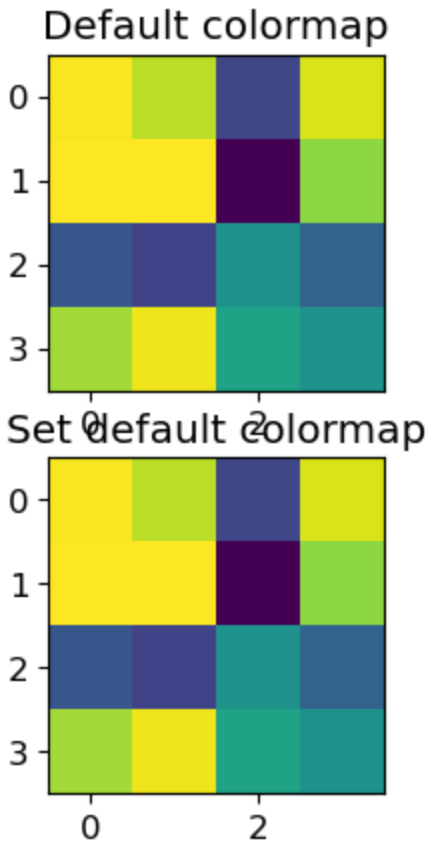
Lecture 02: Clustering class demo

Contents

To run the code below, you need to install pytorch and torchvision in the course conda environment.

```
conda install pytorch torchvision -c pytorch
```

```
import numpy as np
import pandas as pd
import sys
import os
import torch, torchvision
from torchvision import datasets, models, transforms, utils
from PIL import Image
from torchvision import transforms
from torchvision.models import vgg16
import matplotlib.pyplot as plt
DATA_DIR = os.path.join(os.path.abspath(".."), "data/")
sys.path.append(os.path.join(os.path.abspath(".."), "code"))
from plotting_functions import *
```



Hierarchical clustering on recipes dataset

Let's create a dendrogram on a more realistic dataset.

We'll use Kaggle's Food.com recipes corpus. This corpus contains 180K+ recipes and 700K+ recipe reviews. In this lab, we'll only focus on recipes and **not** on reviews. The recipes are present in `RAW_recipes.csv`. Our goal is to find categories or groupings of recipes from this corpus based on their names.

```
orig_recipes_df = pd.read_csv(DATA_DIR + "RAW_recipes.csv")
orig_recipes_df.shape
```

```
(231637, 12)
```

```
def get_recipes_sample(orig_recipes_df, n_tags=300, min_len=5):
    orig_recipes_df = orig_recipes_df.dropna()  # Remove rows with NaNs.
    orig_recipes_df = orig_recipes_df.drop_duplicates(
        "name"
    )  # Remove rows with duplicate names.
    orig_recipes_df = orig_recipes_df[orig_recipes_df["name"].apply(len) >= mi
    first_n = orig_recipes_df["tags"].value_counts()[0:n_tags].index.tolist()
    recipes_df = orig_recipes_df[orig_recipes_df["tags"].isin(first_n)]
    return recipes_df
```

```
recipes_df = get_recipes_sample(orig_recipes_df)
recipes_df.shape
```

```
(9100, 12)
```

```
recipes_df["name"]
```

```
42            i yam what i yam  muffins
101             to your health  muffins
129       250 00 chocolate chip cookies
138                     lplermagronen
163             california roll   salad
                    ...
231430      zucchini wheat germ cookies
231514         zucchini blueberry bread
231547           zucchini salsa burgers
231596                    zuppa toscana
231629                     zydeco salad
Name: name, Length: 9100, dtype: object
```

```
from sentence_transformers import SentenceTransformer
embedder = SentenceTransformer('all-MiniLM-L6-v2')
```

```
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
File ~/miniforge3/envs/jbook/lib/python3.12/site-packages/transformers/utils/im
   1862 try:
-> 1863     return importlib.import_module("." + module_name, self.__name__)
   1864 except Exception as e:

File ~/miniforge3/envs/jbook/lib/python3.12/importlib/__init__.py:90, in import
     89         level += 1
---> 90 return _bootstrap._gcd_import(name[level:], package, level)

File <frozen importlib._bootstrap>:1387, in _gcd_import(name, package, level)

File <frozen importlib._bootstrap>:1360, in _find_and_load(name, import_)

File <frozen importlib._bootstrap>:1331, in _find_and_load_unlocked(name, impo

File <frozen importlib._bootstrap>:935, in _load_unlocked(spec)

File <frozen importlib._bootstrap_external>:995, in exec_module(self, module)

File <frozen importlib._bootstrap>:488, in _call_with_frames_removed(f, *args,

File ~/miniforge3/envs/jbook/lib/python3.12/site-packages/transformers/integra
     23 import numpy as np
---> 24 from tokenizers import Tokenizer, decoders, normalizers, pre_tokenizers
     25 from tokenizers.models import BPE, Unigram

File ~/miniforge3/envs/jbook/lib/python3.12/site-packages/tokenizers/__init__.p
     78 from .tokenizers import (
     79     AddedToken,
     80     Encoding,
   (...)
     92     __version__,
     93 )
---> 94 from .implementations import (
     95     BertWordPieceTokenizer,
     96     ByteLevelBPETokenizer,
     97     CharBPETokenizer,
     98     SentencePieceBPETokenizer,
     99     SentencePieceUnigramTokenizer,
    100 )

File ~/miniforge3/envs/jbook/lib/python3.12/site-packages/tokenizers/implementa
----> 1 from .base_tokenizer import BaseTokenizer
      2 from .bert_wordpiece import BertWordPieceTokenizer

File ~/miniforge3/envs/jbook/lib/python3.12/site-packages/tokenizers/implementa
      3 from tokenizers import AddedToken, EncodeInput, Encoding, InputSequence
----> 4 from tokenizers.decoders import Decoder
      5 from tokenizers.models import Model

File ~/miniforge3/envs/jbook/lib/python3.12/site-packages/tokenizers/decoders/_
     14 Sequence = decoders.Sequence
---> 15 DecodeStream = decoders.DecodeStream
```

```
AttributeError: module 'decoders' has no attribute 'DecodeStream'

The above exception was the direct cause of the following exception:

RuntimeError                              Traceback (most recent call last)
File ~/miniforge3/envs/jbook/lib/python3.12/site-packages/transformers/utils/ir
   1862 try:
-> 1863     return importlib.import_module("." + module_name, self.__name__)
   1864 except Exception as e:

File ~/miniforge3/envs/jbook/lib/python3.12/importlib/__init__.py:90, in impor
     89         level += 1
---> 90 return _bootstrap._gcd_import(name[level:], package, level)

File <frozen importlib._bootstrap>:1387, in _gcd_import(name, package, level)

File <frozen importlib._bootstrap>:1360, in _find_and_load(name, import_)

File <frozen importlib._bootstrap>:1331, in _find_and_load_unlocked(name, impo

File <frozen importlib._bootstrap>:935, in _load_unlocked(spec)

File <frozen importlib._bootstrap_external>:995, in exec_module(self, module)

File <frozen importlib._bootstrap>:488, in _call_with_frames_removed(f, *args,

File ~/miniforge3/envs/jbook/lib/python3.12/site-packages/transformers/models/a
     22 from typing import List, Union
---> 24 from ...configuration_utils import PretrainedConfig
     25 from ...dynamic_module_utils import get_class_from_dynamic_module, res

File ~/miniforge3/envs/jbook/lib/python3.12/site-packages/transformers/configu
     28 from .dynamic_module_utils import custom_object_save
---> 29 from .modeling_gguf_pytorch_utils import load_gguf_checkpoint
     30 from .utils import (
     31     CONFIG_NAME,
     32     PushToHubMixin,
   (...)
     41     logging,
     42 )

File ~/miniforge3/envs/jbook/lib/python3.12/site-packages/transformers/modelin
     21 from tqdm.auto import tqdm
---> 23 from .integrations import (
     24     GGUF_CONFIG_MAPPING,
     25     GGUF_TOKENIZER_MAPPING,
     26     _gguf_parse_value,
     27 )
     28 from .utils import is_torch_available

File <frozen importlib._bootstrap>:1412, in _handle_fromlist(module, fromlist,

File ~/miniforge3/envs/jbook/lib/python3.12/site-packages/transformers/utils/ir
   1850 elif name in self._class_to_module.keys():
-> 1851     module = self._get_module(self._class_to_module[name])
```

```
   1852     value = getattr(module, name)

File ~/miniforge3/envs/jbook/lib/python3.12/site-packages/transformers/utils/im
   1864 except Exception as e:
-> 1865     raise RuntimeError(
   1866         f"Failed to import {self.__name__}.{module_name} because of the
   1867         f" traceback):\n{e}"
   1868     ) from e

RuntimeError: Failed to import transformers.integrations.ggml because of the fo
module 'decoders' has no attribute 'DecodeStream'

The above exception was the direct cause of the following exception:

RuntimeError                              Traceback (most recent call last)
Cell In[6], line 1
----> 1 from sentence_transformers import SentenceTransformer
      2 embedder = SentenceTransformer('all-MiniLM-L6-v2')

File ~/miniforge3/envs/jbook/lib/python3.12/site-packages/sentence_transformers
      7 import os
      9 from sentence_transformers.backend import (
     10     export_dynamic_quantized_onnx_model,
     11     export_optimized_onnx_model,
     12     export_static_quantized_openvino_model,
     13 )
---> 14 from sentence_transformers.cross_encoder.CrossEncoder import CrossEncod
     15 from sentence_transformers.datasets import ParallelSentencesDataset, Se
     16 from sentence_transformers.LoggingHandler import LoggingHandler

File ~/miniforge3/envs/jbook/lib/python3.12/site-packages/sentence_transformers
      1 from __future__ import annotations
----> 3 from .CrossEncoder import CrossEncoder
      5 __all__ = ["CrossEncoder"]

File ~/miniforge3/envs/jbook/lib/python3.12/site-packages/sentence_transformers
     12 from torch.utils.data import DataLoader
     13 from tqdm.autonotebook import tqdm, trange
---> 14 from transformers import AutoConfig, AutoModelForSequenceClassificatio
     15 from transformers.tokenization_utils_base import BatchEncoding
     16 from transformers.utils import PushToHubMixin

File <frozen importlib._bootstrap>:1412, in _handle_fromlist(module, fromlist,

File ~/miniforge3/envs/jbook/lib/python3.12/site-packages/transformers/utils/im
   1850 elif name in self._class_to_module.keys():
   1851     module = self._get_module(self._class_to_module[name])
-> 1852     value = getattr(module, name)
   1853 elif name in self._modules:
   1854     value = self._get_module(name)

File ~/miniforge3/envs/jbook/lib/python3.12/site-packages/transformers/utils/im
   1849     value = Placeholder
   1850 elif name in self._class_to_module.keys():
-> 1851     module = self._get_module(self._class_to_module[name])
   1852     value = getattr(module, name)
```

```
   1853 elif name in self._modules:

File ~/miniforge3/envs/jbook/lib/python3.12/site-packages/transformers/utils/i
   1863     return importlib.import_module("." + module_name, self.__name__)
   1864 except Exception as e:
-> 1865     raise RuntimeError(
   1866         f"Failed to import {self.__name__}.{module_name} because of th
   1867         f" traceback):\n{e}"
   1868     ) from e

RuntimeError: Failed to import transformers.models.auto.configuration_auto bec
Failed to import transformers.integrations.ggml because of the following error
module 'decoders' has no attribute 'DecodeStream'
```

```
recipe_names = recipes_df["name"].tolist()
embeddings = embedder.encode(recipe_names)
recipe_names_embeddings = pd.DataFrame(
    embeddings,
    index=recipes_df.index,
)
recipe_names_embeddings
```

