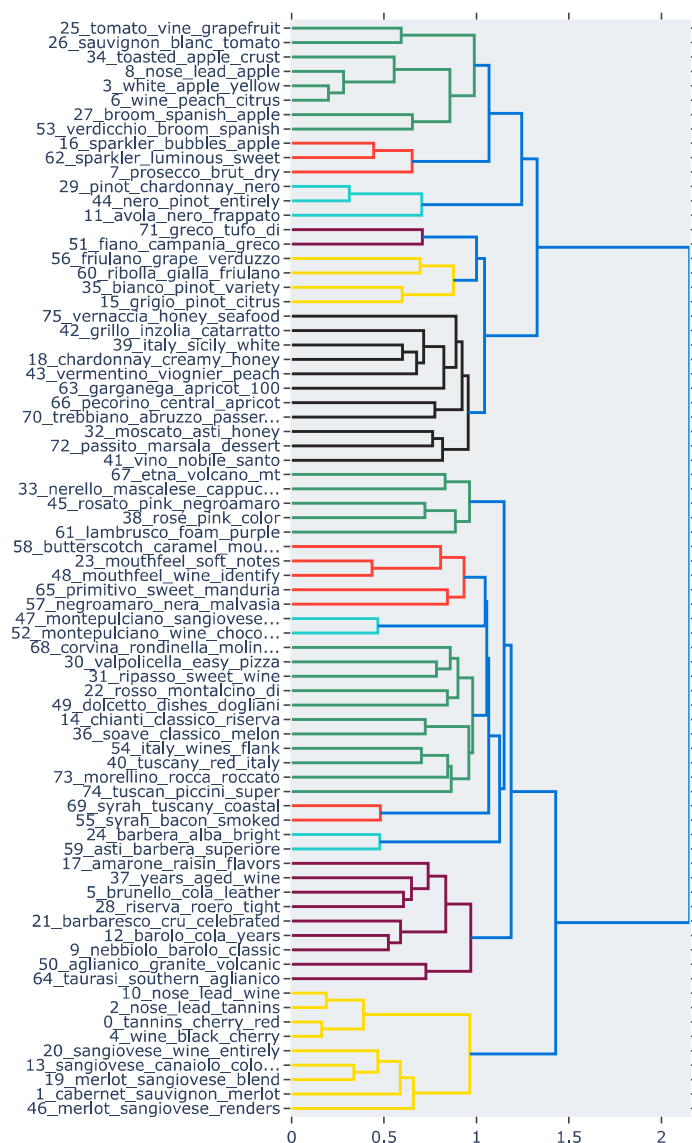
The group 0 is broken up into reviews that focus on texture and flavor i.e. subgroups but there are outlier reviews that talk about drinking the wine before it has breathed or call it something negative like 'funky'.

✓ 2. Hierarchical view - No question here just write code and run it to generatre the viz.

```
fig = topic_model.visualize_hierarchy()
fig.update_layout(width=600, height=1000)
fig.show()
```




Hierarchical Clustering



3. Bar chart - display 10 topics with 10 word labeling (task)

```
# Create bar chart
fig = topic_model.visualize_barchart(top_n_topics=10, n_words=10)
fig.update_layout(width=800, height=800)
fig.show()
```



Topic Word Scores



4. Heatmap - this displays the similarity matrix. By analyzing the similarity matrix, you can identify clusters of topics that are closely related to each other and those that are more distant.

The group of 39 to 42 makes a plaid looking texture in the heatmap where the grouping has similar 'similarity' to groups 3,6,54, and 70. Another trend is groups 4-6 is similar to this 39-42 group but also has similarity to 10, 57,63 which is neat because I can understand how these groups may overlap like a venn diagram if we were to compare only 2-3 similar wines.

```
# Visualize topic similarity heatmap
topic_model.visualize_heatmap()
```



Similarity Matrix

