Tried it using only 10k rows as it was taking very very long

In [ ]:
```
!pip install bertopic
```

In [2]:
```
from bertopic import BERTopic
```

In [6]:
```
import pandas as pd
import re
import string
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer
from nltk.probability import FreqDist
from wordcloud import WordCloud
from collections import defaultdict
from wordcloud import STOPWORDS
import seaborn as sns
from sklearn.feature_extraction.text import TfidfVectorizer,CountVectorizer
from sklearn.decomposition import LatentDirichletAllocation
from sklearn.decomposition import TruncatedSVD
import matplotlib.pyplot as plt
```

In [7]:
```
nltk.download('punkt')
nltk.download('stopwords')
nltk.download('wordnet')
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
[nltk_data] Downloading package wordnet to /root/nltk_data...
```

Out[7]:
```
True
```

In [8]:
```
def cleanse(string):
    # Lower casing
    string = str(string).lower()

    # Getting rid of mentions
    string = re.sub(r"@\S+", " ", string)

    # Removing HTML
    string = re.sub(r"&.*?;|<.*?>", " ", string)

    # URL removal
    string = re.sub(r"https?://\S+|www\.\S+", " ", string)

    # Handling abbreviations
    # string = convert_abbrev_in_text(string)

    # Non-word removals (special chars)
    string = re.sub(r"[^a-z]", " ", string)

    # Stop word removal
    string = " ".join(word for word in nltk.tokenize.word_tokenize(string)
if word not in nltk.corpus.stopwords.words('english'))

    # Lemmatization
    lemma = nltk.stem.WordNetLemmatizer()
    string = " ".join(lemma.lemmatize(word) for word in nltk.tokenize.word_
tokenize(string))

    # Single char removal
    string = re.sub(r"\b\w\b", "", string).strip()

    return string
```

In [4]:
```
df = pd.read_csv('Reviews10k.csv')
```

In [5]:
```
df.head()
```

Out[5]:

	Text
0	We have used Thai Kitchen Peanut sauce many ti...
1	I love this product and I share it with my fam...
2	I love tea. I drink 4-6 cups a day. I've been ...
3	I purchased this coffee because both the Newma...
4	These garlic cloves are amazing! I snack on th...

In [10]:
```
# Example usage
text_to_clean = "This is an example text with @mentions, <html> tags, and h
ttps://example.com links."
cleaned_text = cleanse(text_to_clean)
print(cleaned_text)
```

```
example text tag link
```

In [11]:
```
df['cleaned']=df['Text'].apply(cleanse)
```

In [12]:
```
docs = list(df.loc[:,'cleaned'].values)
```

In [13]:
```
docs[:5]
```

Out[13]:
```
['used thai kitchen peanut sauce many time love chicken shrimp hard find s
tarted looking online best price packing well done glass jar little worrie
d also worried expiration date food jar good well next year worry either g
ood experience',
 'love product share family co worker jv way amazon best online purchase d
efinitely continue buy currently preparing third order many month great up
set stomach sore throat stressful day cool night make hot water milk take
chill',
 'love tea drink cup day tea snob usually make tea loose leaf sometimes pr
efer convenience tea bag really liked mint blend tea although usually drin
k green tea find black tea welcome change every also found extra energy ki
ck tea refreshing give jitter would definitely use',
 'purchased coffee newman sumatran reserve priced near per  cup month need
ed something dark bold replace price came back coffee brewed strong bit em
pty bitter finish caffeine content acceptable package could treated little
better shipping lot dent crushed corner side evidence bending side  cup no
ne cup thus far shown broken seal though overall acceptable extra bold sty
le  cup better option favorite newman sumatran reserve tried black diamond
cup bit pricey right also probably end using coupon buying preferred flavo
r brick mortar store nearby better price ordering coffee people count extr
a bold flavor box pretty stable pricing past month option cheaper substitu
te pricier extra bold flavor wish amazon could sort supply issue keep  cup
price little stable lastly buy similar priced option tl dr okay go extra b
old flavor priced',
 'garlic clove amazing snack right jar think nothing ordering case sent wa
y hawaiian island kaua yes good love garlic even check little morsel white
gold sorry']
```

BERTopic

In [16]:
```
vectorizer_model = CountVectorizer(min_df=2)
topic_model = BERTopic(nr_topics=10)
topics, probs = topic_model.fit_transform(docs)
```