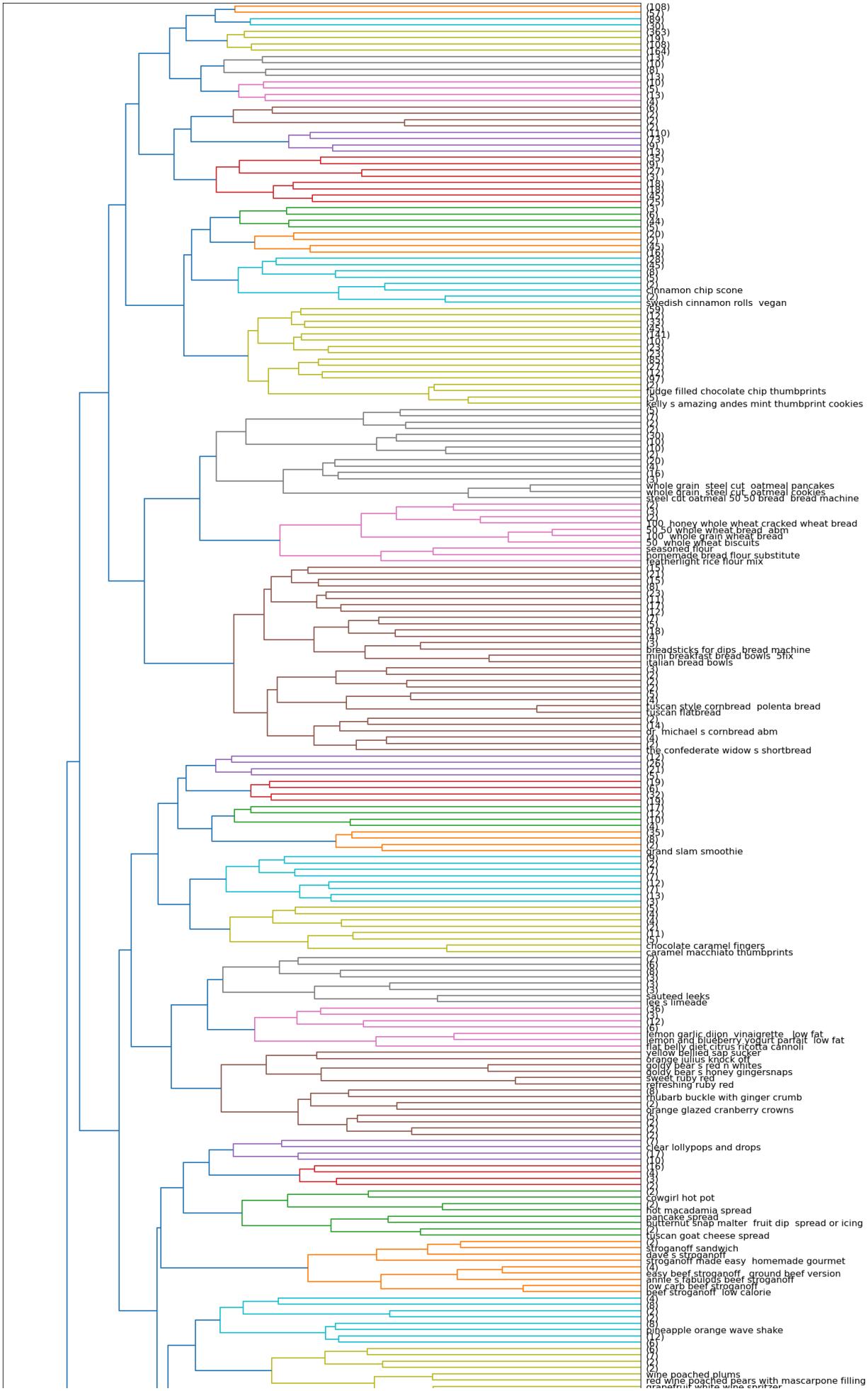
	0	1	2	3	4	5	
42	0.019592	-0.088336	0.072677	-0.034575	-0.048741	-0.049801	0
101	-0.000567	-0.011825	0.073199	0.058175	0.031688	-0.015428	0
129	-0.022604	0.065034	-0.033065	0.014450	-0.105039	-0.050559	0
138	-0.066915	0.025988	-0.087689	-0.006847	-0.012861	0.049035	0
163	-0.007068	-0.007308	-0.026629	-0.004153	-0.052810	0.011126	0.
...	...	...	...	...	...	...	
231430	-0.036366	0.087173	-0.039641	0.002705	0.097142	-0.075385	0.
231514	-0.052718	0.008980	-0.046014	0.030194	0.005201	0.009964	-0.
231547	-0.080801	0.004295	-0.044325	0.038307	-0.030125	-0.063566	0.
231596	-0.060801	0.111672	-0.108755	0.052323	-0.099851	-0.027532	0.
231629	-0.085059	0.009065	-0.088442	-0.008907	0.002234	-0.055801	0.

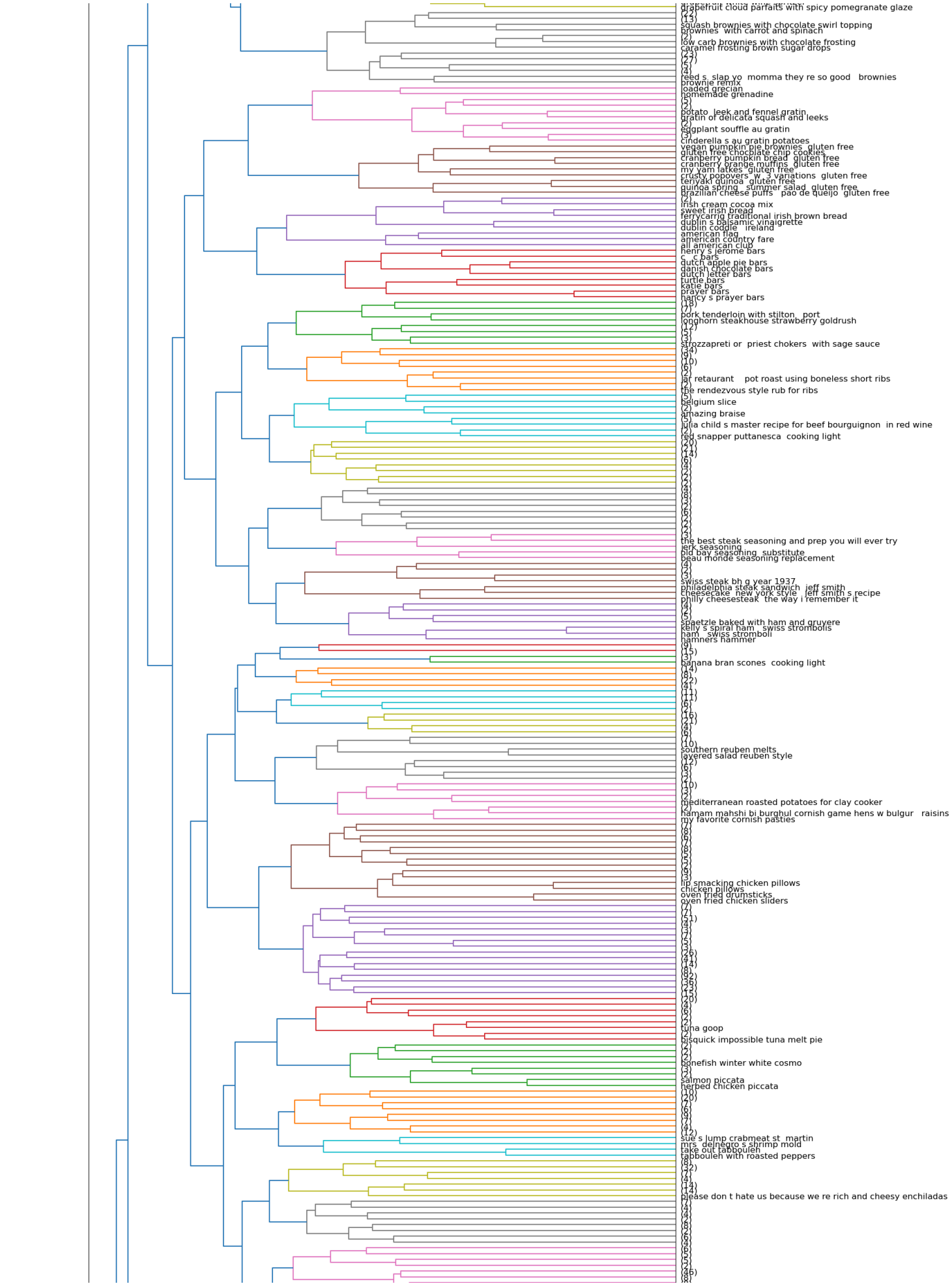
9100 rows × 384 columns

```
def plot_dendrogram(Z, labels, w=15, h=20, p=10):
    fig, ax = plt.subplots(figsize=(w, h))
    dendrogram(
        Z,
        p=p,
        truncate_mode="level",
        orientation="left",
        leaf_font_size=12,
        labels=labels,
        ax=ax,
    )
    ax = plt.gca()
    ax.set_ylabel("Cluster distances", fontsize=w)
```

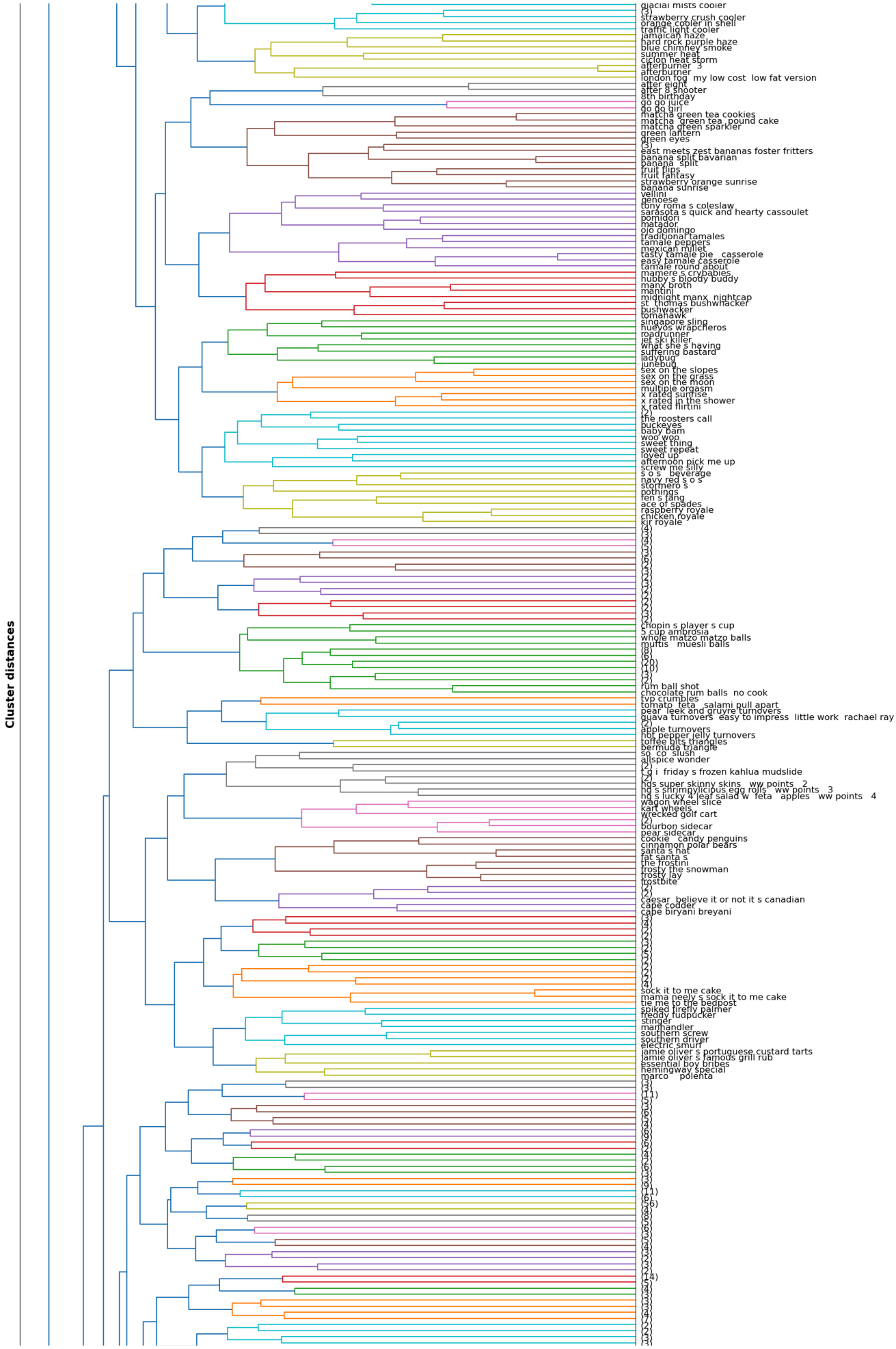
```
from scipy.cluster.hierarchy import dendrogram, linkage

recipe_names = recipes_df["name"].tolist()
Z = linkage(embeddings, method="complete", metric="cosine")
```

```
plot_dendrogram(Z, labels=recipe_names, w=15, h=230, p=10)
```