View Topics

In [17]:
```
topic_model.get_topic_info()
```

Out[17]:

	Topic	Count	Name	Representation	Representative_Docs
0	-1	3057	-1_like_product_taste_good	[like, product, taste, good, great, flavor, on...	[flavor tea light even though steeped long tim...
1	0	5194	0_coffee_tea_taste_like	[coffee, tea, taste, like, flavor, good, great...	[drink lot tea dollar store plain black tea ex...
2	1	713	1_cat_food_dog_treat	[cat, food, dog, treat, love, one, like, toy, ...	[felt needed review cat food getting one star ...
3	2	521	2_product_amazon_order_arrived	[product, amazon, order, arrived, box, plant, ...	[great product great price arrived time packag...
4	3	262	3_baby_milk_formula_son	[baby, milk, formula, son, organic, month, lov...	[much research formula best child wanted choos...
5	4	104	4_hair_shampoo_conditioner_scalp	[hair, shampoo, conditioner, scalp, oil, use, ...	[first review shampoo colored damaged hair ver...
6	5	44	5_china_fda_made_dog	[china, fda, made, dog, product, gmo, treat, j...	[ordered three package price good saw asking a...
7	6	43	6_wine_kit_herb_yeast	[wine, kit, herb, yeast, oak, bottle, gallon, ...	[blended week chianti rosso wine kit blended w...
8	7	34	7_trap_fly_gopher_moth	[trap, fly, gopher, moth, ant, dust, sticky, w...	[reading article gardening forum regarding dif...
9	8	28	8_tea_sleep_cough_eye	[tea, sleep, cough, eye, calm, help, also, thr...	[tea great lot stress lately tea definitely he...

trying to fine tune

In [21]:
```
vectorizer_model = CountVectorizer(stop_words="english", ngram_range=(1, 3), max_df=0.95, min_df=2)
topic_model.update_topics(docs, vectorizer_model=vectorizer_model) #straigh
t update no need to retrain
```

In [22]:
```
topic_model.get_topic_info()
```

Out[22]:

	Topic	Count	Name	Representation	Representative_Docs
0	-1	3057	-1_taste_flavor_love_food	[taste, flavor, love, food, dog, bag, sugar, b...	[flavor tea light even though steeped long tim...
1	0	5194	0_coffee_tea_taste_flavor	[coffee, tea, taste, flavor, cup, love, chocol...	[drink lot tea dollar store plain black tea ex...
2	1	713	1_cat_food_dog_treat	[cat, food, dog, treat, love, dog food, toy, e...	[felt needed review cat food getting one star ...
3	2	521	2_arrived_box_plant_received	[arrived, box, plant, received, ordered, shipp...	[great product great price arrived time packag...
4	3	262	3_baby_milk_formula_son	[baby, milk, formula, son, organic, love, old,...	[much research formula best child wanted choos...
5	4	104	4_hair_shampoo_conditioner_scalp	[hair, shampoo, conditioner, scalp, oil, cocon...	[first review shampoo colored damaged hair ver...
6	5	44	5_china_fda_dog_gmo	[china, fda, dog, gmo, treat, jerky, chicken j...	[ordered three package price good saw asking a...
7	6	43	6_wine_kit_herb_yeast	[wine, kit, herb, yeast, gallon, basil, fermen...	[blended week chianti rosso wine kit blended w...
8	7	34	7_trap_fly_ant_dust	[trap, fly, ant, dust, sticky, set, catch, hol...	[reading article gardening forum regarding dif...
9	8	28	8_tea_sleep_cough_eye	[tea, sleep, cough, eye, calm, help, throat, p...	[tea great lot stress lately tea definitely he...