(108)
(57)
(89)
(38)
(363)
(19)
(108)
(164)
(13)
(10)
(8)
(13)
(10)
(5)
(13)
(4)
(6)
(2)
(2)
(2)
(110)
(13)
(9)
(13)
(35)
(9)
(27)
(3)
(18)
(18)
(18)
(45)
(25)
(5)
(6)
(4)
(44)
(5)
(20)
(2)
(45)
(18)
(28)
(45)
(8)
(6)
(5)
(2)
cinnamon chip scone
(2)
swedish cinnamon rolls vegan
(59)
(12)
(33)
(45)
(21)
(141)
(10)
(23)
(53)
(25)
(85)
(27)
(12)
(97)
(2)
fudge filled chocolate chip thumbprints
(5)
kelly s amazing andes mint thumbprint cookies
(5)
(2)
(2)
(2)
(2)
(30)
(10)
(10)
(2)
(20)
(4)
(16)
(3)
whole grain steel cut oatmeal pancakes
whole grain steel cut oatmeal cookies
steel cut oatmeal 50 50 bread bread machine
(2)
(5)
(3)
(2)
100 honey whole wheat cracked wheat bread
50 50 whole wheat bread abm
100 whole grain wheat bread
50 whole wheat biscuits
seasoned flour
homemade bread flour substitute
featherlight rice flour mix
(15)
(21)
(15)
(8)
(23)
(111)
(17)
(21)
(12)
(7)
(5)
(18)
(4)
(3)
breadsticks for dips bread machine
mini breakfast bread bowls 5fix
italian bread bowls
(3)
(2)
(2)
(2)
(5)
(4)
tuscan style cornbread polenta bread
tuscan flatbread
(2)
(14)
dr michael s cornbread abm
(4)
(6)
the confederate widow s shortbread
(12)
(26)
(21)
(5)
(19)
(6)
(32)
(19)
(17)
(12)
(10)
(4)
(35)
(8)
(2)
grand slam smoothie
(6)
(9)
(2)
(7)
(7)
(12)
(7)
(13)
(3)
(5)
(3)
(4)
(4)
(3)
(11)
(5)
chocolate caramel fingers
caramel macchiato thumbprints
(2)
(6)
(8)
(3)
(3)
(3)
sauteed leeks
lee s limeade
(36)
(3)
(12)
(6)
lemon garlic dijon vinaigrette low fat
lemon and blueberry yogurt parfait low fat
flat belly diet citrus ricotta cannoli
yellow bellied sap sucker
orange julius knock off
goldy bear s red n whites
goldy bear s honey gingersnaps
sweet ruby red
refreshing ruby red
(8)
rhubarb buckle with ginger crumb
(2)
orange glazed cranberry crowns
(5)
(2)
(2)
(2)
(7)
clear lollypops and drops
(17)
(10)
(18)
(16)
(4)
(3)
(3)
(2)
(2)
cowgirl hot pot
(2)
hot macadamia spread
pancake spread
butternut snap malter fruit dip spread or icing
(2)
tuscan goat cheese spread
(2)
stroganoff sandwich
dave s stroganoff
stroganoff made easy homemade gourmet
(4)
easy beef stroganoff ground beef version
anne s fabulous beef stroganoff
low carb beef stroganoff
beef stroganoff low calorie
(4)
(8)
(2)
(2)
(8)
pineapple orange wave shake
(12)
(6)
(9)
(7)
(2)
wine poached plums
red wine poached pears with mascarpone filling
grapefruit white wine spritzer

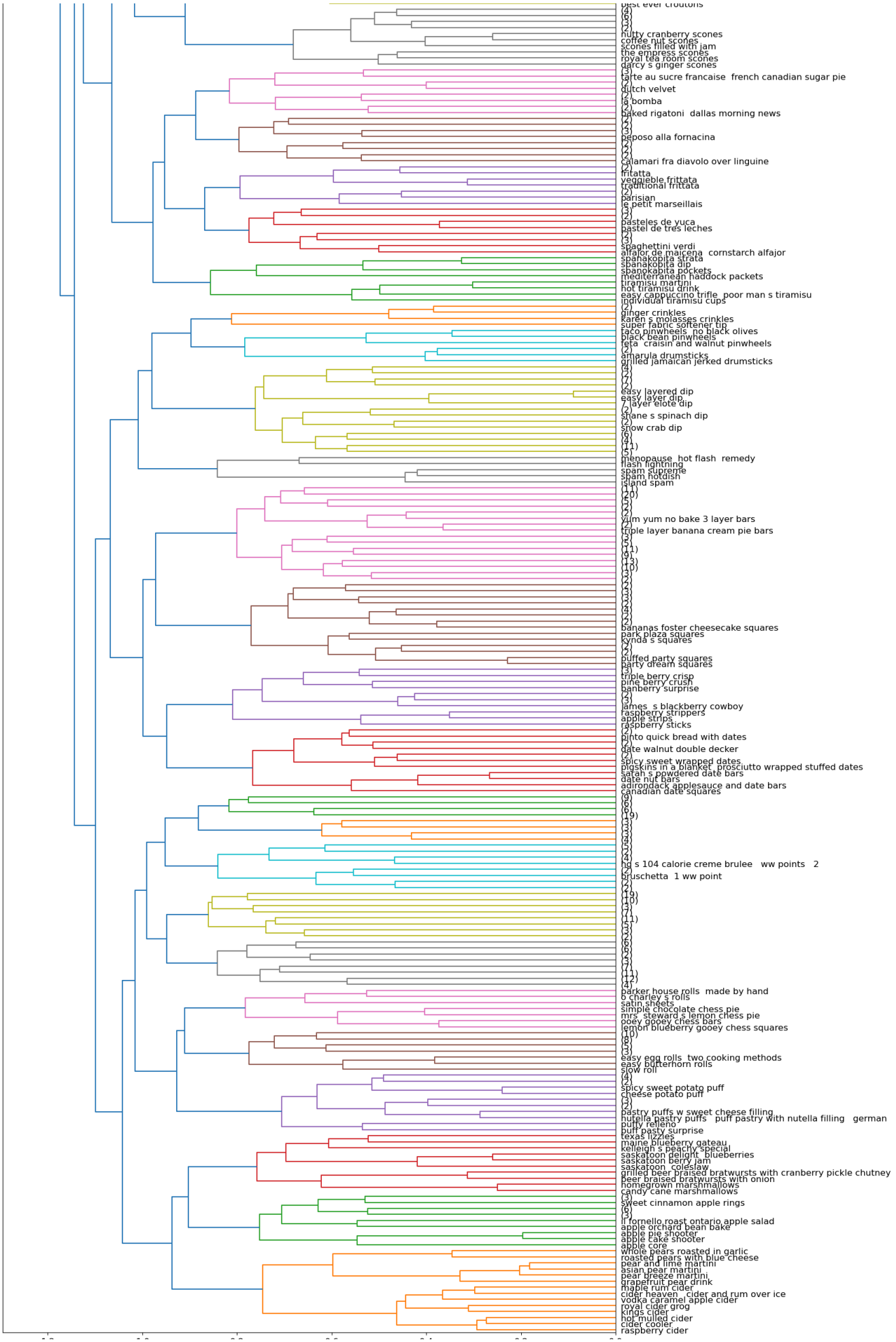

grapefruit cloud parfaits with spicy pomegranate glaze
(22)
(13)
squash brownies with chocolate swirl topping
brownies with carrot and spinach
(2)
low carb brownies with chocolate frosting
caramel frosting brown sugar drops
(23)
(27)
(7)
(2)
(4)
reed s slap yo momma they re so good brownies
brownie remix
loaded grecian
homemade grenadine
(5)
(2)
potato leek and fennel gratin
gratin of delicata squash and leeks
(2)
eggplant souffle au gratin
(3)
cinderella s au gratin potatoes
vegan pumpkin pie brownies gluten free
gluten free chocolate chip cookies
cranberry pumpkin bread gluten free
cranberry orange muffins gluten free
my yam latkes gluten free
crusty popovers w 3 variations gluten free
teriyaki quinoa gluten free
quinoa spring summer salad gluten free
brazilian cheese puffs pao de queijo gluten free
(2)
irish cream cocoa mix
sweet irish bread
ferrycarrig traditional irish brown bread
dublin s balsamic vinaigrette
dublin coddle ireland
american flag
american country fare
all american club
henry s jerome bars
c c bars
dutch apple pie bars
danish chocolate bars
dutch letter bars
turtle bars
katie bars
prayer bars
nancy s prayer bars
(18)
(7)
pork tenderloin with stilton port
longhorn steakhouse strawberry goldrush
(12)
(5)
(3)
strozzapreti or priest chokers with sage sauce
(34)
(9)
(10)
(6)
(2)
jar retaurant pot roast using boneless short ribs
(4)
the rendezvous style rub for ribs
(5)
belgium slice
(2)
amazing braise
(5)
julia child s master recipe for beef bourguignon in red wine
(2)
red snapper puttanesca cooking light
(20)
(21)
(14)
(6)
(4)
(2)
(2)
(2)
(4)
(8)
(3)
(2)
(6)
(2)
(2)
(3)
the best steak seasoning and prep you will ever try
jerk seasoning
bid bay seasoning substitute
beau monde seasoning replacement
(4)
(2)
(3)
swiss steak bh g year 1937
philadelphia steak sandwich jeff smith
cheesecake new york style jeff smith s recipe
philly cheesesteak the way i remember it
(4)
(2)
(5)
(3)
spaetzle baked with ham and gruyere
kelly s spiral ham swiss stromboli s
ham swiss stromboli
hamners hammer
(9)
(15)
(3)
banana bran scones cooking light
(14)
(8)
(22)
(4)
(11)
(11)
(6)
(2)
(16)
(21)
(4)
(6)
(7)
(10)
southern reuben melts
layered salad reuben style
(12)
(6)
(3)
(2)
(10)
(3)
(2)
mediterranean roasted potatoes for clay cooker
(2)
hamam mahshi bi burghul cornish game hens w bulgur raisins
my favorite cornish pasties
(7)
(8)
(6)
(7)
(8)
(6)
(5)
(2)
(6)
(3)
lip smacking chicken pillows
chicken pillows
oven fried drumsticks
oven fried chicken sliders
(7)
(7)
(51)
(4)
(3)
(7)
(5)
(3)
(26)
(41)
(14)
(8)
(92)
(36)
(18)
(23)
(15)
(20)
(4)
(6)
(2)
(2)
tuna goop
(2)
bisquick impossible tuna melt pie
(2)
(2)
(2)
bonefish winter white cosmo
(3)
(2)
salmon piccata
herbed chicken piccata
(10)
(20)
(7)
(6)
(9)
(7)
(4)
(12)
sue s lump crabmeat st martin
mrs delnegro s shrimp mold
take out tabbouleh
tabbouleh with roasted peppers
(8)
(32)
(2)
(4)
(14)
(14)
please don t hate us because we re rich and cheesy enchiladas
(7)
(4)
(4)
(2)
(8)
(2)
(6)
(4)
(6)
(5)
(5)
(2)
(46)
(8)


(10)
(13)
(13)
(26)
(15)
(10)
(3)
(2)
(2)
jalapeno boats
(2)
dukkah duqqa pronounced doo kah
(3)
mini muffuletta
schwenkbraten
schuylkill river punch
savory parsnip mash
monstrous mash
star s sangria
shortcut sangria
lady cello
hpnotiq rita
cranberry rita
bjermagronen
(4)
(4)
(7)
zippity zingity coleslaw
dragonsbreath slabs
mirin
nachynka
hop frog
froggies aka new york delight
(2)
(2)
(2)
(2)
kimikaze
kamikaze on the rag
(2)
(2)
saketini
lychee saketini blush
nitro
nuan gao
(2)
ciabatta
vanita s insalata verde
insalata caprese
tuscan minestrone soup
tuscan minestrone
mamma mia minestrone
(2)
revani
(3)
(3)
(2)
(2)
bagno di poppea
(2)
bossanova
ouzo long drink
adios amigos
pap or sadza or ugali or
(3)
(2)
flor de cana tribal cana in celebration of survivor nicaragua
voodoo
illusion
the undercurrent
mob witnesses
(2)
bloated bag of monkey spunk
abehjerne monkey brains
(2)
procession of the pink elephants
elephant shake
elephant gun cocktail
(2)
(2)
(6)
scarlett kiss
the seductive swan
midsummers night dream
passion potion
hana passion
golden dream
american dream
frozen mixed dream
dreamsicle daiquiri
dreamsicle cocktail
vanilla dreamsicle
dreamsicle creamsicle on the rocks
bobby flay s bourbon cherry slushie
bobby flay throwdown blondies
texas roadhouse cowboy quencher
my favorite rouladen
chicken roulade susie
chocoflan
bailey s chocotini
(2)
carrol s family goulash
shane s goulash
ron s gourmet dawgs
susan d s biscotti
jana s church enchiladas
bridesmaid
melanie s moussaka
gail s goop
connie s stroganoff
cindy s stromboli
the margareta
melissa s fried taunga
(2)
pudina
time capsule
cold shoulder
bad attitude
sunstroke
neopolitan
blizzard
jalousie
exotic
tequilibrium
irrational exuberance
sweating colombian
columbian gold
latin connection
el ritual
el nino
the saint charles
montreal saint germain
holyfield s ear
grouper verzcruz
absolut trouble
poulet bonne femme
filet sin
profiteroles
pumpkin hermits
hermits
mincemeat hermits
silkey s mai tai
midori mai tai
p f chang s poolside
chanel no 6
peugeot
220 btu
(2)
northwood 2
washington apple
cafe seattle
china white
black and whites
tracy s white trash
(2)
(2)
the realtor s black forest pie quick and easy
full house
(2)
malibu zinger
the 69
57 chevy
(2)
(2)
(2)
(3)
sunset passion colada
golden star
blue star
milky way galaxy
southern hurricane pitcher
new orleans hurricane
calgary stampede beergarita
(2)
vista blue
(2)
(2)
ice blue hpnotiq
hpnotiq blue ibiza
(2)
(2)
tropical life saver
rainforest trail mix
vodka gilligan s island
gilligan s island
virgin islands orange colada
(2)
caribbean connection
caribbean cassis
isle of pines
isle of coconut
hawaiian sugar plantation
rose water face mask
blushing lady
egyptian roses
pierced fuzzy navel
fuzzy navel
an elbow in the boobs
(2)
floating on a cloud
dunking platters
(4)
a day at the beach
sand in your shorts
sand drops
(2)
cariad rhudd crimson love
(2)
big red hooter
(5)
(2)
purple gecko
purple crayon
lurkers
leg spreader
(2)
kerry simon iron chef slider
white castle like sliders
(4)
thanksgiving s a breeze seabreezes
(2)
(4)
cuervo caribbean mist
caribbean mist cocktail