In [34]:
```
# topic_labels = topic_model.generate_topic_labels(nr_words = 10,topic_pref
ix = False, separator = " - ")
# topic_model.set_topic_labels(topic_labels)
# topic_model.get_topic_info()
```

In [43]:
```
topic_model.get_document_info(docs)
```

Out[43]:

	Document	Topic	Name	CustomName	Representation	Representa
0	used thai kitchen peanut sauce many time love ...	-1	-1_taste_flavor_love_food	taste - flavor - love - food - dog - bag - sug...	[taste, flavor, love, food, dog, bag, sugar, b...	[flavor tea though st
1	love product share family co worker jv way ama...	-1	-1_taste_flavor_love_food	taste - flavor - love - food - dog - bag - sug...	[taste, flavor, love, food, dog, bag, sugar, b...	[flavor tea though st
2	love tea drink cup day tea snob usually make t...	0	0_coffee_tea_taste_flavor	coffee - tea - taste - flavor - cup - love - c...	[coffee, tea, taste, flavor, cup, love, chocol...	[drink lo store plai
3	purchased coffee newman sumatran reserve price...	0	0_coffee_tea_taste_flavor	coffee - tea - taste - flavor - cup - love - c...	[coffee, tea, taste, flavor, cup, love, chocol...	[drink lo store plai
4	garlic clove amazing snack right jar think not...	0	0_coffee_tea_taste_flavor	coffee - tea - taste - flavor - cup - love - c...	[coffee, tea, taste, flavor, cup, love, chocol...	[drink lo store plai
...	...	...	...	...	...	
9995	would say favorite cereal one super filling su...	0	0_coffee_tea_taste_flavor	coffee - tea - taste - flavor - cup - love - c...	[coffee, tea, taste, flavor, cup, love, chocol...	[drink lo store plai
9996	like hummus stuff nearly good fresh tired quic...	-1	-1_taste_flavor_love_food	taste - flavor - love - food - dog - bag - sug...	[taste, flavor, love, food, dog, bag, sugar, b...	[flavor tea though st
9997	great treat schnauzer love however make sure m...	5	5_china_fda_dog_gmo	china - fda - dog - gmo - treat - jerky - chic...	[china, fda, dog, gmo, treat, jerky, chicken j...	[or package saw
9998	got gel time box little dented big deal got wi...	2	2_arrived_box_plant_received	arrived - box - plant - received - ordered - s...	[arrived, box, plant, received, ordered, shipp...	[great pr price a
9999	become difficult find senseo pod decided give ...	0	0_coffee_tea_taste_flavor	coffee - tea - taste - flavor - cup - love - c...	[coffee, tea, taste, flavor, cup, love, chocol...	[drink lo store plai

10000 rows × 9 columns

In [23]:
```
topic_model.get_topic_freq()
```

Out[23]:

	Topic	Count
1	0	5194
0	-1	3057
3	1	713
4	2	521
5	3	262
2	4	104
7	5	44
8	6	43
9	7	34
6	8	28

In [30]:
```
topic_model.get_topic(3)
```

Out[30]:
```
[('baby', 0.06292980341837269),
 ('milk', 0.05969024035000255),
 ('formula', 0.047611211572692276),
 ('son', 0.03143695613619083),
 ('organic', 0.029902093606828176),
 ('love', 0.027865180617748713),
 ('old', 0.025904653729729953),
 ('kid', 0.025298340190530646),
 ('food', 0.025099937439142603),
 ('daughter', 0.024854483022166134)]
```

In [29]:
```
for i in range(-1, 9):
    topic = topic_model.get_topic(i)
    words = [word for word, _ in topic]  # Extracting words, ignoring weigh
ts
    print(f"Topic {i}: {', '.join(words)}")
```

```
Topic -1: taste, flavor, love, food, dog, bag, sugar, buy, sweet, box
Topic 0: coffee, tea, taste, flavor, cup, love, chocolate, best, drink, su
gar
Topic 1: cat, food, dog, treat, love, dog food, toy, eat, cat food, pill
Topic 2: arrived, box, plant, received, ordered, shipping, item, flower, b
uy, service
Topic 3: baby, milk, formula, son, organic, love, old, kid, food, daughter
Topic 4: hair, shampoo, conditioner, scalp, oil, coconut, skin, coconut oi
l, clear, smell
Topic 5: china, fda, dog, gmo, treat, jerky, chicken jerky, chicken, compa
ny, kellogg
Topic 6: wine, kit, herb, yeast, gallon, basil, fermentation, garden, batc
h, fresh
Topic 7: trap, fly, ant, dust, sticky, set, catch, hole, wire, area
Topic 8: tea, sleep, cough, eye, calm, help, throat, pain, drinking, sleep
aid
```

In [38]:
```
topic_model.visualize_barchart()
```

In [27]:
```
topic_model.visualize_topics()
```

In [40]:
```
# topic_model.visualize_hierarchy()
```