Cluster distances
glacial mists cooler
(3)
strawberry crush cooler
orange cobler in shell
traffic light cooler
jamaican haze
hard rock purple haze
blue chimney smoke
summer heat
ciclon heat storm
afterburner 3
afterburner
london fog my low cost low fat version
after eight
after 8 shooter
8th birthday
go go juice
go go girl
matcha green tea cookies
matcha green tea pound cake
matcha green sparkler
green lantern
green eyes
(3)
east meets zest bananas foster fritters
banana split bavarian
banana split
fruit flips
fruit fantasy
strawberry orange sunrise
banana sunrise
vellini
genoese
tony roma s coleslaw
sarasota s quick and hearty cassoulet
pomidori
matador
ojo domingo
traditional tamales
tamale peppers
mexican millet
tasty tamale pie casserole
easy tamale casserole
tamale round about
mamere s crybabies
hubby s bloody buddy
manx broth
martini
midnight manx nightcap
st thomas bushwhacker
bushwacker
tomahawk
singapore sling
huevos wrapcheros
roadrunner
jet ski killer
what she s having
suffering bastard
ladybug
lunabug
sex on the slopes
sex on the grass
sex on the moon
multiple orgasm
x rated sunrise
x rated in the shower
x rated flirtini
(2)
(4)
the roosters call
buckeyes
baby bam
woo woo
sweet thing
sweet repeat
loved up
afternoon pick me up
screw me silly
s o s beverage
navy red s o s
stormero s
nothings
ten s ring
ace of spades
raspberry royale
chicken royale
kir royale
(4)
(3)
(4)
(5)
(3)
(3)
(6)
(2)
(4)
(3)
(2)
(3)
(2)
(4)
(2)
(2)
(2)
(3)
(2)
chopin s player s cup
5 cup ambrosia
whole matzo matzo balls
muffis muesli balls
(8)
(6)
(20)
(10)
(3)
(2)
rum ball shot
chocolate rum balls no cook
tvp crumbles
tomato feta salami pull apart
pear leek and gruyre turnovers
guava turnovers easy to impress little work rachael ray
(2)
apple turnovers
hot pepper jelly turnovers
toffee bits triangles
bermuda triangle
so co slush
allspice wonder
(2)
(2)
i d i friday s frozen kahlua mudslide
(2)
hgs super skinny skins ww points 2
hg s shrimpylicious egg rolls ww points 3
hg s lucky 4 leaf salad w feta apples ww points 4
wagon wheel slice
kart wheels
wrecked golf cart
(2)
bourbon sidecar
pear sidecar
cookie candy penguins
cinnamon polar bears
santa s hat
fat santa s
the frostini
frosty the snowman
frosty jay
frostbite
(2)
(2)
caesar believe it or not it s canadian
cape codder
cape biryani breyani
(3)
(2)
(2)
(2)
(2)
(3)
(2)
(4)
(5)
(2)
(2)
(2)
(2)
(2)
(4)
sock it to me cake
mama neely s sock it to me cake
tie me to the bedpost
spiked firefly palmer
freddy fudpucker
stinger
manhandler
southern screw
southern driver
electric smurf
jamie oliver s portuguese custard tarts
jamie oliver s famous grill rub
essential boy bribes
hemingway special
marco polenta
(3)
(3)
(11)
(5)
(3)
(4)
(6)
(5)
(4)
(6)
(9)
(6)
(2)
(4)
(3)
(2)
(6)
(3)
(3)
(3)
(9)
(11)
(6)
(56)
(4)
(8)
(5)
(6)
(3)
(5)
(4)
(3)
(2)
(3)
(2)
(14)
(5)
(4)
(3)
(3)
(3)
(3)
(4)
(7)
(2)
(2)
(3)
(3)


(6)
(9)
(11)
(5)
(3)
(4)
(2)
(2)
licorice stick shot
batter
90 minute batter rolls
the marv albert
soup for the hearts rsc
cheesy heartattacks 5fix
treasure chest bars
captain morgan s treasure
(2)
the angel in your pantry pasta dinner
fallen angel
angel face
old fashioned
bourbon old fashioned
2000 flushes
cheese and sage damper
apricot damper
fantastic glass cleaner
(3)
(3)
(57)
(19)
(53)
(54)
(2)
(16)
(9)
(15)
(9)
(3)
douglas fir sparkletini
(2)
(3)
(2)
(2)
(2)
amaretti
(7)
(2)
(2)
(2)
memphis style rub
memphis rub by steven raichlen
(3)
storming pink bourbon
smokin drunkin flame broiled turkey turkey on a beer can
beer butt chicken beer can
(2)
(2)
(4)
(4)
(3)
(2)
victory s homemade root beer schnapps
gin buck
tiki john s rum runner
holiday isle rum runner
rumarita
campechana rum
(4)
(2)
(5)
(5)
(2)
(2)
fresh vegetable juice v8 double
homemade pedialyte
tootsie roll float cocktail
koala float with alcohol
(2)
old fashioned root beer float cake
tipsy coke float
diet soda floats
(42)
(77)
(19)
(20)
(2)
salt free poultry seasoning
sauerkraut rye bread sugar free
hachrippe und sauerkraut spareribs and sauerkraut
(55)
(30)
(12)
(9)
(8)
(3)
(9)
(12)
(11)
(7)
(23)
(8)
(21)
(51)
(5)
(5)
(3)
weight watchers chile cheese dip
weight watchers white bread
weight watchers garlic herb bread
weight watchers apple cole slaw 1 point
mom s apple crisp weight watchers
weight watcher s king ranch
(7)
(13)
(16)
(7)
(73)
(8)
(10)
(16)
(3)
(3)
(4)
(2)
(6)
(5)
(6)
(5)
(19)
(15)
(13)
(17)
(6)
(13)
(10)
(4)
(3)
(126)
(26)
(55)
(78)
(9)
(6)
(17)
(10)
(59)
(24)
(19)
(40)
(39)
(27)
(64)
(10)
(23)
(23)
(5)
(4)
(4)
(10)
(19)
(3)
(2)
(7)
(26)
(9)
(2)
(4)
(12)
(4)
(3)
(6)
(2)
(23)
(3)
(8)
(4)
(5)
franks in foil
(2)
panko encrusted tilapia fillets
(7)
(11)
lower fat new england style clam chowder
clam chowda mug o soup
stovetop clam bake
(3)
ham spinach and fontina or gruyere strata
(2)
foil pack tilapia with lemon rosemary tomato
cucumbers in sour cream
cucumber mint gimlet
light mexican custard
custard millefeuille
(10)
(4)
(5)
(13)
(10)
(8)
(3)
easy sticky rice in a pasta pot
(14)
(62)
(2)
fig preserves cake
(4)
chia protein packed chocolate orbs raw vegan healthy
vegan applesauce carrot cake
chocolate pudding cake vegan
(3)
(2)
(2)
(2)
(2)
(2)
(3)
sardines parmesan and olive tapenade on pide
(2)
tyler florence s shrimp scampi with linguine
(2)
(2)
grand marnier sauce
grand marnier marinated chicken
chicken marbella silver palate cookbook
herb marinated tomato salad with basil
basil tomato salad
fresh basil vinaigrette


basil garlic vinaigrette
lemon basil syrup for beverages
peach basil sangria
asian basil pesto
cowboy coffee steak rub
coffee body scrub
stir soap
dry rub for pork
3 ingredient moroccan dry rub
wet rub for poultry
rotisserie chicken rub
simple marinade and rub for fish
greek herb rub for fish
vagabond fish rub
sassy middle eastern fish rub
(2)
garlic free ketchup
homemade catsup
(2)
(2)
mocha coffee cooler
iced coffee mocha
(2)
lactose free indian curry
country style curry with ground beef and green beans
blue eye cod with curry sauce
balti sauce   basic sauce for anything goes curry
(8)
(3)
(2)
lemongrass lime   thai basil mojito
killer kool aid
canadian killer kool aid recipe
kool aid cocktail
monte cristo coffee drink
monte cristo cocktail
turmeric broth for sipping
rooibos juice
reef juice
orange drink concentrate
lemonade concentrate
(6)
(3)
(4)
(3)
i am sick as a dog tea
comforting cold and flu tea
herbal tea for bronchitis
(7)
(9)
(2)
peach and mint iced tea   adult
iced green tea with jasmine
green jasmine mint iced tea with lemon
zinger green tea drink
(2)
(2)
(4)
(2)
(3)
(2)
(2)
(2)
emeril lagasse s moroccan mint tea
wild mint tea mix
sage   mint tea for sore throats and more
mango mint lassi with indian sweet spices
mint tea
mint chocolate coffee
(7)
(3)
(4)
(3)
tzatziki barefoot contessa ina garten
skip and go barefoot
(4)
(2)
(12)
(11)
(9)
(3)
traditional antipasti platter
antipasto squares
sirloin tips with mushrooms
broiled sirloins
(3)
(2)
(3)
(3)
(2)
(3)
guilt free iced mocha eclair
java chip frappuccino
(6)
(4)
prize winning festive phyllo tarts
(2)
mock filet mignon
filet sorrentino
(7)
(6)
(12)
(29)
(31)
(48)
(48)
(20)
(9)
(8)
(3)
(14)
(6)
(3)
(2)
(7)
(3)
(5)
(5)
(3)
sfenj   north african beignets
(4)
(2)
(5)
(3)
(4)
(9)
(5)
(6)
(6)
xinjiang spice mix
furikake mix
(3)
chex mix booka booka
(10)
(4)
(37)
(15)
(13)
(18)
(6)
(18)
(9)
(11)
(8)
(7)
(65)
(36)
(79)
(11)
(5)
(3)
(3)
(4)
(5)
(3)
(7)
(7)
cheap and dirty garden mary
tsr version of chili s margarita presidente by todd wilbur
queso dip with tequila by rick bayless
tsr version of claim jumper meatloaf by todd wilbur
(2)
chocolate french toast pain perdu by melissa d arabian
tony roma s carolina honeys bbq salmon by todd wilbur
sublime scrambled eggs by gordon ramsay
(5)
(2)
(3)
(3)
(8)
(21)
(3)
waffle of insane greatness
(4)
(5)
toffee almond blondies
toffee blondies
twix blondies
blondies with variations
(3)
(5)
(4)
(4)
(2)
white lady no egg
lazy eggs
pit egg bake
baked egg boats
greek deviled eggs
deviled egg chicks
the geezer s deviled eggs
emeril s devilish eggs
(2)
5 c s crescents
easy artichoke crescent squares
ciboia
(8)
(7)
(2)
(2)
(6)
(9)
(2)
butter pecan bundt cake marry me cake
(5)
(3)
(6)
(3)
(4)
julianna s macaroon recipe
(3)
high alititude french macarons
(4)
(3)
(2)
(2)
latin quinoa
(3)
(2)
mini mexican cherry pepper quiches
(4)
homemade herbed chevre spread with grilled crostini
crantini
banana crantini
cramosa   variation on a theme
french crullers
crepe batter a la julia child


best ever croutons
(4)
(6)
(3)
(2)
nutty cranberry scones
coffee nut scones
scones filled with jam
the empress scones
royal tea room scones
darcy s ginger scones
(3)
tarte au sucre francaise french canadian sugar pie
(2)
dutch velvet
(2)
(4)
(8)
(2)
baked rigatoni dallas morning news
(2)
(2)
(3)
peposo alla fornacina
(2)
(2)
(2)
calamari fra diavolo over linguine
(2)
frittata
veggieble frittata
traditional frittata
(2)
parisian
le petit marseillais
(3)
(2)
pasteles de yuca
pastel de tres leches
(2)
(3)
spaghettini verdi
alfajor de maicena cornstarch alfajor
spanakopita strata
spanakopita dip
spanokabita pockets
mediterranean haddock packets
tiramisu martini
hot tiramisu drink
easy cappuccino trifle poor man s tiramisu
individual tiramisu cups
(2)
ginger crinkles
karen s molasses crinkles
super fabric softener tip
taco pinwheels no black olives
black bean pinwheels
feta craisin and walnut pinwheels
(2)
amarula drumsticks
grilled jamaican jerked drumsticks
(4)
(2)
(7)
(2)
easy layered dip
easy layer dip
7 layer elote dip
(2)
shane s spinach dip
(2)
snow crab dip
(6)
(4)
(11)
(5)
menopause hot flash remedy
flash lightning
spam supreme
spam hotdish
island spam
(11)
(20)
(5)
(2)
(2)
yum yum no bake 3 layer bars
(2)
triple layer banana cream pie bars
(3)
(5)
(11)
(9)
(13)
(10)
(3)
(2)
(2)
(3)
(3)
(2)
(4)
(2)
(2)
bananas foster cheesecake squares
park plaza squares
kynda s squares
(2)
(2)
puffed party squares
party dream squares
(3)
triple berry crisp
pine berry crush
banberry surprise
(2)
(3)
james s blackberry cowboy
raspberry strippers
apple strips
raspberry sticks
(2)
pinto quick bread with dates
(2)
date walnut double decker
(2)
spicy sweet wrapped dates
pigskins in a blanket prosciutto wrapped stuffed dates
satah s powdered date bars
date nut bars
adirondack applesauce and date bars
canadian date squares
(9)
(6)
(6)
(19)
(3)
(3)
(3)
(4)
(5)
(2)
(4)
hg s 104 calorie creme brulee ww points 2
(2)
bruschetta 1 ww point
(2)
(2)
(19)
(10)
(3)
(7)
(11)
(5)
(3)
(3)
(2)
(6)
(6)
(2)
(3)
(7)
(11)
(12)
(4)
parker house rolls made by hand
o charley s rolls
satin sheets
simple chocolate chess pie
mrs steward s lemon chess pie
ooey gooey chess bars
lemon blueberry gooey chess squares
(10)
(8)
(5)
(3)
easy egg rolls two cooking methods
easy butterhorn rolls
slow roll
(4)
(2)
spicy sweet potato puff
cheese potato puff
(3)
(2)
pastry puffs w sweet cheese filling
nutella pastry puffs puff pastry with nutella filling german
puffy relleno
puff pasty surprise
texas lizzies
maine blueberry gateau
kelleigh s peachy special
saskatoon delight blueberries
saskatoon berry jam
saskatoon coleslaw
grilled beer braised bratwursts with cranberry pickle chutney
beer braised bratwursts with onion
homegrown marshmallows
candy cane marshmallows
(3)
sweet cinnamon apple rings
(6)
(3)
il fornello roast ontario apple salad
apple orchard bean bake
apple pie shooter
apple cake shooter
apple core
whole pears roasted in garlic
roasted pears with blue cheese
pear and lime martini
asian pear martini
pear breeze martini
grapefruit pear drink
maple rum cider
cider heaven cider and rum over ice
vodka caramel apple cider
royal cider grog
kings cider
hot mulled cider
cider cooler
raspberry cider

1.2 1.0 0.8 0.6 0.4 0.2 0.0

Let's cluster images!!

For this demo, I'm going to use the following image dataset:

1. A tiny subset of Food-101 from last lecture (available here).
2. A small subset of Human Faces dataset (available here).

Let's start with small subset of birds dataset. You can experiment with a bigger dataset if you like.

```
device = torch.device("cuda:0" if torch.cuda.is_available() else "cpu")
```

```
import random
def set_seed(seed=42):
    torch.manual_seed(seed)
    np.random.seed(seed)
    random.seed(seed)
```

```
set_seed(seed=42)
```

```

```

```
import glob
IMAGE_SIZE = 224
def read_img_dataset(data_dir, BATCH_SIZE):
    data_transforms = transforms.Compose(
        [
            transforms.Resize((IMAGE_SIZE, IMAGE_SIZE)),
            transforms.ToTensor(),
            transforms.Normalize([0.5, 0.5, 0.5], [0.5, 0.5, 0.5]),
        ])

    image_dataset = datasets.ImageFolder(root=data_dir, transform=data_transfo
    dataloader = torch.utils.data.DataLoader(
         image_dataset, batch_size=BATCH_SIZE, shuffle=True, num_workers=0
    )
    dataset_size = len(image_dataset)
    class_names = image_dataset.classes
    inputs, classes = next(iter(dataloader))
    return inputs, classes
```

```
def plot_sample_imgs(inputs):
    plt.figure(figsize=(10, 70)); plt.axis("off"); plt.title("Sample Training
    plt.imshow(np.transpose(utils.make_grid(inputs, padding=1, normalize=True)
```

```
def get_features(model, inputs):
    """Extract output of densenet model"""
    model.eval()
    with torch.no_grad():  # turn off computational graph stuff
        Z = model(inputs).detach().numpy()
    return Z
```

```
densenet = models.densenet121(weights="DenseNet121_Weights.IMAGENET1K_V1")
densenet.classifier = torch.nn.Identity()  # remove that last "classification"
```

```
data_dir = "../data/food"
file_names = [image_file for image_file in glob.glob(data_dir + "/*/*.jpg")]
n_images = len(file_names)
BATCH_SIZE = n_images  # because our dataset is quite small
food_inputs, food_classes = read_img_dataset(data_dir, BATCH_SIZE)
n_images
```

```
350
```

```
X_food = food_inputs.numpy()
```

```
plot_sample_imgs(food_inputs[0:24,:,:,:])
```


```
Z_food = get_features(
    densenet, food_inputs,
)
```

```
Z_food.shape
```

```
(350, 1024)
```

```
from sklearn.cluster import KMeans

k = 7
km = KMeans(n_clusters=k, n_init='auto', random_state=123)
km.fit(Z_food)
```

```
huggingface/tokenizers: The current process just got forked, after parallelism
To disable this warning, you can either:
        - Avoid using `tokenizers` before the fork if possible
        - Explicitly set the environment variable TOKENIZERS_PARALLELISM=(true
```

▼ KMeans

```
KMeans(n_clusters=7, random_state=123)
```

```
for cluster in range(k):
    get_cluster_images(km, Z_food, X_food, cluster, n_img=6)
```

```
82
Image indices:  [ 82 334 265  35 114 343]
```

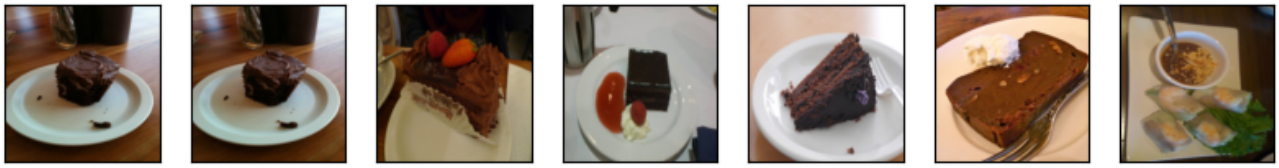
Cluster center 0

```
276
Image indices:  [276 191 186  29 265 343]
```

Cluster center 1

```
161
Image indices:  [161 102 190 241  56  76]
```

Cluster center 2



```
76
Image indices:  [ 76  39 259 295 313 138]
```

Cluster center 3

```
124
Image indices:  [124 253  25  27 223 280]
```

Cluster center 4

```
214
Image indices:  [214  20 141 249 203 274]
```

Cluster center 5

```
48
Image indices:  [ 48  76  61   5  89 138]
```

Cluster center 6

Let's try DBSCAN.

```
dbscan = DBSCAN()

labels = dbscan.fit_predict(Z_food)
print("Unique labels: {}".format(np.unique(labels)))
```

```
Unique labels: [-1]
```

It identified all points as noise points. Let's explore the distances between points.

```
from sklearn.metrics.pairwise import euclidean_distances

dists = euclidean_distances(Z_food)
np.fill_diagonal(dists, np.inf)
dists_df = pd.DataFrame(dists)
dists_df
```

	0	1	2	3	4	5	
0	inf	27.170071	23.031963	29.344770	27.464485	25.905249	28.6
1	27.170071	inf	22.623034	24.378839	25.491880	21.179726	25.2
2	23.031963	22.623034	inf	28.764124	25.544821	23.159750	26.1
3	29.344770	24.378839	28.764124	inf	28.520519	23.345842	28.1
4	27.464485	25.491880	25.544821	28.520519	inf	21.172096	26.4
...	...	...	...	...	...	...	
345	26.458879	23.107458	25.684532	21.547804	25.304211	22.533875	26.8
346	28.636688	21.258495	26.029633	26.480219	25.343542	21.611742	26.2
347	23.513239	19.471357	20.417009	24.750483	23.945749	19.392572	23.9
348	26.565344	20.257633	21.786682	25.170492	23.584101	19.035461	22.5
349	27.202259	22.101439	22.903080	26.520855	25.460316	21.709145	24.9

350 rows × 350 columns

```
dists.min(), np.nanmax(dists[dists != np.inf]), np.mean(dists[dists != np.inf]
```

```
(np.float32(10.06717), np.float32(36.652683), np.float32(24.538565))
```

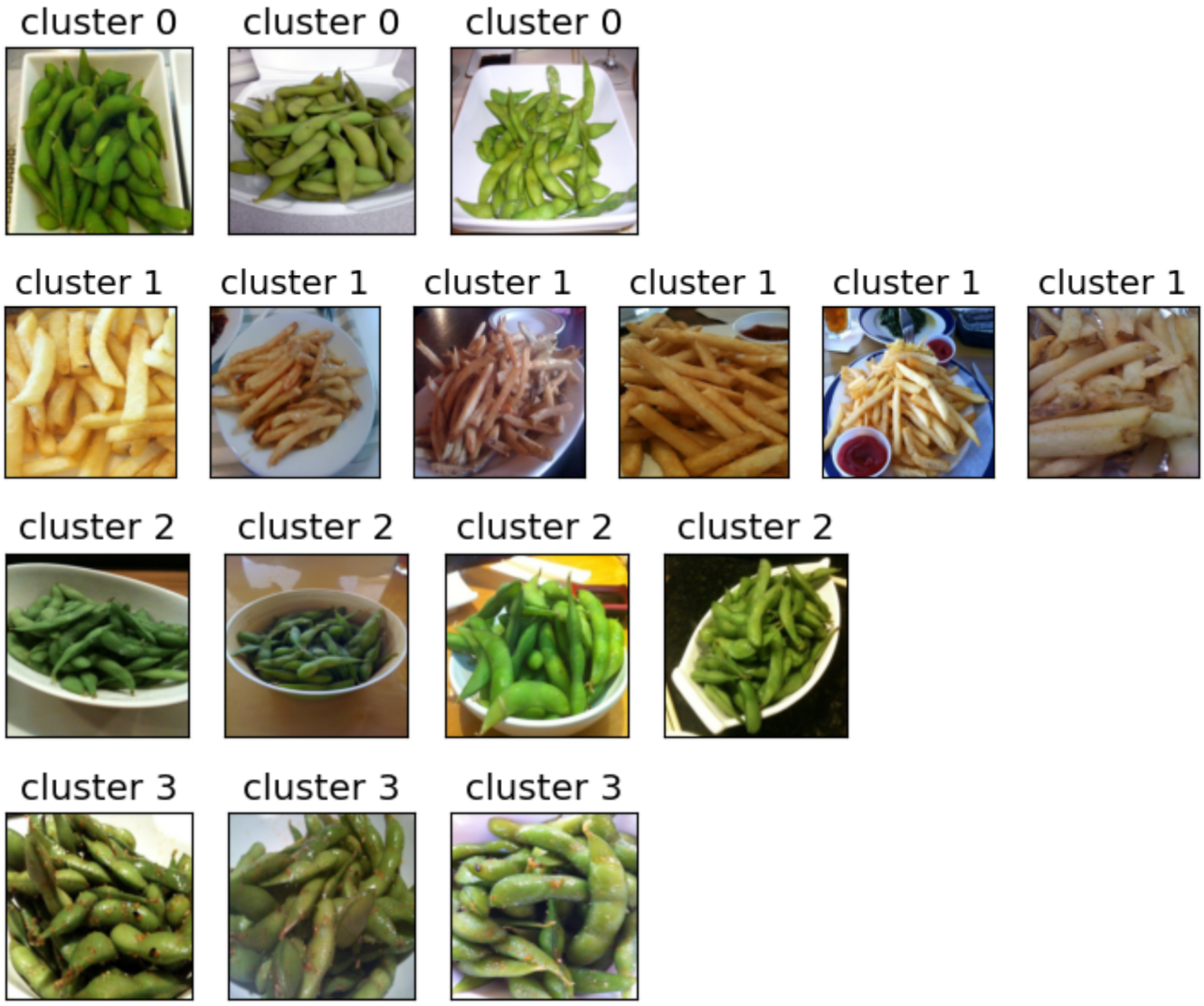
```
for eps in range(13, 20):
    print("\neps={}".format(eps))
    dbscan = DBSCAN(eps=eps, min_samples=3)
    labels = dbscan.fit_predict(Z_food)
    print("Number of clusters: {}".format(len(np.unique(labels))))
    print("Cluster sizes: {}".format(np.bincount(labels + 1)))
```

```
eps=13
Number of clusters: 2
Cluster sizes: [347   3]

eps=14
Number of clusters: 5
Cluster sizes: [334   3   6   4   3]

eps=15
Number of clusters: 4
Cluster sizes: [299  26   8  17]

eps=16
Number of clusters: 4
Cluster sizes: [248  86   3  13]

eps=17
Number of clusters: 2
Cluster sizes: [205 145]

eps=18
Number of clusters: 2
Cluster sizes: [160 190]

eps=19
Number of clusters: 2
Cluster sizes: [116 234]
```

```
dbscan = DBSCAN(eps=14, min_samples=3)
dbscan_labels = dbscan.fit_predict(Z_food)
print("Number of clusters: {}".format(len(np.unique(dbscan_labels))))
print("Cluster sizes: {}".format(np.bincount(dbscan_labels + 1)))
print("Unique labels: {}".format(np.unique(dbscan_labels)))
```

```
Number of clusters: 5
Cluster sizes: [334   3   6   4   3]
Unique labels: [-1  0  1  2  3]
```

```
print_dbscan_clusters(Z_food, food_inputs, dbscan_labels)
```



Let's examine noise points identified by DBSCAN.

```
print_dbscan_noise_images(Z_food, food_inputs, dbscan_labels)
```